POLITECNICO
MILANO 1863

Master of Science in Management Engineering

"How can learning analytics improve educational outcomes? Results from an innovative project."

Supervisor: Prof. Tommaso Agasisti

Candidate: Margherita Mura Matr.: 899991

Academic Year 2018-2019

A very special gratitude goes to my supervisor Professor Tommaso Agasisti, who gave me the possibility to work on a very current and interesting topic, allowing me to work from Sweden.

Furthermore, I would like to thank Marta Cannistrà, research fellow, and Chiara Masci, PhD in Politecnico di Milano, for supporting me in the analyses and for answering to every doubt I had.

Another acknowledgement goes to Professor Anna Paganoni and Professor Francesca Ieva, for having shared their technical knowledge with me.

Thank you also to Politecnico di Milano, for letting me spend the most challenging and rewarding years of my life.

A mio nonno, senza il quale non avrei avuto la passione per i numeri,

Alla mia grande famiglia, che mi ha sopportata in momenti in cui il Politecnico mi ha messo a dura prova e insegnato che non esiste solo lo studio in questo mondo,

A Riccardo, senza il quale lo studio sarebbe stato noioso e monotono e mi ha insegnato che quando si vuole veramente una cosa, con determinazione, la si conquista,

Ai miei amici internazionali di Linkoping, senza i quali questa tesi sarebbe stata una pura applicazione delle teorie già esistenti, che mi hanno fatto comprendere gli ostacoli che affrontano tutti i giorni per frequentare un'università lontana da casa.

Abstract

In an always more globalized world, with an always more significant economy driven by international students, universities, policymakers, and job recruiters need to have a more in-depth and a more unobstructed view on the characteristics by which their university's performance are affected, trying later to give them the proper support.

From the literature, a discrepancy in the performance between international and domestic students is highlighted. To further explore this, in this project a multi-dimensional model will be designed. It will be tested on the students belonged to a Master of Science in a Technical University in Italy thanks to two machine learning algorithms.

The aim of this thesis is giving a contribution to the still unexplored field of learning analytics for international students.

Keyword

Learning analytics, International students' performance, Economics of Education, Academic performance predictions, classification algorithms

Abstract

In un mondo sempre più globalizzato, con una crescita dell'economia verso il mercato degli studenti internazionali, le università, i policy makers, le agenzie di reclutamento per il lavoro hanno bisogno di avere una visione più precisa delle caratteristiche che condizionano le performance degli studenti internazionali nelle università, provando a dare loro il supporto necessario.

Da un punto di vista letterario, il gap tra le performance degli studenti internazionali e quelli che invece risiedono nella nazione dell'università è evidente. Con l'obiettivo di investigare questo punto, un modello multi-dimensionale sarà sviluppato. Sarà in seguito testato sugli studenti iscritti ad una laurea magistrale in una università tecnica italiana attraverso l'utilizzo di due algoritmi di machine learning.

L'obiettivo di questa tesi è quello di cercare di dare un contributo nell'ancora poco esplorato ambito di learning analytics per gli studenti internazionali.

Keyword

Learning analytics, performance degli student internazionali, Economics of Education, predizione delle performance accademiche, metodi di classificazione

Executive summary

This thesis aims to give a contribution to the field of learning analytics in higher education, with the focus on international students at university.

In an always more globalized world, with an always more significant economy driven by international students, universities, policymakers, and job recruiters need to have a more in-depth and a more unobstructed view on the characteristics by which their university's performance are affected, trying later to give them the proper support.

From the literature, a discrepancy in the performance between international and domestic students is highlighted. To further explore this, a multi-dimensional model will be designed, basing on casual effects between the factors that might influence students' academic behaviour and students' performance.

It will be tested on the students belonged to Master of Science's program in a Technical University in Italy using data from the academic year 2013 to the academic year 2015, after having assessed that international students do perform worse than domestic ones, also in this university, in term of grade average points (GPA), retention rate and time spent for graduating.

In particular, the result of this highlights that usually the culture of the international students and the course of study chosen affect dramatically their performance. *Indians* and *Pakistanis* are the ones who dropout more, *Iranians* are the highest percentage of students who took more to graduate but they are the lowest percentage of students who have a lower GPA, *Colombians* and *Turkish* have almost no students who have a high time to graduation. Regarding the course of study, it seems that *Computer Science* is the course with the highest percentage of the international students' dropouts, *Energy engineering* has the highest percentage of students who took more years to graduate and *Aeronautical engineering* is the course with the highest percentage of students who have a final GPA lower.

International students will be further analysed with two different algorithms of machine learning: logistic regression with mixed effects and decision tree with mixed effects both with random effects on the nationality of the bachelor of the students and on the course of study chosen to study in Italy.

This will allow understanding which variables are the most relevant in predicting their performance. What emerged from the results was that the most significant variables to predict their performance were: *GPA in the first year* affects positively both the probability of dropout and the probability of being taking more to graduate, *the total number of credits taken in the first year* affects positively all the performance, *the average number of attempts per exam in the first year* affects negatively the students who take more time to graduate and the students who have a lower GPA.

Furthermore, thanks to the random effects of the algorithms, the readers will have the chance to understand which are the countries and courses of studies in which the students have the most significant probability to have a lower o higher performance. In particular, the most significant nesting belongs to the students who have a lower GPA, in which *the course of study* appears to be very relevant in determining their performance. Less significant was the *nationality* for the time to degree. Regarding the students who dropout, neither the nationality nor *the course of study* were particular critical in predicting this performance.

Finally, the choices about which of the algorithms perform better related to the different performance will be taken: the models chosen were the logistic regression with random effects on the course of study for the dropout, the logistic regression with random effects on the nationality for the students who took more years to graduate and the classification tree with random effects on the nationality for the students who have a lower GPA.

At the end, in the last paragraph the managerial and policy implications will be explained.

Table Of Contents

List of Tables

List of Figures

Chapter 1. INTRODUCTION

With an increase in data availability in the current world, a significant focus of the research is taking place in the field of data analytics. In recent years, researchers have found the potentiality of studying data in higher education in order to predict and take actions on the students who more need help.

Siemens and Long (2011) stated that "analytics is seen as one of the most dramatic factors in shaping the future of higher education".

Data has always been a significant asset for academic institutions, and it has been used to ease their daily operational decisions as well as to plan longer-term business decisions. On a strategic scale, data is used to advise business planning and overall strategy for institutions. Students' enrolment data influences the decisions of schools and universities to design specific programs, while financial data influence strategic decisions on expanding or reducing particular faculties or services provided.

The hardest challenge is to use available data to gain further insights from an educational point of view: data help university's managers and employees in taking their decisions, such as which entry barriers to build in order to keep only the most motivated students, or the preventive actions to take, based on the predictions trained with the past data, in order to let the students succeed in their careers.

Indeed, the learning process of a student is influenced by his/her characteristics, their family, their peers, the neighborhood in which s/he lives, as well as by the characteristics of the school that s/he is attending and his/her early academic life. Moreover, how the various variables affect educational outputs is likely to change substantially across the educational systems operating in different countries. Ideally, academic data can be collected, linked together and analysed to provide insights into students' behaviour and identify patterns that could allow predicting their future performance.

Researchers and developers from the educational community started exploring the potential adoption of predicting techniques for gaining insight into academic activities.

Even the list of goals and objectives that can be pursued with the application of analytics to students' big data can be very long:

- Improve students' results: during his/her academic life, each student generates a unique data path. This can be analysed in real-time to deliver a customized learning environment for the student, as well as to gain a better understanding of his/her individual behaviour. Besides, with the help of appropriate algorithms, it could be possible to determine the strengths and weaknesses of each student.
- Create Mass-customized Programs: students will be given the opportunity to develop their personalized program, as well as mass-customized programs, which aggregate students with the same needs in the same cluster and allow them to receive what they need. An example that has already taken place is the Massive Open Online Courses (MOOC), courses of study available over the Internet without charge to a vast number of people.
- Improve the learning experience in real-time: each student learns differently, and this affects his/her final grade. Some students learn very efficiently while others may be extremely inefficient. If available, this information could be used to provide a customized program or real-time feedback to learn more effectively and, thus, improve the results.
- Reduce dropouts, increase results: using predictive analytics on all the collected data can give educational institutes insights into future student's outcomes. Moreover, using a predictive model based on historical data, they can take early actions on those students that are more at risk and try to avoid their retention from the school.
- Understand difficulties faced by international students: with an increase of globalization, there has been an increasing number of students studying in a university that does not belong to their home country. This allows the researches to understand that in many cases they perform worse than home students for broad reasons.

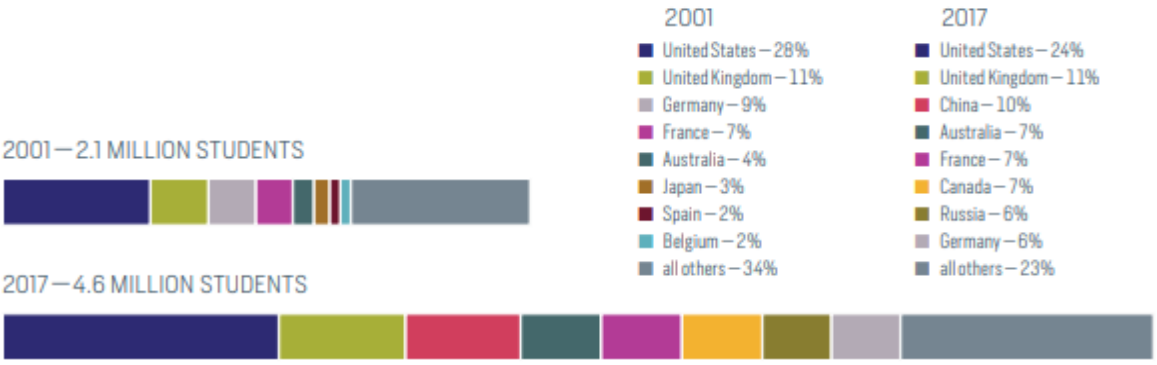
In this work, there will be a focus on the performance of international students, and, based on the previous works, a framework will be designed in order to find the variables which impact the most on their outcomes.

It all started when more than 2,000 years ago Horace went to Athens to join Plato's Academy and when in 1190 Oxford University admitted its first known international student, Emo of Friesland (The Economist, 2016).

According to the UNESCO Institute of Statistics, the number of globally international students increased to 3.4 million students in 2009, up from 2.1 million students in 2002. The two leading destination countries are the U.S.A. and the UK.

The table below shows the number of students by country of origin.

Figure *1.1. Host destination for globally mobile students 2001-2017*



source: UNESCO 2017

One in five of the world's international students is from either China or India. In the U.S. alone, these two countries contributed to 84% of all increases in international student enrolment between 2000 and 2011.

A common belief among educators is that international students are insufficiently adjusted to higher education in their host country, both academically and socially. Furthermore, several groups of international students experience considerable amounts of stress while adapting to the culture of the host institute. Several researchers argue that studies on adaptation of international students should widen its focus to the underlying mechanisms that lead towards this ''misalignment''.

To better understand this potential gap, the model designed will be applied to a real case, focusing on the data available of a technical university in Italy.

Students enrolled in an engineering Master of Science University from 2010 to 2015 are the ones I will analyse and train the model with.

After a careful data merging, preparation and transformation, I will analyse in a descriptive way the variables that I create in order to find insights about the differences across national and international students.

Secondly, after some proper data preparation and transformation, thanks to some machine learning algorithms, I will find the variables that impact the most on students' outcomes.

I will use different machine learning algorithms to predict the final performance of the international students, with students who enrolled in 2015 as a test.

In the end, there is the comparison between my results and the hypothesis I thought could be true.

Unfortunately, it is not possible to test all the hypothesis in the model, since the data I have does not have all the necessary information.

Moreover, there is a try to give the reader some inputs in order to find the best solution to international students' problems.

1.1. Motivation

Thanks to the use of predictive analytics from various data sources, problems can be predicted early, and intervention plans can be designed. This gives a full view of each student and provides a one to one learning experience to the student.

Eduventures (2013) also stated that by investigating historical data, predictive analytics could let an institution know which applicants are expected to enrol: far ahead in the student life cycle which students are likely to go on and graduate and which students will have the best performance.

The increased presence of international students has raised new pedagogic questions such as whether the way teachers currently teach can effectively get the learning needs of both home and international students, and whether special skills in teaching across cultures need to be developed. With much work emerging on how institutions can manage this diversity and ensure quality learning for all students, little research has been undertaken in the area of assessment of learning in multicultural educational settings.

This is striking in that such information may serve as a useful indicator to monitor the appropriateness of current entry requirements for international students, and to make inferences on the extent to which adopted assessment methods are culturally fair indicators of ability across diverse student groups.

Although the choice to study abroad has a series of advantages, it also involves considerable difficulties: compared to local students, international students have fewer resources to meet the needs they face, and they experience major adaptation problems.

In fact, they have to face a double challenge: in one hand, the academic difficulties experienced by every student who joins a university system; on the other hand, they face the difficulties related to the process of acculturation and adaptation to the new context of life, experimenting in many cases acculturation stress, often associated with negative emotional experiences and on the onset of psychological symptoms.

Moreover, they can experience academic cultural shock, led by a different way of studying between countries and different way of behaving in a classroom. For all these reasons, they might perform worse than their home peers.

After having read the literature, it is understood that in the majority of the works, just one cluster of variables is taken into consideration when analysing the causal link between

the cluster, such as social integration, or cultural shock or students' background, and students' outcome. Consequently, there is a need to explore a multi-dimensional model in order to see the overall relationships between the variables of different clusters that affect the performance and in which way the variables contribute to the final performance.

Further exploring the background in which my case of study is designed, I found that Italian universities are hosting always more international students principally on their Master of Science or Doctor of Philosophy.

According to La Stampa, international students mainly choose the linguistic, architectural, political-social, economic-statistical and engineering sectors. They graduate on average with a little delay compared to their Italian colleague and with a lower marks.

Once the qualification has been obtained, however, many returns abroad, often to their country of origin, attracted by labor contracts that are more durable, profitable and obtained in less time.

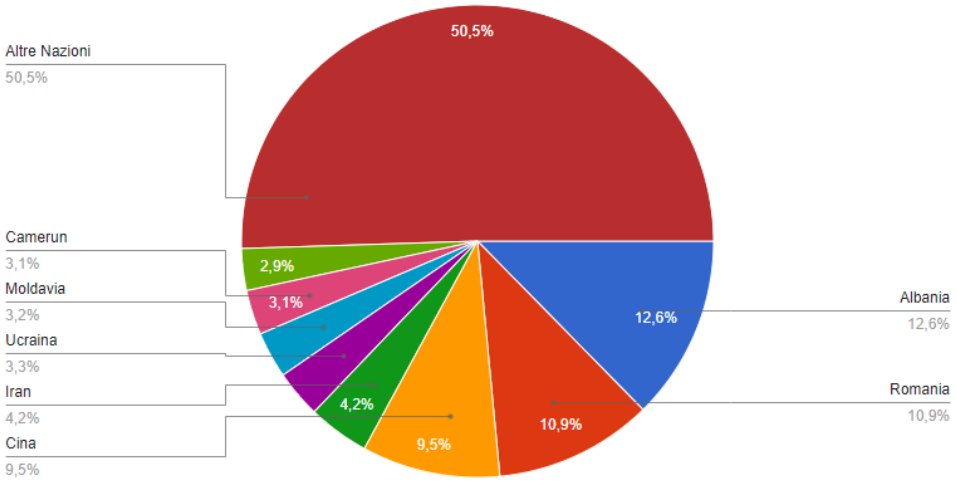
The increase of them, whom today represent 3.5% of the total number of graduates, places Italy on the eighth position in the OECD ranking on the attractiveness of the university system. In fact, out of 100 students who choose to study abroad, the podium goes to the United States (26.3%), the United Kingdom (15%) and France (10.5%). Followed by Germany (9.8%), Australia (8.3%), Japan (2.9%), Canada (2.7%) and Italy (2.6%).

Moreover, according to University to Business, the internationalization of the Italian universities has long been included among the objectives of the Universities by the reform law 240 of 2010: internationalization that would allow Italian universities to compete with the bests in the world, and that would allow to overcome language barriers with the inclusion of courses in foreign languages, in particular with the use of the most common English .

This gave rise to the 'The university case', which in 2012 established that from 2014 all the master's degree courses and doctorates would be in English, a decision that had opened a diatribe especially between teachers that was taken up to the exam before the TAR and then the Council of State. Recently, Minister Fedeli intervened directly, affirming that English can never replace Italian in universities.

Analysing deeper where international students in Italy come from, thanks to Ustat's data, I can notice in figure 1.2.1. that the first country of origin is Albania, followed by Romania and China, data is from 2016/2017.

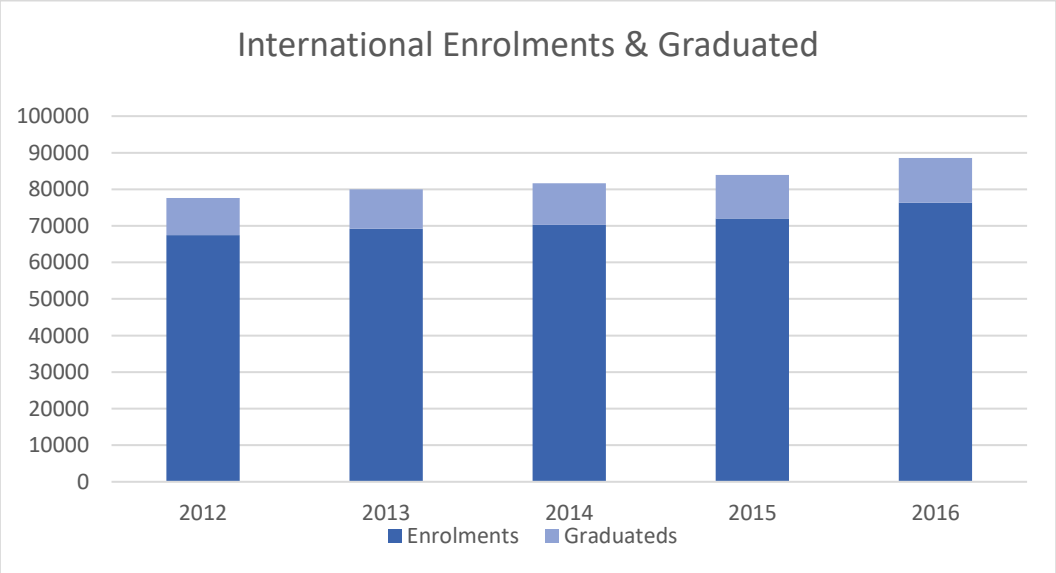
Figure 1.2.1. Percentage distribution of international students coming to Italy to attend their higher education



Source: Ustat

Moreover, figure 1.2.2. shows the number of enrolled and graduated of these international students.

Figure 1.2.2. Number of international students' enrolments over the number of graduated in Italy over the years



Source:Ustat

According to the shown increase of international students in Italy, my first aim will be supporting policy-makers to take decisions regarding them, such as which ones to enrol, which ones need help, which kinds of help do they need, which kinds of tool could be valid on their performance.

In particular, this can be divided into subobjectives as follow.

Firstly, there is a need to understand how deep the literature is going into the issues faced by international students when going in a foreign country and try to find the gaps that the literature has not covered yet.

The second objective is to further analyse the performance of international students compared to the home country's one, catching the differences and the similarities and trying to reason why there are these gaps.

Thirdly, there is a need to understand which are the causes of these differences and the relationship between the dimensions of the most relevant variables.

The final objective is to suggest some methods or tool policy makers can apply in order to increase the international students' performance and to give them helps to succeed, according to the relations found within the previous objective.

1.2. Research questions

With my work, I will try to focus on the following research questions.

1. Which are the differences between international and home country students?
   The answer to this will be given in chapter 2.5 and 7.2.
2. Which factors are the most relevant in predicting the performance of international students?
   The answer to this will be given in chapter 2.5, 3.3 and 7.3.
3. Will I be able to predict international students' performance based on machine learning algorithms?
   The answer to this will be given in chapter 7.4.

1.3. Chapters

The following chapters will include:

Chapter 2. Received literature: firstly, it is shown an analyse of the studies in learning analytics, and then there is a focus on international students and on identifying the most common factors that are presented in the literature that affect international students' outcomes.

Chapter 3. Theoretical framework: in this chapter the theoretical model is shown, and the hypotheses related to this are presented. Here, the model is a three-dimensional model, in which each dimension is characterized by its own variables that together contribute to affect the outcome.

Chapter 4. Background, Data available and Data Pre-processing: the field of application of the model is presented, there is a focus on the analysis of international student in the university. Moreover, here there is a section dedicated to the description of the dataset and of the data pre-processing. Moreover, the most significant variables are shown.

Chapter 5. Methods: in this chapter the theoretical algorithm of machine learning used to hold the analysis are presented. Moreover, a section on how to fit the models and how to read their results are shown.

Chapter 6. Results: the answers for each research question are provided, by means of an adequate empirical analysis.

Chapter 7. Discussion and conclusion: in this chapter there are presented the limits of the analysis held, the further developments and the impact that the analysis could have in the literature about international students and its managerial and policy implications.

1.4. Definitions and abbreviations

International students	Students enrolled in a university in a country different from the one of their previous study
GPA	Grade average point: is calculated in two steps: 1. the grade awarded for each course is multiplied by the credit value for each course. 2. The aggregate score is divided by the total number of credits for all courses completed in the defined period of study.
Dropout	Students whose carrier status is not active and not graduated
ETCS	European Credit Transfer System: represents the workload and defined learning outcomes ("what the individual knows understands and can do") of a given course or program.
CFU	Credito Formativo Universitario: is a method used in Italian universities to measure the workload required of the students
LA	Learning analytics
MOOC	Massive Open Online Courses: are courses designed for distance learning involving a large number of users
EU	European Union
UK	United Kingdom
USA	United States of America
MSC	Master of Science
PhD	Doctor of Philosophy
CQ	Cultural intelligence: describes the cultural knowledge and ability of an individual to adapt his/her interactions with persons and countries of other cultures.
TOEFL	Test of English as a Foreign Language
IELTS	International English Language Testing System
Home students	Students who attended the previous education in the same country as the current one
GPA<23/ GPAinf23	Students who had the final GPA lower than 23 (23/30)
T2D>3	Students who took more than 3 academic years to graduate in their Master of Science If a student enrolled in September 2017 and graduate in April 2021, it is considered to took 3 years. The "plus one" academic year is calculated from July of the following year.
CDS	Course of study
A.A.	Academic Year Usually it lasts from September to the following September
Slow student	A student who took more than 3 academic years to graduate

Fast student	A student who took less than 3 academic years to graduate
University, Italian and International	University: student who attended his/her bachelor in the university selected Italian: student who attended his/her bachelor in another Italian university International: student who attended a not Italian university in the bachelor
Big university	A university with a high number of students coming to finish their academic year in the considered university (more than 45 students from 2010 to 2018)
Small university	A university with a small number of students coming to finish their academic year in the considered university (less than 45 students from 2010 to 2018)
#	Number
AVG	Average
ICC	Interclass correlations
VPC	Variance Partition Coefficient
RSS	Residual Sum of Squares
TP	True positive
TN	True negative
FP	False positive
FN	False negative

Chapter 2. RECEIVED LITERATURE

2.1. Overview

The main aim of the literature review is to provide the state-of-art situation about international students' performance and problems when studying abroad. The information has been retrieved from existing scientific literature in the field.

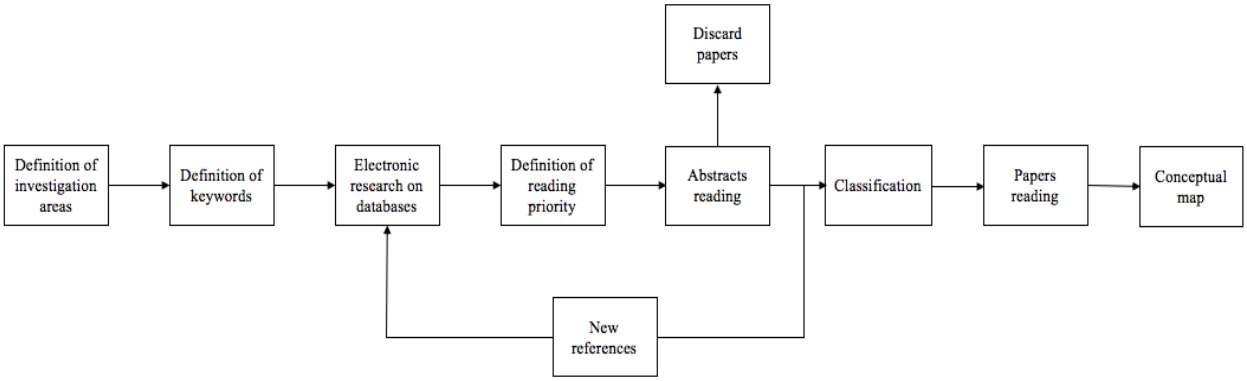
This chapter will be divided into seven main sections: firstly, the methods used to hold the analysis are presented, secondly an introduction about the topic is explained, then big data in higher education is explored. Later on, the predictive models of learning analytics used by the previous literature are presented, then a section regarding a focus on international students is shown. Here the reader can find the main cluster of variables already explained by the literature. In the following topic the tools and methods used to contrast the lower outcomes of students are shown. Finally, there is a subchapter about the conclusion and the take-off of the literature analysed.

2.2. Method for selecting academic papers

The literature review was divided into two different types of research: the first one is related to the determinants of students' success in higher education, and the second one related to learning analytics approaches.

The different phases of the process are described by the Figure below.

Figure 2.2.1.1. Process of selecting the paper for the literature review



Source: Author's release

The literature review started by identifying the keywords needed in order to cover this analysis. In the different steps of the literature review there was been looking for different keywords and filters.

At the beginning they were:

Table 2.2.1. Filter for articles' selection

KEYWORDS	Performance in higher education, learning analytics in higher education, dropout, the economics of education
TIME FRAME	2000- 2018
DATABASE	Scopus, JStore
TYPE OF LITERATURE	Academic journals, book chapters and articles, dissertations

Source: Author's release

Notes: about 2700 papers for performance in higher education, 500 for learning analytics in higher education, dropout in higher education 150 articles, the economics of education about 3700 articles.

After having read the abstracts of the ones that appear to be more relevant for the project's purpose and having selected the ones that appear to be more relevant according to the number of citations they have (more than 50), there were selected about , there had been selected about ten articles and they had been read.

Afterward, there had been looked to their references and their own cited articles, in order to find some interesting articles to go deeper into some topics. Overall, there had been read in this stage about 55 articles.

The next step was to look for a new keyword: "international student performance" (100 articles available following the criteria mentioned above in the table) and trying to find the related articles. After having selected about ten articles and read them, there had been looking at their references and their own cited articles, in order to find some interesting articles to go deeper into some themes.

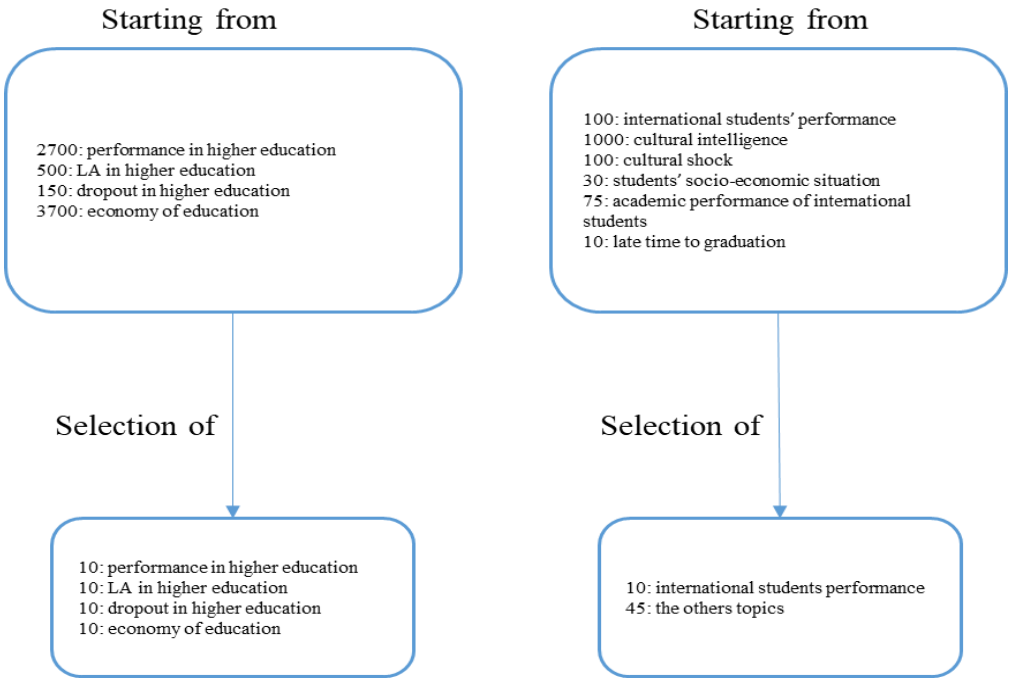
Finally, after having understood the most recurrent causes of international students' performance, there had been looked for new articles with the keywords: "cultural intelligence (1000 articles available), cultural shock (100 articles available), the socio-economic situation for international students (30 articles available), academic performance of international students (75 articles available), late time to graduation (10 articles available)".

There had been selected about ten articles to read and then there had been looked into their references and their own cited articles. Overall, in these second steps, about 45 articles were deeply read.

At the end, before starting to write the literature discussion, a conceptual map about how to develop it was designed.

The Figure below shows the process of filtering the literature received.

Figure 2.2.1.2. Number of literature's papers found with the criteria shown in table 2.2.1. with different keywords



Source: Author's release

Among all the articles read, the literature selected at the end was:

Table 2.2.2. Papers chosen and deeply analysed

TITLE	AUTHORS	YEAR	ACADEMIC JOURNAL/BOOK	TOPIC FACED IN THE ARTICLE (P:performance of studets, A: learning analytics approaches)
Inter-university variations in undergraduate noncompletion rates: A statistical analysis by the subject of study	Johnes	1997	Journal of Applied Statistics	A
The Determinants of Undergraduate Grade Point Average: The Relative Importance of Family Background, High School Resources, and Peer Group Effects	Betts, Morell	1999	The Journal of Human Resources,34	P
International Students, Learning Environments and Perceptions: a case study using the Delphi technique	Robertson, Line, Jones, Thomas	2000	Higher Education Research & Development	A
Inclusive approaches to effective communication and active participation in the multicultural classroom	De Vita	2000	Active Learning In Higher Education	P
Determinants of degree performance in UK universities: a statistical analysis of the 1993 student cohort	Smith, Naylory	2001	Oxford Bulletin Of Economics And Statistics	A/P

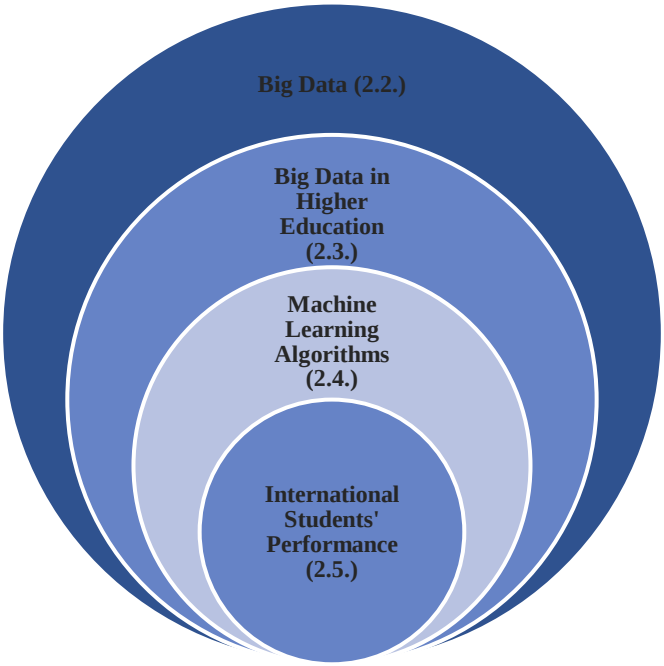
The Relative Effect of Family Characteristics and Financial Situation on Educational Achievement	Chevalier, Lanot	2002	Education Economics	P
Student performance in tertiary-level accounting: an international student focus	Hartnett, Römcke, Yap	2004	Accounting and Finance 4	P
Higher education outcomes, graduate employment and university performance indicators	Bratti, McKnight, Smith	2004	J. R. Statist. Soc.	P
Degree Performance Of Economics Students In Uk Universities: Absolute And Relative Performance In Prior Qualifications	Naylor, Smith	2004	Scottish Journal of Political Economy,	A/P
Researching the performance of international students in the UK	Morrison, Merrick, Higgs, Le Métais	2006	Studies in Higher Education	P
Why Students Leave Engineering: The Unexpected Bond	Fleming, Engerman,, Williams	2006	American Society for Engineering Education Conference	A
Relative Success? Determinants Of College Graduation Rates In Public And Private Colleges In The U.S.	Bailey, Kienzl	2006	Research in Higher Education,	A
Data-driven Personalization of Student Learning Support in Higher Education	Liu, Bartimote-Aufflick, Pardo, Bridgeman	2006	Learning analytics: Fundaments, applications, and trends	A
Toward a Cultural Advancement of Tinto's Theory	Guiffrida	2006	The Review of Higher Education	P
Student attrition and academic and social integration: Application of Tinto's model at the University of Papua New Guinea	Mannan	2007	Higher Education	P/A
Factors influencing university drop out rates	Araque, Roldán, Salguero	2009	Computers & Education	P/A
University drop-out: an Italian experience	Belloc, Maruotti, Petrella	2009	High Educ	A
Cultural Equivalence in the Assessment of Home and International Business Management Students: a UK exploratory study	De Vita	2010	Studies in Higher Education	P
Determinants of International Students' Academic Performance A Comparison Between Chinese and Other International Students	Li, Chen, Duanmu	2010	Journal of Studies in International Education 14	P/A
Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities	Willcoxson, Cotter, Sally Joy	2011	Studies in Higher Education	P/A
Course Signals at Purdue: Using Learning Analytics to Increase Student Success	Arnold, Pistilli	2012	Proceedings of the 2nd international conference on learning analytics and knowledge	A/P

Mapping world cultures: Cluster formation, sources and implications	Shankar	2013	Journal of International Business Studies 44	-
Goal-efficacy framework: an examination of domestic and international accounting students' academic performance	Phang, Shireenjit, Cooper	2014	Accounting and Finance 44	P/A
Exploring the influence of individual and academic differences on the placement participation rate among international students A UK case study	Crawford, Wang, Andrews	2015	Education + Training	P
Foreign Travel Experience and Cultural Intelligence: Does Country Choice Matter?	Engle, Nash	2016	Journal of Teaching in International Business	P
Improving students' performance in quantitative courses: The case of academic motivation and predictive analytics	Rahal, Zainuba	2016	The International Journal of Management Education	P
Handbook of Learning Analytics	Lang, Siemens, Wise, Gašević	2017	Solar	A
Predicting Four-Year Student Success from Two-Year Student Data	Nadasen, List	2017	Big Data and Learning Analytics in Higher Education,	A/P
The current landscape of learning analytics in higher education	Viberg, Hatakka, Bälter, Mavroudi	2018	Computers in Human Behaviour	A
International Students' Academic Achievement and Progress in Turkish Higher Education Context: Students' and Academics' Views	Yükselir	2018	Universal Journal of Educational Research 6	P
Does Cross-cultural Competence Matter when Going Global: Cultural Intelligence and Its Impact on Performance of International Students in Australia	Iskhakova	2018	Journal of Intercultural Communication Research	P
Supporting student experience management with learning analytics in the UK higher education sector	Kika	2018	University Of Bedfordshire	A
Beyond English language proficiency scores: understanding the academic performance of international undergraduate students during the first year of study	Neumann, Padden, McDonough	2019	Higher Education Research & Development	P

Source: Author's release

The analysis of the literature is shown in Figure 2.2.1., as the reader can see it has a funnel shape. About the last topic, three main causes have been analysed in the literature: cultural shock, social integration and early factors, shown in Figure 2.2.2.

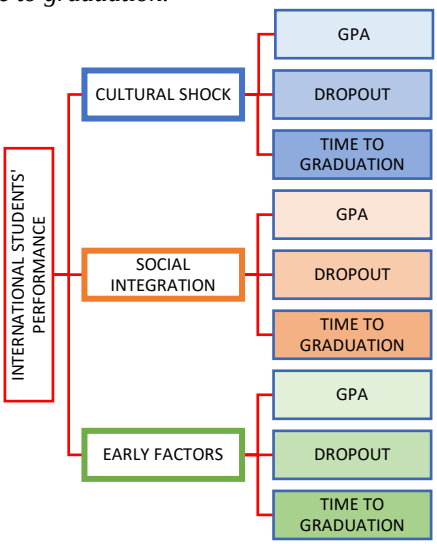
Figure 2.2.1.3. Funnel representation about the way the literature review was hold



Source: Author's release

Notes: the numbers next to the title of the subchapters correspond to each subchapter.

Figure 2.2.1.4. Clusters of variables that, according to the literature affect students' performance in term of GPA, dropout and time to graduation.



Source: Author's release

Notes: they can be found in chapter 2.5

2.3. Big Data

Big Data and technological approaches in the field of Business Intelligence and Analytics are increasingly important in both business and academic communities (Chen et al., 2012). They are techniques and applications used for the analysis of business data to enhance business decisions (Davenport and Harris, 2007).

Big Data is a term invented to focus on the challenges that new data streams have brought to business firms and higher education institutions.

It is often regarded by three factors: volume, velocity and variety. Many authors and researchers have given their definitions of each factor:

1. Volume: the quantity of data obtainable by a firm, which does not necessarily have to own all of it as long as it can access it (Kaisler et al., 2013). For example, Facebook processes 500 TB per day.
2. Velocity: the growing rate at which data moves within a firm and in the World, speed of generation and processing of data (Daniel, 2015).
   For example, the speed at which the financial market changes.
3. Variety: not just a single type of data but also semi-structured data from a variety of sources like web pages, web log files, social media sites, emails, documents, and sensor devices from both dynamic and inert devices (Katal et al., 2013).

2.4. Big data in higher education

The combination of digital technology into higher education (HE) impacts both instructing and learning practices, and enables access to information, basically available from online learning environments, that can be used to improve students' learning (Schumacher and Ifenthaler, 2018). This gave to Learning Analytics (LA) the possibility to rise.

Learning Analytics is "the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data" (Cooper, 2012). It was defined by Del Blanco et al. (2013) as "a discipline that gathered and analysed educational data with different purposes such as seeking a pattern in the learning process and trends or problems in student performance", and by Greller and Drachsler (2012) as an academic area that concentrated on learners, their learning procedures and behaviours.

LA, emerged as a fast-growing and multi-disciplinary area of Technology-Enhanced Learning (Ferguson, 2012). In LA, information about learners and learning environments is used to "access, elicit, and analyse them for modelling, prediction, and optimization of learning processes" (Mah, 2016).

LA reflects the rise of data-intensive approaches to education (Siemens and Baker, 2012). One of the most recent review (Ferguson and Clow, 2017) incorporated the investigation of the proof of whether LA improves learning practice in HE dependent on four recommendations of LA: they improve learning results, they bolster learning and educating, they are conveyed broadly, and are utilized morally.

Moreover, learning analytics requires to generate new big data infrastructures that include both human and nonhuman actors: performance indicator metrics, data warehouses, data files, spreadsheets, information and records systems, visualization software, algorithm-led analytics packages and institutional dashboards, plus data managers, data stewards, business managers, financial officers and deans. Building the data infrastructure of higher education also sometimes requires private sector outsourcing companies, software developers, cloud hosting firms, data analytics designers and a host of other technical specialists (Williamson, 2017).

Focusing on the different sides of LA, table 2.4.1. introduces the classifications utilized by Grönlund and Andersson (2006) for the strategies for information gathering.

Table 2.4.1. The strategy of information gathering

Research approach	Definition
Descriptive	Describes a phenomenon in its appearance without any use of theory
Philosophical	Reflects upon a phenomenon without data and any use of theory
Theoretical	Reflects on a phenomenon based on some theory but without empirical data
Theory use	Applies a theory/theories or models as a framework for the conducted study
Theory generating	Analyse data in a systematic manner with a purpose of building a theory
Theory testing	Test a theory using data in systematic manner

Source: Grönlund and Andersson (2006)

There are several ways to collect data: most of the papers (57%) undertook a descriptive research approach, followed by theory use research (26%). There is no dominating theory, but they try to explain different aspects of LA, such as human behaviour (Gilmore, 2014), learning and knowledge outcomes (Kovanović et al., 2015), or technology acceptance and use (Nistor et al., 2014).

Furthermore, different methods of data collection such as interpretative study, experiment, literature study, survey, product description, argument and ethnography are used. The most common method is interpretative methods, 68% of the studies use interpretative data collection methods, such as interviews or focus groups. The second most commonly applied method is an experiment (18%), the third one is product description (15%), followed by surveys (11%) (Merceron, 2015).

Table 2.4.2. Methods to analyse data

Methods	Description
Prediction	A major task tackled by prediction methods is to predict performance of students. The most common methods include regression and classification.
Clustering	Clustering techniques are used to group objects so that similar objects are in the same cluster and dissimilar objects in different clusters.
Relationship Mining	This category includes such methods as association rule mining, correlation mining, sequential pattern mining and casual data mining.
Distillation of data for human judgement	This category includes statistics and visualizations that help humans make sense of their findings and analyses.
Discovery with models	This category encompasses approaches in which the model obtained in a previous study is included in the data to discover more patterns.

Source: Merceron (2015)

Focusing now on another aspect, the methods used to analyse data are different and they are summarized by Merceron (2015) and shown in the table 2.4.2.

The study held by Vilberg et al. (2018) showed that predictive methods (including regression and classification) were the most frequent methods used by 32% of the literature analysed, followed by relationship mining with 24% of the studies and distillation of data for human judgment including statistics and visualizations that help people make sense of their findings.

Another crucial issue of LA, that is faced by a few learning analytics papers, is if the analysis held is ethic or not, and how to deal with data protection.

Slade and Prinsloo (2013), for example, recognized ethical contemplations in LA and proposed six standards to manage higher instructive organizations related to these issues. Prinsloo and Slade (2016) investigated undergraduates' defencelessness in connection to cultivating the capability of LA. Rubel and Jones (2016) contended that LA presented good and strategic issues for understudies' security.

2.5. Predictive modelling for learning analytics

Predictive modelling is one of the most prominent methodological approaches in educational data mining.

Firstly, there is a need to understand the type of data considered to predict.

In applied statistics, there are usually four types of variables: categorical, ordinal, interval, and ratio.

Ordinal variables are often treated as categorical, interval and ratio are considered as numeric. Categorical values may be binary, for example having or not a characteristic, or multivalued, such as predicting the class of average grades a student might have. Two distinct classes of algorithms exist for these applications; classification algorithms are used to predict categorical values, while regression algorithms are used to predict numeric values (Lang et al.,2017).

Secondly, it is essential to distinguish predictive modelling from explanatory modelling.

In explanatory modelling, the goal is to use all available evidence to explain a given outcome. For example, age, gender, and socioeconomic status of a student might be used in a regression model to explain how they contribute to the student's achieved result.

In predictive modelling, the purpose is to create a model that will predict the values or class of new data based on observations. It is based on the assumption that a set of known data can be used to predict the new data based on observed.

Several different algorithms exist for building predictive models.

With educational data, it is common to see models built using methods such as these:

- Linear Regression predicts a continuous numeric outcome from a linear combination of variables.
- Logistic Regression predicts the odds of two or more outcomes, allowing for categorical predictions.
- Nearest Neighbours Classifiers use only the closest labelled data points in the training dataset to determine the appropriate predicted labels for new data.
- Decision Trees are repeated partitions of the data based on a series of single attribute "tests." Each test is chosen algorithmically to maximize the purity of the classifications in each partition.

- Naïve Bayes Classifiers assume the statistical independence of each attribute given the classification and provide probabilistic interpretations of classifications.
- Bayesian Networks feature manually constructed graphical models and provide probabilistic interpretations of classifications.
- Support Vector Machines use a high dimensional data projection in order to find a hyperplane of most significant separation between the various classes.
- Neural Networks are algorithms that propagate data input through a series of interconnected layers of computational nodes to produce an output.
- Ensemble Methods use a voting pool of either homogeneous or heterogeneous classifiers. Two prominent techniques are bootstrap aggregating, in which several predictive models are built from random sub-samples of the dataset, and boosting, in which successive predictive models are designed to account for the misclassifications of the prior models.

Most of these methods have tuneable parameters that change the way the algorithm works depending upon expectations of the dataset: when building decision trees, a researcher might set a minimum leaf size or maximum depth of the tree (Lang et al.,2017).

Now, the literature review will focus on some cases of study that apply these algorithms.

Regarding students' performance, Lykourentzou et al., (2009) used neural networks to accurately cluster students at early stages of a multiple-choice quiz activity.

Romero-Zaldivar et al. (2012) tracked events and analysed the gathered data with multiple regression for the estimation of the variance of performance.

The prediction of the dropout was explored by Lykourentzou et al. (2009) who applied a combination of three machine learning techniques on detailed students' profiles from an LMS environment.

Dekker et al. (2009) tried to predict students' dropout and to identify the factors of their success based on the use of different classification algorithms. They used classifiers for the prediction based on simple early data, from the first-year enrolment, and boosted the accuracy with cost-sensitive learning.

More recently, Kizilcec et al. (2013) classified learners according to their interactions with course content in learning activities in MOOCs: they clustered engagement patterns, and they compared clusters based on learners' characteristics and behaviour.

Finally, Guo (2010) used neural network techniques for prediction of students' dropout, examining the number of students enrolled in each course and the distinction in their final grades.

2.6. International students' performance' review

There are a few investigations directed for learning examination in advanced education, the vast majority of them centres around dropout of undergraduates, others on students' GPA. A niche section of these, which is currently taking plenty of considerations, is about International Students.

In this section, I will try to answer, thanks to the past literature review, to research question number 1: the difference of performance between international and other students. Furthermore, I will focus on the reason why they usually have lower performance, exploring the research question number 2 in a theoretical way.

Focusing on this topic, the definition of 'international' student, and consequently of 'home' student, can vary according to the different articles.

In the UK the distinction may be made based on fee status, so 'home' students in that sense could include other European Union (EU) students, whereas in other contexts 'international' students include non-UK EU students. In Australia 'home' students usually includes New Zealanders, and vice versa. In the UK, definitions could also be based on domicile, while many other countries distinguish based on nationality or immigration status. For many others, 'international' students in the UK are defined as 'students not domiciled in the UK' (Morrison et. all, 2006).

Some studies have found no vast differences in performance between home and international students.
For instance, Ackers (1997) found that overseas-domiciled Master of Education students at the University of Newcastle achieved pass and distinction rates equivalent to their UK domiciled peers.
Several studies have found that international students performed *better than* national students. For instance, a study at the Curtin University of Technology, Western Australia (Pauley, 1988) found that overseas-fee-paying students outperformed the general students and had lower drop-out rates. A study of overseas students at Murdoch University, Australia, found that they performed as well as, or better than, their 'home' counterparts (Williams, 1989).
A study of Singaporean students of engineering at the University of Surrey (Marshall and Chilton, 1995) found that this group was more likely to gain a good degree than their

British peers. Singaporean students reported undertaking a significantly higher amount of private study than that reported by their home counterparts.

Bie (1976) found that, overall, Norwegian students in the UK had a lower failure rate than their UK peers and comparable examination results in both annual and final examinations. Bie also commented that previous studies, which looked at the performance of international students as a single group, were of limited use, given the differences between national groups.

At postgraduate level, Wright and Cochrane's (2000) study of doctoral students at the University of Birmingham found that the submission rates of overseas domiciled students of arts and humanities subjects were better than those of UK students in these subjects.

On the other hand, some research indicates that international students perform *less well* than home students.

A Dutch study of engineering students at Delft University of Technology (Jochems et al., 1996) found that, compared with Dutch students, international students achieved about the same pass rate, but they took more attempts to pass their examinations. Makepeace and Baxter (1990) studied overseas students in the UK polytechnic sector who had failed first-year examinations and found that they usually did less well than their UK peers.

Later research had proceeded onward from straightforward examinations in execution to attempt to recognize explanations behind the distinctions. De Vita (2002) found that worldwide students on a first-year business studies program in a UK college accomplished lower marks in certain types of assessment than home students.

More recently, the gap between the international and domestic students' performance was investigated in order to find the reasons why it exists.

For example, Hanus and Fox (2015) found that self-determined students are more likely to gain higher self-efficacy and better performance, language barrier is perceived as one of the primary hurdles of the performance of international students (Turner and George, 2011). Moreover, international students may not have much confidence in asking questions or seek for clarification from their professors (Medved et al., 2013). International students may also experience culture shock in sense of isolation and acculturation (Smith & Khawaja, 2011).

Going deeper into the studies about this segment, it is noticeable some recurrences in variables selected by different studies to find out the correlations between these variables

and the performance of the students. The output is usually students' dropout, GPA or time to graduation.

Some studies correlated this performance with social integration, others with cultural shock and others with the early performance a student have in his/her carrier, and many of them also use demographic data.

2.6.1. Social integration

Focusing on the first approach, in 1975, Tinto published an interaction model of student attrition that laid the theoretical foundation for research about student attrition. Spady (1970) first suggested the application of Durkheimain's classical analysis of social factors involved in suicide (1951) to the phenomenon of student attrition. The model was further extended and refined by Tinto (1993).

This stated that students enter into higher education institutions with a variety of attributes, family and community backgrounds, educational experiences and achievements, skills and value orientations.

These background characteristics and individual attributes develop educational expectations and commitments. The critical issue of the Tinto model is the level of a student's integration into the social and academic systems of the college, which determines persistence or dropout. The higher the degree of integration of the individual into the college system, the higher will be the commitment to the specific institution and the goal of college completion leading to persistence.

Pascarella et al. (1986) and Bers and Smith (1991) in their studies of community college students observed that both social and academic integration measures were significantly related to persistence.

Tinto's postulation of parity between social and academic integration in their effects on dropout was fought by Munro (1981), who in his study found strong effects of academic integration on persistence and social integration had no significant effect. Studies which reported an insignificant effect of social integration on persistence did not take into account the compensatory nature of academic and social integration (Mannan, 2007).

Tinto (1993) stated that sufficient attention should be given to the development of group-specific models or methods to study student attrition to make the research more policy relevant.

In his study, Johnson (1996), examined the GPA and learning experiences as a measure of academic integration, and found that dropout students from arts and education got a higher GPA than the science students. Dropout science students reported higher negative learning experiences than arts and education students.

Studies by Scot et al. (1996) attempted to measure the differences of dissatisfaction as a reason for leaving university between science/technology, arts/humanities or business/law students.

Spady (1971) suggested that the effects of forces that lead to dropout during the first year will continue to have an impact on the attrition process during the ensuing year.

Mannan (2007) held an analysis of the application of Tinto's model at the University of Papua New Guinea. After having designed a factor analysis to identify the clusters of variables to consider, he found out five main groups of variables that affect students' performance related to social integration:

- Informal contact with academic staff
- Academic staff concern for students' development and learning
- Peer interaction
- Extracurricular activities
- Peer group interaction

Except for the second factor, which represented academic integration, all the others represented social integration. This study showed a strong negative relationship between academic and social integration, which indicates that less integration in the social domain of the university was compensated by higher academic integration leading to student persistence.

Similarly, less academic integration might be compensated by higher social integration influencing students to continue to study. It appears from the findings of this paper that students' integration in the academic and social system of the university leading to their persistence differs according to their subject areas of studies and their year of studies. These significant differences between different groups of students implied that general retention policy's measures would not be adequate to address the group-specific problems of student retention. Thus, higher education institutions must define retention policy measures according to the differing needs of specific groups.

Furthermore, related to Tinto's point of view, overseas students are under greater pressure from families to succeed and less competent with academic skills (New Zealand Ministry of Education, 2004).

This was confirmed by Searle and Ward (1990), cross-cultural adjustment is a function of psychological/emotional adjustment and sociocultural adaptation. The former is associated with the social support people receive, and the latter depends on cultural knowledge and cultural identity. It has been found that supportive communication practices with friends and family were useful in releasing stress (Misra et al., 2003) and therefore facilitating the cross-cultural adjustment. A preference for a mentor who is interpersonally involved in the student's life has been found among international research students compared with their home counterparts in the United States (Rose, 2005). This finding highlights the social barriers faced by many international students and the primacy of social support as a coping strategy.

Further in the analysis, a study by Marra et al. (2007) showed that students who reported higher GPAs indicated their feeling of not belonging in engineering was more of a factor in their transfer decision. It seems to indicate that the more academically competent students perceived a lack of belonging in engineering as a more critical factor in their decision to leave engineering. "Of these non-persisting students, why do the more academically successful feel like they don't belong?" asked the authors. This analysis was made with a survey in which they asked five universities' students in USA different kind of questions that vary from academic background to socioeconomic status to social integration and sense of belonging. After having collected data, they held a factor analysis.

More recently, Zhou and Zhang (2014) have studied the challenges that students have to face during their first year at a university in Canada: they found that international students have developed patterns of socialization to foster their own integration in their universities.

Many studies confirmed the issue of social integration with the domestic peers. Fozdar and Volet (2012) showed that sometimes, although students are positive about working with peers from other countries, they found the work challenging. It was confirmed in the UK, by Harrison and Peacock (2009) who found that many domestic students felt negative about working with international students. Furthermore, Moore and Hampton (2015) highlighted that many domestic students preferred to work with others from their background.

Moreover, Mittelmeier et al. (2017) found that high-performing international students often had a stronger relationship with the group they worked with, holding the assignments with more fun and help from the others. At the same time, low performing students declared to have social tensions with the other teammates. They though that their

academic experience could be improved with an increasement of their relationship with their domestic peers.
This was confirmed by other researchers who showed that lower performing students are more likely to have fewer social relationships with peers in comparison to high-performing students (Gasevic et al. 2013; Hommes et al. 2012).

2.6.2. Cultural shock

Relating to the second approach, it is possible to englobe different points of view basing on the different cultures students come from.
Culture as a concept has been defined in many ways. Examples of such definitions include: "Culture is the collective programming of the mind that distinguishes the members of one group or category of people from another" (Hofstede, 2001); "Culture is the way in which a group of people solves problems and reconciles dilemmas" (Trompenaars and Hampden-Turner, 1998); and culture has also been defined as "a learned pattern of behaviour representing shared values and beliefs within a particular group" (He et al., 2007).
This last definition is perhaps the most applicable for cultural intelligence, as the definition suggests a range of specific things a person must know and do in order to be culturally intelligent.

2.6.2.1. Cultural intelligence

Firstly, one of the significant factors influencing cultural shock is cultural intelligence. There is a need to understand what it is and how it impacts on international students' performance.
Cultural intelligence (CQ) is defined as an "individual's capability to function effectively in culturally diverse settings" (Ang and Van Dyne, 2008). Cultural Intelligence may be seen as related to other measures of "real-world" intelligence, such as social intelligence, emotional intelligence, and practical intelligence, but is unique since it describes the cultural knowledge and ability of an individual to adapt their interactions with persons

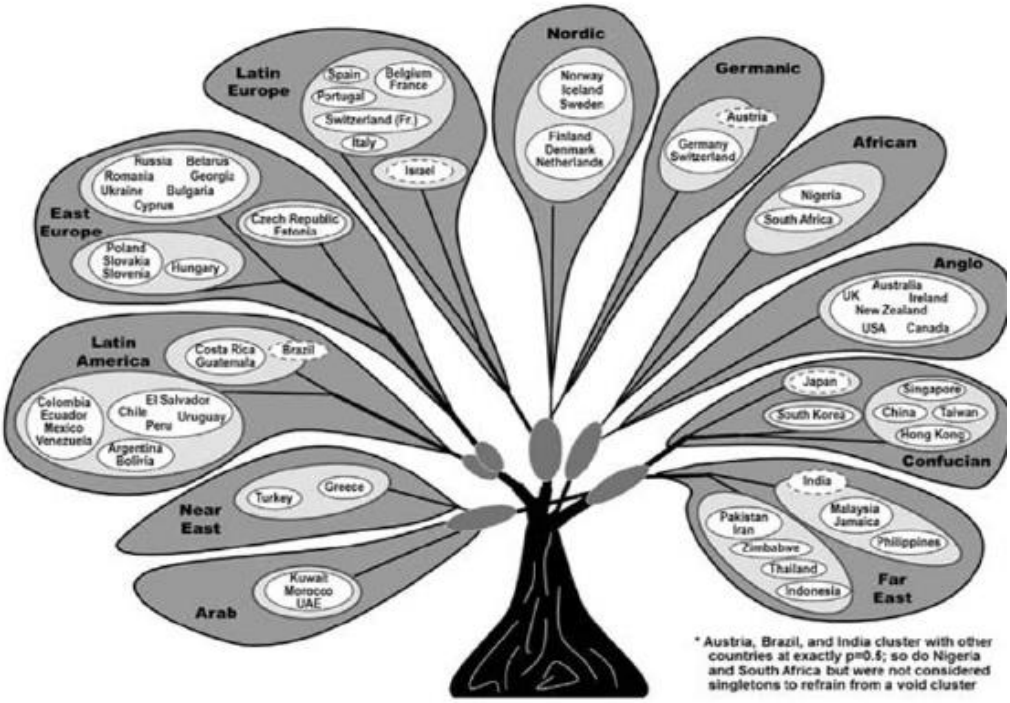
and countries of other cultures. This differentiation is crucial because it is possible for an employee or manager to perform successfully within their own culture, but to struggle to conduct themselves in culturally foreign circumstances.

Earley and Ang (2003) suggested a Cultural Intelligence Scale, consisting of Metacognitive, Cognitive, Motivational and Behavioural components, that remains one of the most widely used scales to measure cultural intelligence. Metacognitive reflects the mental processes that individuals use to acquire and understand cultural knowledge. Cognitive reflects knowledge of the norms, practices and conventions in different cultures acquired from education and personal experiences. Motivational reflects the capability to direct attention and energy toward learning and functioning in situations characterized by cultural differences. Behavioural reflects the capability to exhibit appropriate verbal and non-verbal actions when interacting with people from different cultures.

The phenomenon of mass international student education is relatively recent, and thus there is significantly less cultural intelligence research in this context. However, there has been an increase in theoretical and empirical attention directed to the study of international experiences in recent years. The intercultural adjustment has traditionally been one of the most important themes of expatriate research.

Ronen and Shenkar (2013) used a cultural mapping approach to group culturally similar countries based on multiple sets of inputs rather than only a single set of criteria such as Hofstede's (2001) dimensions. Including data from 11 studies, using different country cultural classification's approaches, they conducted a principal statistical analysis of 70 countries resulting in 11 well-defined regional cultural clusters. Their approach was based on similarity rather than dissimilarity, with the clusters being formed considering religion, language, and geography, as well as economic correlates.

Figure 2.6.1.1.1. A cluster of countries



Source: Ronen and Shenkar (2013)

They argued that such an approach potentially allowed to capture the broadest possible concept of culture, as these variables were favoured for their representation of contextual factors of countries and they were considered to be "core variables" which reflect not only culture, but the broader institutional framework, as the figure 2.6.11.1. shows.

Studies showed that cultural intelligence is affected by the previous traveling experiences a student had before living in a country where s/he will study. Social learning theory is the theory that shows that cultural intelligence is affected by international traveling.
Shaffer and Miller (2008) found the previous international experience contributes to the individual's ability to make dealing with other cultures easier. Osland and Osland (2005) saw that the previous international experiences help to develop the cultural intelligence and potentially they can favour future international experiences for the individual. They also saw previous international experiences providing the individual a history of dealing with new situations and increasing their abilities to suspend judgments until enough information is available, minimizing misunderstanding and confusion resulting from cultural differences.
For example, Crowne (2008) found work-related and educationally related international experiences to have a significant positive influence on a person's level of cultural

intelligence but found general travel or vacation experience to have no significant impact. He also concluded that extensive or longer-term exposure might better develop these abilities. On the other hand, Moon et al. (2012) found nonwork rather than work experience to be more related to cultural intelligence and concluded that "previous non-work experience is more important than work experience for developing expatriates' CQ than international work experience".

Some others have also suggested that the cultural intelligence development benefits of international travel can also occur with short-term events, as DeLoach et al. (2015) found short-term study-abroad programs are significantly and positively correlated with global awareness. Besides, Engle and Crowne (2014) found a specific type of program ranging from 7 to 12 days to have significant positive impacts on cultural intelligence development.

On the base of these studies, Engle and Nash (2016) studied a sample of students of 20 years old on average and who travel internationally mostly because of vacation travel and study abroad. They found that, given the USA identified as a part of the Anglo regional cultural cluster, and using a sample of USA citizens, individuals who have identified the country within which they have spent the most total time to be a non-Anglo country have developed a higher level of cultural intelligence than demographically similar U.S. citizens in the sample who have spent the most significant amount of time in an Anglo cluster country.

Consequently, if one of the objectives for international travel, either study abroad or general nonwork travel, is to develop cross-cultural competence, then the choice of the country appears to matter and the better choice may be to travel to a country which is outside of the individual's native country's regional cluster.

Another study in this field was held by Iskhakova (2018), basing on a survey from international students of a large university in Sidney, Australia, and having as predictor the academic average grade. She tested two main hypotheses, the first one was "does CQ correlate positively with students' academic performance?", while the second one was "does cultural exposure of students in pre-study period correlate positively with their academic performance?". Regarding the first hypothesis, what she found out is that Metacognitive and Behavioural cultural intelligence negatively correlate with academic performance, while Motivational and Cognitive do not show a significant relationship with academic performance. This is a surprising result that contradicts her hypothesis. However, potential explanations can be found in the literature.

Mannan (2007) said that students who are extensively involved in extracurricular activities might devote less time to their academic study, which eventually leads to lower study performance. Another view explored by the author is that students who demonstrated a high level of the Metacognitive facet tend to demonstrate cultural overconfidence and significantly overestimate their actual Cross-Cultural Competence. They tend to be resistant to change and not flexible regarding adjustments to a new culture: international students who are over-confident in their perception of cultural competence are not as successful as other students. Related to the second hypothesis, the study recognizes three dimensions of Cultural Exposure (Geographic, Environmental, and Inherited). Only the Geographic and Environmental dimensions of cultural exposure exhibit a positive correlation with academic performance, while the Inherited does not.
Consequently, cross-cultural experience gained through friends and relatives in other cultures might be considered a valuable source and it is highly recommended during the pre-study period.

2.6.2.2. Languages

Another essential aspect of identifying cultural differences is languages (Webb and Read,2000). For international students who do not speak in English as the first language, their proficiency in English plays a crucial role in their final performance while studying in an English-speaking learning environment.
In particular, research on the Test of English as a Foreign Language (TOEFL) and scores on the International English Language Testing System (IELTS) have shown mixed results in predicting students' performance.
Hill at al. (1999) found students' TOEFL scores to be weakly correlated with the students' academic achievement as measured by course grades and to have the weak predictive ability when they regressed the TOEFL score on GPA.
The impact of academic discipline on the relationship between TOEFL scores and GPA was confirmed by Wait and Gressel (2009), with the TOEFL score being a better predictor of academic performance in non-engineering students than for engineering students.
TOEFL scores have not been shown to predict whether students complete a master's degree program (Van Nelson at al., 2004). However, when they split the students into two

groups with either a high or a low GPA, TOEFL scores could predict these two performance categories with 75% to 84% accuracy.

In sum, relationships between TOEFL scores and students' overall GPA tend to be weak or non-existent: the strength of this relationship increases when examined by major or type of course, among other factors.

Focusing on IELTS test, some studies found weak correlations between IELTS test scores and measures of academic achievements, such as students' GPA.

Other studies found statistically significant relationships between IELTS test scores and students' academic performance.

Hill et al. (1999), in the same study as before, found that IELTS scores overall only had a weak predictive ability on academic success, but when they divided students into groups based on their overall IELTS scores to explore academic success for each group separately, the researchers discovered that for each group, GPA was higher compared to the groups below. However, only academic achievement in the highest ability was statistically significantly different from those of the other groups.

Especially in the first semester of study, Yen and Kuzma (2009) found moderate correlations between students' overall IELTS scores as well as the listening and reading sub-scores and their GPA.

Kerstjens and Nery (2000) found low correlations between academic performance and the reading and writing sub-scores. In the multiple linear regression, however, only the reading sub-score was a significant predictor of students' GPA.

This issue was faced also by Neumann et al. (2019) with a study at an English-medium university in the French-speaking province of Quebec in Canada. Here one of the authors' research question was how English proficiency and academic self-concept relate to international students' performance during their first year of study at an English-medium university.

The results from this study indicate that there is a statistically significant correlation between students' GPA and their IELTS and that this score predicts about 12% of the variance in students' GPA at the end of the first academic year. However, overall, international students' GPA was not very different from the one of non-international students at the end of their first year of study. Moreover, the students' academic self-concept also has an impact on how students perform during their first year, as both the results and the findings from the interviews indicate.

Moreover, De Vita (2000), while examining intercultural communication taking place in the classroom, understood that primarily verbal communication, can be a significant

source of misunderstandings in communicating with international students: pronunciation, for example, can severely inhibit effective communication.

It is also essential to pay attention to the pace of delivery; a rate of speech that is too slow can lead to boredom and cause attention's falls, while too fast delivery can cause frustration and disengage students. There is a need to never forget that in order to listen, process the messages caught and take notes, time is needed, especially for students for whom English is their second or third language.

Another language-related factor concerns the use of colloquialisms and idiomatic expressions, which are often foreign even to other cultures using the English language.

2.6.2.3. Culture-specific factors

Finally, there are many other culture-specific factors associated with academic behaviour and achievement.

For example, many studies have suggested that effort and hard work are emphasized in the Chinese culture (Hau and Salili, 1996) and Chinese learners attribute their performance more to their effort than to their ability. A significant number of papers focused their attention on the difference between Asian and Western students supports the view that "Asian students have difficulties in adjustment to an educational environment that was more characterized by independent learning and less instructor supervision and guidance" (Smith and Smith, 1999).

Some cultural shocks can be caused by different classroom behaviour, for example Asian students tend to be less willing to participate in group discussions and do not like to ask or answer questions.

Watkins and Biggs (1996) tested this thesis: they showed that Chinese students are more likely to adopt an in-depth approach to learning than their Westerns, it seems that classroom performance does not necessarily reflect the approach to learning.

Another issue faced by international student is that they have to cope with a range of obstacles that home students do not.

For example, it has been shown that international students, especially those traveling from the Far East to the Western world, may face culture shocks and difficulties in cross-cultural adjustment. Robertson et al. (2000) found that the most common factor of

international students were feelings of isolation from local people, homesickness, and the need for social activities.

According to Searle and Ward (1990), cross-cultural adjustment is a function of psychological/emotional adjustment and sociocultural adaptation. It is associated with the social support people receive.

Furthermore, international students may also suffer from "academic culture shock", which is, according to Gilbert (2000), a subset of culture shock and it "is a case of incongruent schemata about higher education in the students' home country and the host country".

Li at al. (2010) conducted a study to face all these topics in the School of Management at the University of Surrey. There were selected a proportion of international students, especially at the postgraduate level.

This study suggested that the Chinese students who never studied abroad before were more likely to have higher achievement in their current studies than their home classmates who had studied overseas before. It also found that Chinese students were likely to show lower proficiency in English than other students and they had a less active learning strategy. Their English explained their relatively low performance in comparison to other international students. However, no evidence showed that this learning strategy had a significant impact on their academic achievement.

2.6.3. Early factors

Finally, regarding the third approach, as Tinto (1993) and many others have noted, first-year students are the group at most considerable risk of attrition from colleges or universities. For first-year students, research indicates attrition is mainly related to prior academic performance and GPA (Willcoxson et.al, 2011). Moreover, first-year performance could affect final performance of students.

There have been decades of studies that show a consistent relationship between college academic achievement and retention, with higher performing students persisting to a greater extent than lower performing students.

First-semester GPA is preferable as a predictor to cumulative GPA after one or more years, as first-semester GPA provides an earlier alert for students who are at risk of not

graduating. The sooner students and support staff can be aware of a potential problem, the sooner interventions can be deployed.

Research on retention and persistence has examined GPA at different points in time and under different circumstances during college education. From first-year students to last year, researchers found that cumulative GPA, which in these studies includes the first two semesters, is a significant predictor of retention (Willcoxson et.all, 2011).

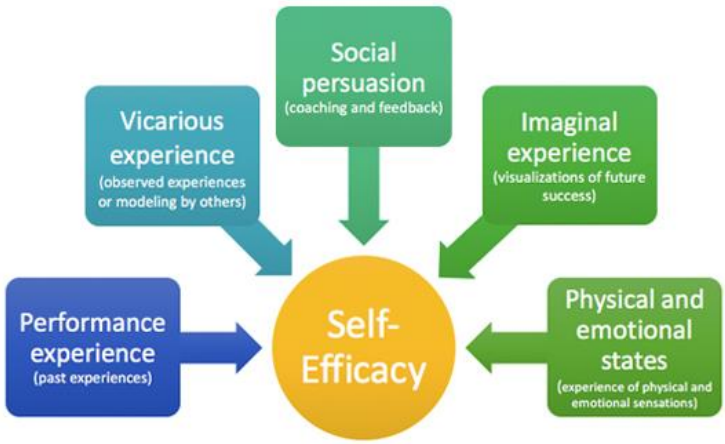
The focus of these studies on retention and persistence confirms the importance of GPA and, in some case, first-semester GPA. There are even fewer studies, however, that link the first-semester GPA to graduation.

The most relevant research in this field is a dissertation that utilizes institutional data of a university with a focus on low-income students (Yizar, 2010). Here, GPA was decomposed into categorical variables to identify the points above the academic cut-off, where students continued to be at risk of not graduating.

Studies examining the self-efficacy of first-year students and its relationship to GPA have produced somewhat mixed results. Elias and Loomis (2002) found significant correlations between self-efficacy and GPA.

In figure 2.6.3.1., the reader can find an explanation about what self- efficacy is according to Albert Bandura (2010).

Figure 2.6.3.1. Factor affecting self-efficacy



Source: Transforming Education

Zajacova, Lynch, and Espenshade (2005) found that self-efficacy was a strong predictor of GPA, even taking into account high school performance and background variables. Students' attitudes toward their college environments have also been identified as an essential factor influencing first-year student success.

Rientis et al. (2012) showed that the differences in academic and social integration between home and international students have an impact on academic performance, such as GPA and ETCS points obtained after their first year of study. This research was done among Dutch and international students using a dataset that was composed of five business schools of first-year bachelor students. GPA was positively correlated with academic adjustment, personal emotional adjustment, attachment and the perception of the faculty, while the ETCS obtained after one year correlated with academic adjustment and attachment. They found that Western students obtained a higher GPA and a higher number of ECTS than Mixed-Western, Dutch and non-Western students. GPA and ECTS scores of mixed-Western students were similar to Dutch students but lower than those belonging to Western students. Non-Western students scored lower on both GPA and ECTS compared to Western students. This difference disappeared when compared non-Western with Dutch or mixed-Western students. This could be justified reasoning that Western students who studied abroad are in general one or two years older than their home peer students and their reason to study abroad is usually a more conscious choice.

Golding et al. (2006) studied the relationship between students' GPA and matriculation requirements performance in first-year courses in the Bachelor of Science and Information Technology program at the University of Technology, Jamaica. The findings point out that performance in first-year gateway courses had some level of significance in predicting students' final performance. One of their hypotheses was that performance in first year programming and computer science courses did not have an impact on student's performance. After having designed a regression model, it had been found out that the hypothesis was wrong, in fact to predict the overall GPA it was very significant the results of some particular classes taken by the students within the first year.

Also, Al-Barrak and Al-Razgan (2016) used educational data mining to predict students' final GPA based on their grades in previous courses, using data for the information technology department, at King Saud University, designing decision trees. They found a consistent result related to the findings of the previous study.

Moreover, some studies focused on the duration of the graduation of students, another important key issue to prevent and to take action on.

One of them is the one held by Carvajal et al. (2018). The aim of their study was determining and validating a predictive model for the excessive duration, in university students of the Playa Ancha University of Education. They found that the causes related to late graduations are due to: a significant number of study programs had been structured with students who have an exclusive dedication to their studies, while there are a lot of

working students; the economic situation of the families which had to sustain the cost of their education; personal factors such as gender, average primary education, university admission score, employment status and institutional factors, such as the curriculum and its administration, student behaviour in course selection, student orientation and the perception of the support received by the university. Moreover, the most important key factor that affected positively the time to graduation was the time between the degree from the university. While the ones which affected negatively the graduation time were gender, secondary education, college dependency group, semester career duration and position of thesis teacher.

2.7. Supports to international students

In this section, the reader will have a general overview of the methods applied by the different university that might allow international students to perform better. Different cases are shown according to the diversity of problems, explained in the previous sub-chapter, a student may face during his/her international career.
Firstly, it will be discussed methods to increase students' academic integration and performance, then I will discuss methods to increase students' social and cultural integration.

Regarding the first issue, there are presented a few cases about how preventing students' lower performance and increasing their academic integration.
Cambruzzi et al. (2015) held a study on the predictive power of a tracking system, MultiTrails, and they were able to forecast the dropout. This tool allowed simultaneous and longitudinal assessment of multiple variables, such as GPA, extracurricular activities, participation in online discussion forums. Teachers were alerted about the risk students. After the system was introduced, there was an 11% reduction in student dropout.
Fritz (2011) adopted a similar approach at the University of Baltimore and designed Check My Activity, that allowed students to assess their online activity relative to their classmates in real time during the semester. A study showed that 91.5% of students who used the tool at least once was about two times more likely to earn a C or above. The increase of students' grades was connected not only to their online activity, but also to their awareness of their online activity compared to their peers.
More recently, Lu et al. (2017) used learning analytics' analysis on student engagement and self-regulation parameters in an online course to identify the at-risk population. Teachers were alerted with a notification for every student at risk and could organize a face-to-face meeting if needed.
Another promising learning analytics' intervention is the Illume program (Milliron et al., 2014). It predicts student performance and risk of discontinuing by assessing student characteristics, similarly to MultiTrial. Administrators, receiving students' data, alerted at-risk students to their risk status by telephone or email and offered further support.

Regarding the second issue, the most aged thought regarding this issue was student counselling. Some researchers believed that it was one of the most essential international students' service (Johnson, 1996), but others (Schneider and Spinler, 1986) did not agree

since the use of international students' services were infrequent and most of them preferred to keep their problems to themselves or ask friends for advice because they did not trust the international students' services staff.

After an analysis held by Zhang and Dixton (2003), most of the international students suggested them that it would be beneficial if their college and department provided orientation programs for international students addressing academic and cultural differences. From there on, the importance of events, such as the welcome week and other events, took place.

Moreover, Leask (2003) in Hong Kong and Australia found that students reported that even if they were allowed to work in culturally diverse groups, they believed that their classes offered little assistance in the development of the skills required to develop their international culture in these groups. After this, a learning guide was provided for Australian students entitled "What Do I Call You? An Introduction to Chinese, Malay and Hindu Names". Thus, it was used by home students who volunteer as buddies, students from the home university who want to help international students with their bureaucratic, academic and welcoming activities, or mentors for visiting international students. An online peer-mentoring system was also established.

Conversation groups for international students were facilitated by language in order to develop their social language skills throughout the academic year. A series of cross-cultural lunches were held on each campus: international and domestic students were invited to attend a range of cross-cultural meetings facilitated by a trained counsellor.

Robertson et al. (2009), after an analysis about the prevention of international students' lower performance, found that the staff suggested solutions in order to increase their performance. The most chosen were: using pair and group work instead of whole class discussion; mixing of international and Australian students in presentation groups; encouraging participation by inviting international students to answer simple questions initially; providing adequate time for students to prepare and providing appropriate guidance before, and following, a presentation.

For increasing the participation and the understanding in class, the solutions were: providing written support material to supplement lectures; taking time to check that international students comprehend material by asking questions; ensuring that oral explanations are not hurried; educating staff about the problems likely to be experienced.

2.8. Conclusion

After having discussed about the role of big data in higher education and about the different definitions of learning analytics from the very first ones, until more recent ones, the future perspective of it and its crucial role in predicting, taking actions, and its policy and managerial implications results very clear. It will have always more a very crucial role in higher education institutions.

Having analysed many different theories and models developed in the years in order to try to predict students' performance, such as retention, GPA or their time to graduate, it is possible to assess which are the factors that affect them the most and build a new framework which includes all these variables.

There have been different results and not coherent in the years among the international students' performance, this could have been caused by the country, in which they study, by the course of study, in which they enrolled, and by many other factors.

The most recursive causes identified by the literature are social integration, cultural shock and early academic factors.

Social integration is seen by many articles to be one of the most significant factors in affecting international students' performance. The first to understand its importance in education was Tinto (1975). Some studies found that there was no correlation between social and academic integration, but many others suggested that the learning experience, as well as informal contacts with academic staff, peers interactions, teachers' readiness to interact with international students, extracurricular activities increase the social integration leading to better students' performance. Furthermore, the pression from students' family is a factor that in many studies appears to be negatively correlated with the academic performance.

Cultural shock is another key factor that affects international students' performance. By the literature it appears to be composed by cultural intelligence, which is sub-set into short traveling experiences in a foreign country which does not belong to the origin country's cluster, there are many theories of studies on this field. Many showed that it is affected by previous travel experiences, long or short, working or pleasure, some others found that to be effective this experience has to be in a country in which the origin is not the same as the students' one, for example a student coming from an anglo speaking country needs to go to a non-anglo speaking country to better exploit this advantage. Moreover, cultural

intelligence is also affected by cross-cultural relationship, which can have been developed for an inheritance, or friendships.

Cultural shock is also affected by the problem of the different languages: in international university English is the most common taught language, but people coming from different countries have different accents and different ways of speaking that drive to meet problems in communication. Many authors measured the capability of speaking in English with the official English test, such as IELTS or TOEFL and found that students with the higher marks resulted to have higher performance, some others did not find the correlations. Many found that it is necessary to decompose the test into the 4 categories (writing, listening, use of English and speaking) to really see a higher correlation between one of these categories and students' performance. Moreover, the pronunciation, the non-verbal communication, the pace of deliveries, and different classroom's behaviours are factors that affect the languages and that affect indirectly students' performance.

Finally, early academic performance and other factors, are subset into first term GPA, first term ETCS, age, scores and number of times of repeating the same exam, high school type and entry test's level. Here, it is possible to find the higher number of different conclusions, in some case ages results to be positively correlated with the performance, in others, it is negatively correlated. It appears in many studies that self-efficacy is a predictor of the GPA, but some others did not find any correlations among the two. Some showed a correlation between the score at the entrance test and the final performance. This category could range more, since it is the most measurable and so, it can vary among different nations or courses of study. That is the reason why I wanted to test this category in a real case of study later on.

These three categories, all together, could become significant predictors of performance and could help universities and policymakers to take actions in order to better them.

Finally, a digression about which are the most common used supports for helping international students in integrating, bettering their language and their performance is exposed. Here, it is clear that the large majority of universities is starting to use the MOOC and to develop welcoming programs and international group assignments in order to let the international people interact and feel more integrated.

Chapter 3. THEORETICAL MODEL

3.1. Overview

After having analysed the different points of view that the literature has had during the years regarding the performance of the international students in higher education, it is possible to derive a theoretical model that englobes all the relevant categories that should be considered, in order to predict their final performance of the students and to take actions to better them.

The focus of this framework will be on the international students of Master of Science who decide to attend their final two years in a country different from the one of their bachelors.

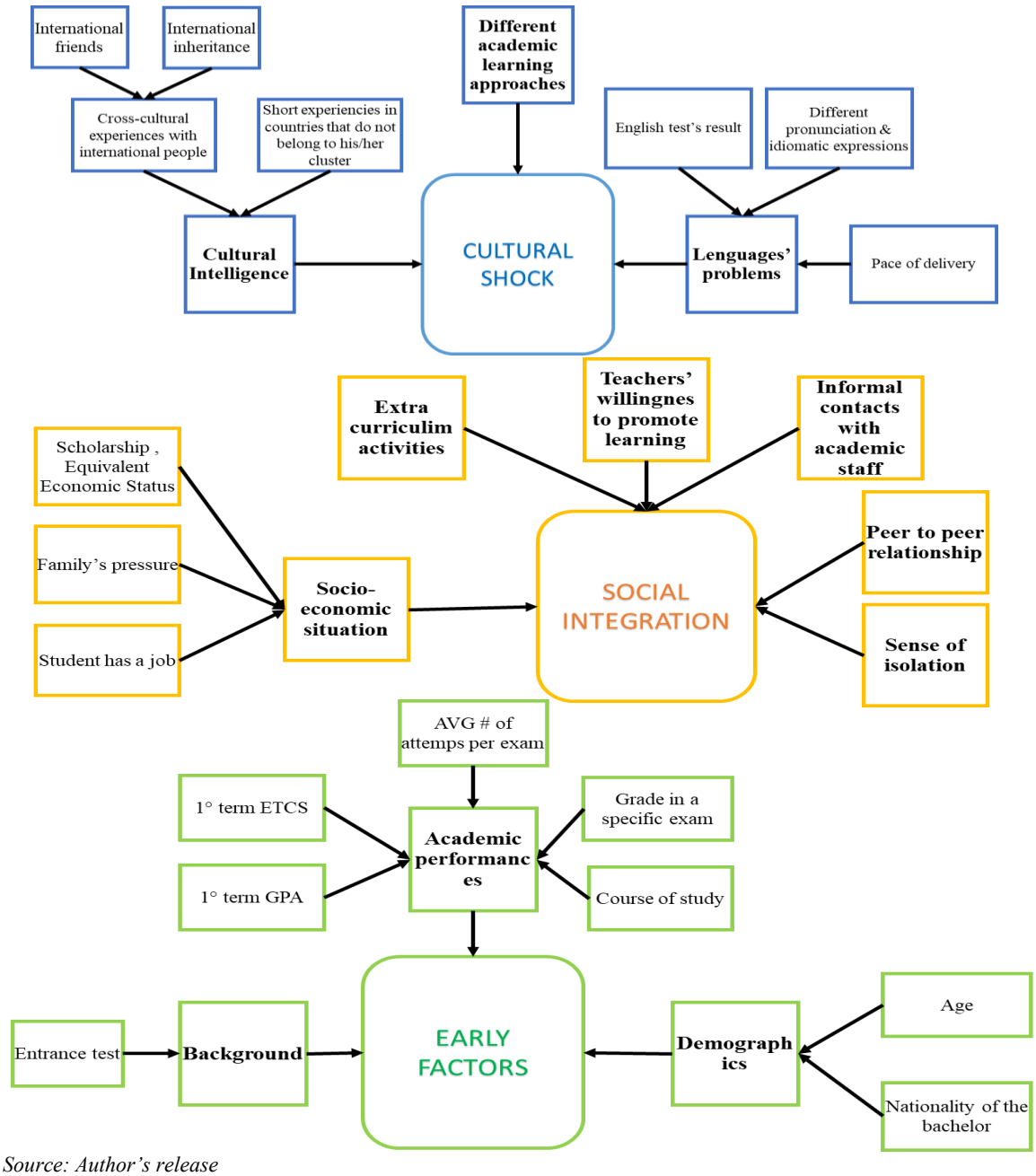
Moreover, the performance that the framework considers are GPA, dropout and time to graduation.

In this chapter, the readers can find the theoretical answer about the research question number 2: factors that influence student's performance. Moreover, in chapter 6.3., there will be a more quantitative answer to this.

3.2. Categories that affect the academic performance

As shown in Figure 3.2.1., and following the existing literature, there are three main categories that affect international students' performance (dropout, lower GPA and high time to graduation): cultural shock, early factors and finally social integration.

Figure 3.2.1. Theoretical framework



Source: Author's release

Notes: The three biggest squares are the main categories that affect students' performance (dropout, GPA and time to graduation), the little squares are the sub-factors that affect the categories, while the little rectangles are the sub sub factors that affect the subfactors. In the following section an explanation is given

Each of these categories contains other sub-factors.

Cultural shock is affected in its turn by cultural intelligence, different academic learning approaches and language's problems.

Cultural intelligence, which is the individual's capability to settle in a different culture, is affected by various and short international experiences in a country that does not belong to the country where the student comes from, such as short lasting travel for pleasure or work, or some educational experiences to better a language, and cross-cultural relationship with international people.

The last factor is composed of international friends, which means having developed a network of international relationships, and international inheritance, which is related to have in the DNA some international roots.

Language's problems are affected by English test results, such as IELTS or TOEFL, different pronunciation and idiomatic expressions, such as the misunderstanding of common sayings or general words, and the different paces of delivery the lessons. For example, a too slow pace could lead to boredom in some students, while a fast pace could lead to not understanding the lessons.

Regarding different academic learning approaches, it is known for example that for example Asian students are less active during lessons that Western ones, so that, they might be shy in participative classes in the Western world.

Early factors are affected by academic performance, background and demographic factors.

Academic performance is affected by first term GPA, first term ETCS (the number of credits obtained by a student in the first term of his/her studies), first term average number of attempts for the exams, grading in a specific exam. Moreover, they are also affected by the course of study a student decides to enrol in. With the first term, it means the first semester or first year. With first term average number of attempts, it means the average of the average number of attempts for each exam, while with grading in a specific exam, it means that, if there is an exam, harder to pass or to get a good grade than the others, student's performance are affected by it.

Background variables are affected by the type of high school/bachelor a student attended in his/her career, and by the grade in his/her entrance test at the current university.

Finally, demographic characteristics are affected by the student's age and students' nationality.

Social integration is affected by extra-curriculum activities a student does in his/her university, peer to peer relationship, if s/he has a friendship network or if s/he interacts with his/her mates, teachers' willingness to promote awareness across their students, which is their willingness to let everyone understand and get the key points of their lessons, informal contacts with the academic staff, when students want to ask questions or for clarifications to their teachers or their tutors or their mentors, and it is also affected by a sense of isolation that international students might have when moving into a new country without any friends or any landmarks.

Moreover, social integration is also affected by the socio-economic situation. This is affected by students' job, if a student is a part-time student and works for half of the time; family's pressure, either if a student is not in a great economic situation and feel the effort of his/her parents in supporting his/her studies, or if his/her parents want him/her to succeed in everything s/he does; and scholarship, if student's studies are fully or partially paid by a scholarship.

3.3. Developing hypotheses to be tested about the determinants of the academic performance

In each sub-chapter of this section, it is shown a different model with its hypotheses according to the different types of variables that affect the different performance. Some hypotheses have already been confirmed by the previous works; others will be confirmed by my analysis in chapter 6.3.

Focusing on the cultural shock and social integration, the performance that the models consider behave in the same way, so the models will be one for each category.

Instead, regarding the early factors, the relationships between them and the performance change depending on which performance are considered. So, there will be three different models and hypotheses according to students' GPA, students' dropout and students' time to graduation.

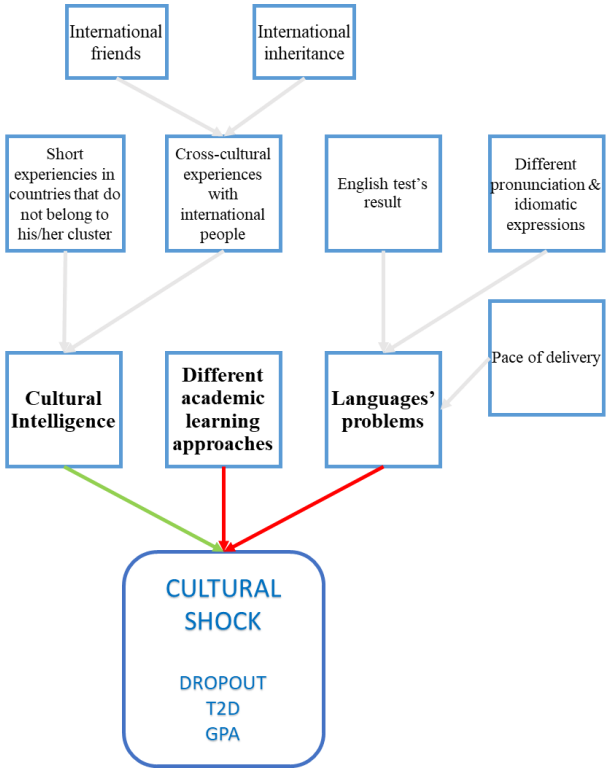
In this section, there are presented the instructions to read the models:

- Green arrows: the variable affects positively the performance
  - It increases the GPA
  - It decreases time to graduation
  - It decreases the probability of dropout
- Red arrows: the variable affects negatively the performance
  - It decreases the GPA
  - It increases time to graduation
  - It increases the probability of dropout
- Dark Grey arrows: it depends on the value of the variable if the performance is affected positively or negatively
- Light Grey arrows: the performance is affected in the way explained by the previous or following arrows

Moreover, HX, where X is a number between 0 and 25, are the hypotheses on which the framework is based on. Behind the hypotheses' statement there are explained the reasons why the variables considered are essential and the justifications, written by the literature selected, about the relationship within the variables.

3.3.1. *Effects of cultural shock's variables on student's performance*

Figure 3.3.1.1. The causal link between cultural shock's variables and students' outcomes



Source: Author's release

- H0: cultural intelligence affects positively the performance

Shannon and Begley (2008) proposed that individuals who have parents from different countries or cultures have greater opportunities to learn about different cultures and to develop a "global mindset", the combination of openness and awareness of diversity. This allows them to increase their academic performance. According to Eisenberg et al. (2013) various and short experiences abroad in different countries affect Cultural Intelligence, they noted that international experience is positively related to students' cultural intelligence at the time of starting their study. Authors found out that cultural intelligence affects positively the performance of the students. Moreover, Iskhakova (2018) found that having culturally exposed friends and relatives helps students in their academic performance.

- H1: different academic learning approach affects negatively the performance

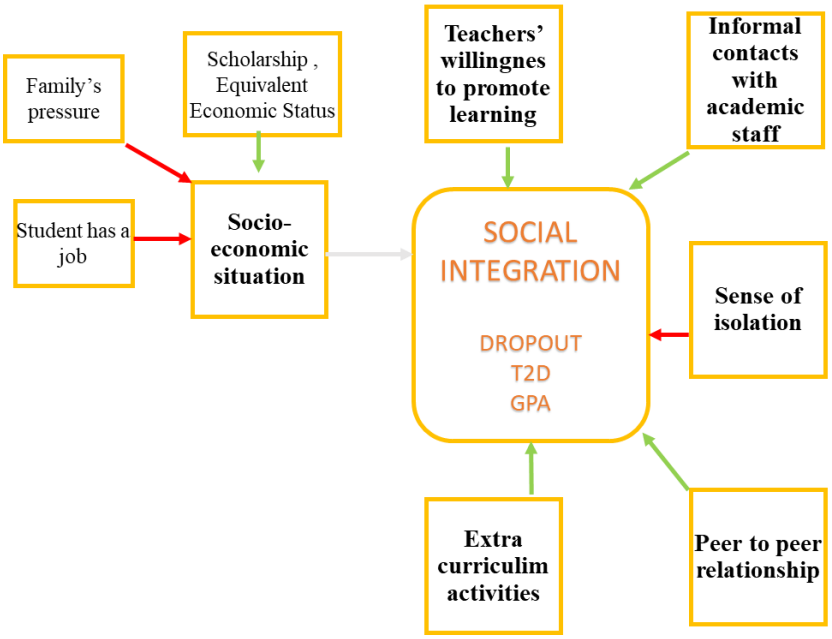
There is a great amount of literature about the difference between Asian culture and Western in their learning approach. Chinese students feel uncomfortable, at least at first, with participatory classroom activities (Pun, 1990). Furthermore, the author found that for some international students, interactive lectures, participatory-based classes and group-work may represent a new way of learning as previous education experiences in their home country. This, according to the article, is very dangerous for their academic results. Moreover, Chalmers and Volet (1997), for example, reported that Southeast Asian students hold different beliefs from many Western students about the appropriateness of speaking out in class, leading to a decreasing of performance.

- H2: language's problems affect negatively the performance

Regarding the languages' difficulties faced by international students in studying in a different language which is not their mother tongue, Van Nelson et al. (2004) split the students into two groups with either a high or a low GPA: TOEFL scores could predict these two categories with an accuracy higher than 75%. Moreover, Bridgeman et al. (2015) discovered complex relationships between students' overall and sub-scores on the TOEFL and their GPA. Li et al. (2010) found that English language proficiency is predictive of international students' academic achievement. Also, Neumann et al. (2019) found a statistically significant correlation between students' GPA and their IELTS scores, these scores predicted about 12% of the variance in students' GPA at the end of the first academic year.

3.3.2. *Effects of social integration's variables on student's performance*

Figure 3.3.2.1. the causal link between social integration's variables and students' outcomes



Source: Author's release

- H4: extra curriculum activities affect positively the performance

Tinto (1993) stated that having extra curriculum activities allows students to establish a social life and so, to have higher academic performance, since this allows them to feel more socially integrated and, consequently, more academic integrated. Later, Pascarella and Terenzini (2005) stated that individual effort and engagement are critical determinants of the impact of academic performance in colleges, it is crucial to focus on how an institution can shape its extracurricular to encourage student engagement. Belonging to a study group, students' fraternity or practicing sports can influence social integration and increase academic performance (Russell, 2010).

- H5: peer to peer relationship affect positively the performance

Regarding this relationship, Furnham and Alibhai (1985) stated that having a sufficient number of friends from the same culture as well as host-culture is important to increase the social integration and consequently, to increase students' performance. More recently, Wilcox et al. (2005) found that social networks of students have a positive influence on

the study-success of first-year students. Having friends, sharing accommodation with other students increase their performance. Later, Meeuwisse et al. (2010) found that students who drop out of higher education often state that their social networks provided insufficient support in order to continue.

- H6: teachers' willingness to promote learning affects positively the performance

First, Burns (1991) found that stress levels were considerably higher among overseas students when compared with local students since they felt misunderstood by academic staff. Robertson at al. (2000) found that that the learning outcomes teachers have in their lessons are the most often ignored by them. He suggested them to better this point in order to increase students' performance. Wilcoxson et al. (2011) added that teaching staff approachability and ability to make courses interesting and challenging contribute significantly to the likelihood of first-year student attrition or retention.

- H7: informal contacts with academic staff affect positively the performance

The topic of the relationship between the academic integration, including student development and willingness to discuss learning tasks with academic staff, and students' academic results, has been discussed by many authors such as Cox et al. (2005) and Hoffman et al. (2002). These authors concluded both that informal contacts from the students with the academic staff influenced positively the performance of the students. Moreover, this was also confirmed by Willcoxson (2011), who found a positive correlation between the two variables: the perceived support for learning and a lower probability of retention.

- H8: scholarships and equivalent economic status affect positively the performance

Cope (1968) discovered that the lack of money is a socially acceptable reason to discontinue attending school, regardless of the actual financial position. Some students who dropped out were probably getting less money from home than other students. Later, Robertson et al. (2000) understood that these problems were frequently worsened by the monetary pressures faced by students, particularly those on scholarships or without independent support. In more recent years, Fleming at al. (2006) showed that maintaining

financial aid has an impact on a student's persistence in engineering: students with sufficient financial assistance who did not have to work, were more likely to persist.

- H9: students' job affects negatively the performance

According to Coleman's (1961), the time spent on work may reduce the time spent studying, leading the student to have lower performance. Crawford and Wang (2014) also noted the existence of the self-selection issue and showed that sandwich students, university students whose courses include a one-year placement within a relevant industry, tended to be higher achievers than full-time students before placements. On the other hand, Tessema et al. (2012), thanks to some T-tests results, showed that student's employment impacts GPA positively, when students work fewer than 10 hours but when they work for more than 11 hours a week, GPAs were found to decline for each additional category of work.

- H10: family's pressure affects negatively the performance

Burns (1991) found out that more pressure from their families to succeed leads to worse performance. Robertson et al. (2000) added that overseas students were under greater pressure from families to succeed and less competent with academic skills. Later, Li et al. (2010) found that the perceived significance of learning success to family influenced negatively student's performance.

- H11: the sense of isolation affects negatively the performance

Robertson et al. (2000) found that the most common international students' feeling was a sense of isolation from local Australian classmates and homesickness. Severiens and Wolff (2008) added that students who feel at home are more likely to graduate. In a recent study conducted among international students in Australia, Russell et al. (2010) found that the 41% of international students who had low performance, experienced substantial levels of stress, which are often a result of homesickness.

3.3.3. *Effects of early factors' variables on student's performance*

In this subchapter the early factors are explained. These hypotheses are the ones that I will test in chapter 6.3., except the ones related to the high grade in a specific exam and the high score in the entrance test. The data I analysed did not give me any information about the entrance test. Moreover, I had the marks relative to all the students enrolled in the Master of Science in the university analysed, but analysing all the courses of study together, there were no standard exams in them. Moreover, there would have been hard training the model for every course of study, because it would have been several observations too small to design a robust model.

The effects of H12 and H13 are the same in all three models, and they are presented below. As stated before the three different models are regarding the following performance: GPA, time to graduation and dropout.

- H12: high score in entrance test affects positively the performance

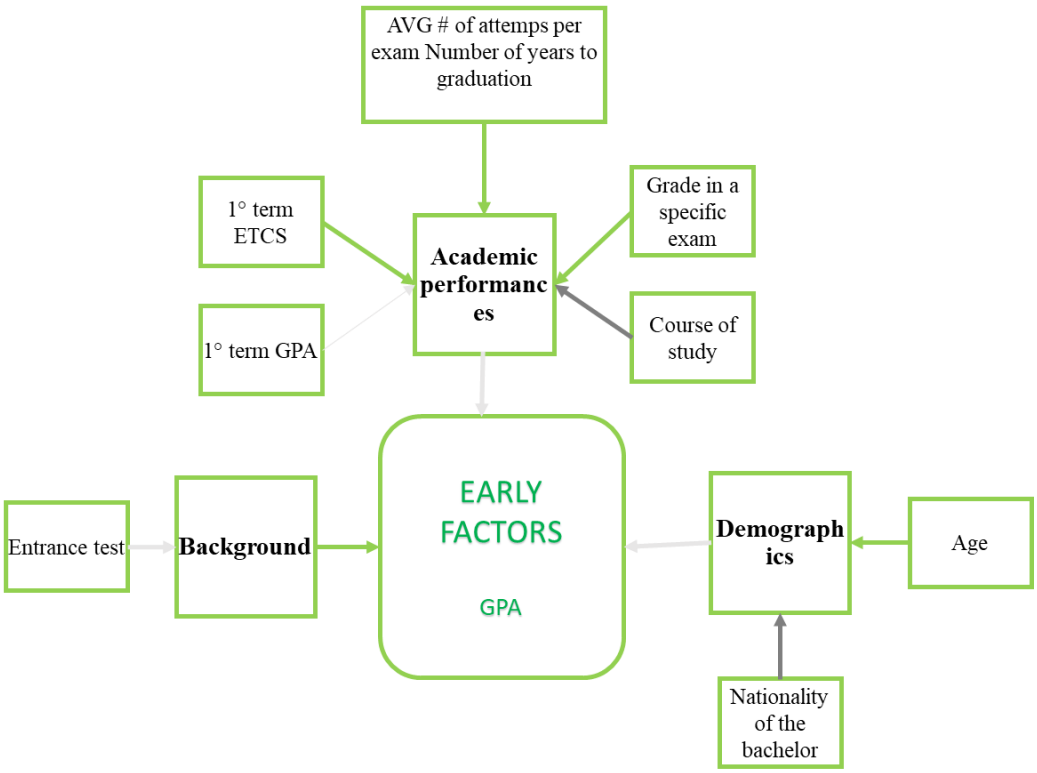
There have been plenty of discussions on this topic. Rantanen (2001) found that performance in the entrance exam predicted student performance only in the field of engineering. While Häkkinen (2004) showed that students with a high rank in the entrance exam have more study credits after four years. Later on, Araque et al. (2007) found that approximately, the risk of abandoning is two times more significant for those students who are admitted by a degree than for those who are admitted by passing the university entrance test.

- H13: first term ETCS affect positively the performance

There are no previous studies that access this relationship, but I will demonstrate through my research the impact and importance of these. The number of credits conquered by a student in his/her first term will affect positively all his/her performance. For sure, s/he will take less time to graduate, then her/his GPA will be higher due to having accepted marks that s/he is satisfied with, and finally having passed more exams allows to have a minor probability of dropping out.

3.3.3.1. *Effects of early factors' variables on student's GPA*

Figure 3.3.3.1.1. The causal link between early factors' variables and students' final GPA



Source: Author's release

- H14: high first term GPA affects positively the GPA

I know that these two variables are correlated and so I need to take them with much caution, that is the reason why I can not go further in their analyses. Moreover, there is no literature about this topic because it would have been statistically tough to justify.

- H15: average number of attempts per exam and the number of years to finish the degree affects positively the GPA

Regarding this issue, there is no literature available, but it is easy to understand the impact of these variables on the GPA. If a student repeats the same exams multiple time there are two cases: the first one is that s/he never passed it, the second one is since s/he did not like the result and wanted to increase it. In the second case, it is undoubtedly impacting

on the final GPA in a positive way and on the time to graduation which will increase. I will test this hypothesis in chapter 6.3.

- H16: high grade in a specific exam affects positively the GPA

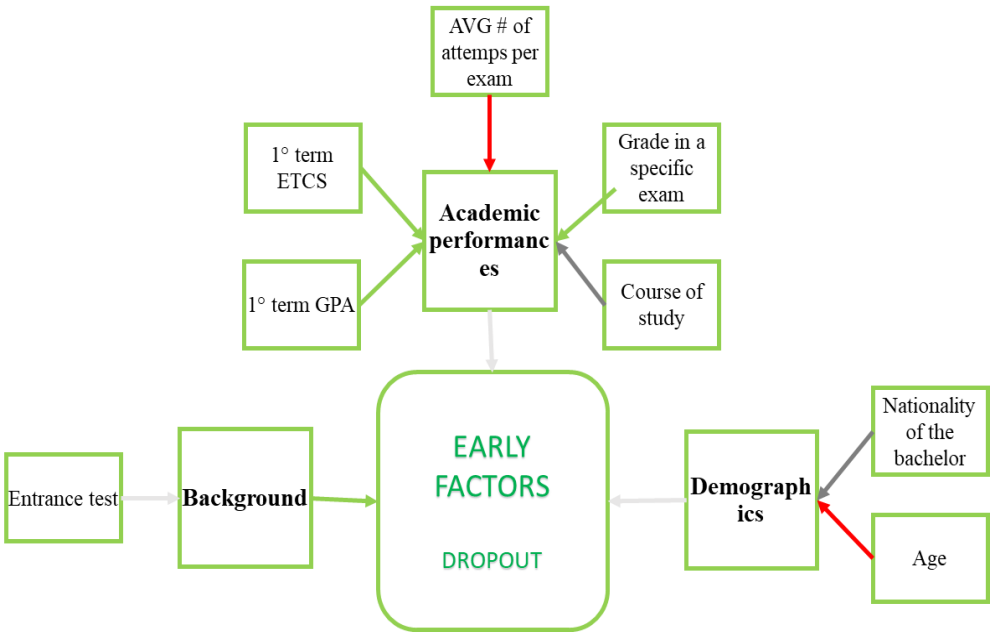
Golding (2006) assessed the importance of the gateway's classes in students' performance in computer science. Al-Barrak and Al-Razgan (2016) found with their decision tree that the root node is the "Software Engineering-1" course taken in the middle of the specialization year. This meant that this course was the most crucial course and it was closely related to students' final GPA.

- H17: student's age affects positively the GPA

Koh and Koh (1999) reported that younger students tend to show better performance in accounting. Two years later, Öckert (2001) studied the completion probability and the effects of university studies: age at entry was negatively related to student performance. More recently, Guney (2009) found that age had a significant positive impact on performance, with older students performing better than their younger counterparts.

3.3.3.2. *Effects of early factors' variables on student's dropout*

Figure 3.3.3.2.1. the causal link between early factors' variables and students' dropout



Source: Author's release

- H18: high 1st term GPA affects positively the dropout

Yizar (2010) decomposed GPA into categorical variables to identify the points above the academic cut-off, where students continued to be at risk of not graduating. Willcoxson et.al (2011) found that cumulative GPA, which in these studies includes the first two semesters, is a significant predictor of retention.

- H19: high grade in a specific exam affects positively the dropout

Fleming et al. (2006) found that performance in calculus courses, a prerequisite for engineering courses, it is believed to be the most significant obstacle for first-year students in engineering programs, correlated strongly with their retention decision.

- H20: average number of attempts per exam affects negatively the dropout

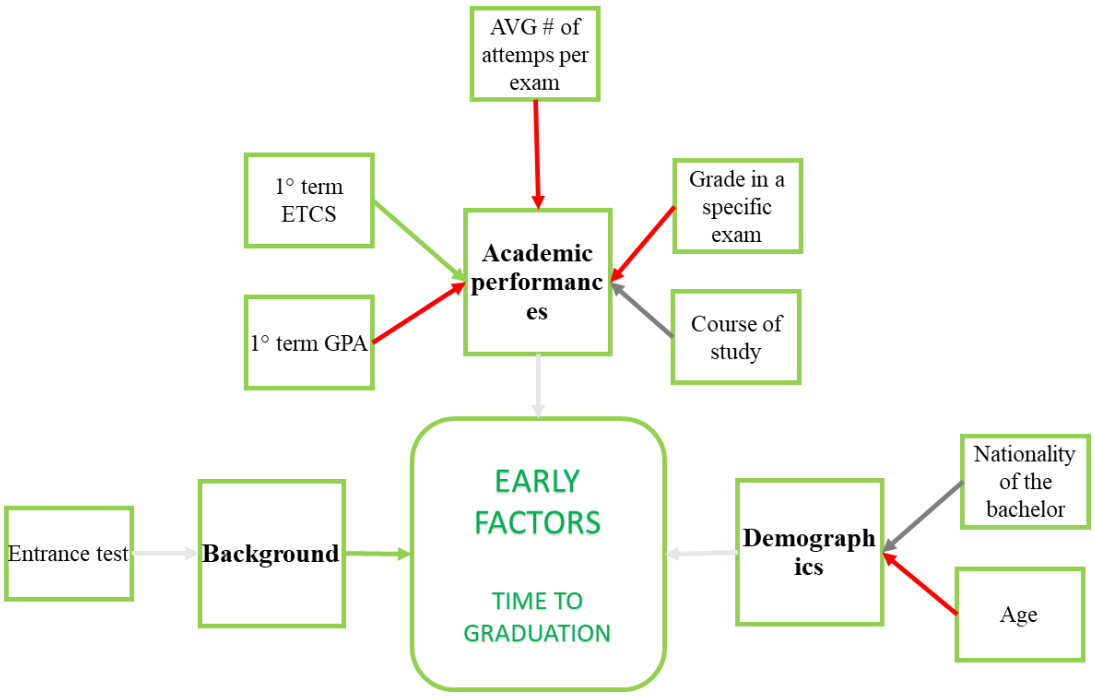
Willcoxson et al. (2011) found that first-year attrition is also associated in at least half of the universities with students' failure to understand academics' expectations, and feelings that they do not have the academic ability and analytic skills to do well in their studies. This leads them to repeat exams more than once, without succeeding in it and finally a student might drop out because of frustration.

- H21: age affects negatively the dropout

Metzger and Bean (1987) find that age and goals have a greater role in the persistence and related outcomes for non-traditional than for traditional students. There is a lot of controversy visions on this topic, there are many works which state the importance of age as a predictor of students' retention, while others affirm that it is not relevant like Goldin at al. (2006).

3.3.3.3. Effects of early factors' variables on student's time to graduation

Figure 3.3.3.3.1. The causal link between early factors' variables and students' final time to graduation



Source: Author's release

- H22: high first term GPA affects negatively the time to graduation

There is no previous evidence from the literature that assesses this topic, in my research I will find out if the model works with these two variables.

I thought it could be interesting to predict the high time to graduation, since if a student has a high GPA in the first term, when the hardest exams are supposed to be, then it means that either s/he is a hard-working student or s/he repeated the exams multiple times in order to increase the mark.

- H23: high grade in a specific exam affects negatively the time to graduation

For the same reasons of H23, I thought that this hypothesis should be relevant in predicting the time to graduation.

- H24: average number of attempts per exam affects negatively the time to graduation

Carvajal et al. (2019) found out that the semester duration impacts negatively the time to graduation, this means that if a student does not pass the exams of a particular semester, s/he has to retake it and consequently the performance gets worse.

- H25: age affects negatively the time to graduation

According to Carvajal et al. (2019) the most important key factor that affected positively the time to graduation was the time between the degree from the university, this means that age affects negatively the time to graduation: the higher the gap between the graduation and the degree, the worse the performance will be.

3.4. Conclusion

After having understood the relationship between the variables in the different models, the next section will be focused on my case of study. I will try to test the hypothesis regarding the early factors, except the ones explained in chapter 3.3.3.

Having a dataset which did not allow us to test all the parts of the model, since I did not have information about the cultural shock students faced when arriving at the university or regarding their social integration, I must trust the literature already written, keeping in mind that this could be an interesting further developed of my analysis.

The table below shows all the hypotheses divided by area of application and causal relationship with the affected performance.

Table 3.4.1. Hypotheses' table

HYPOTESIS	PERFORMANCE	AREA
Cultural intelligence affects positively the performance	DROPOUT GPA T2D	Cultural Shock
Different academic learning approach affects negatively the performance	DROPOUT GPA T2D	
Languages' problems affect negatively the performance	DROPOUT GPA T2D	
Extra curriculum activities affect positively the performance	DROPOUT GPA T2D	Social Integration
Peer to peer relationship affects positively the performance	DROPOUT GPA T2D	
Teachers' willingness to promote learning affects positively the performance	DROPOUT GPA T2D	
Informal contacts with academic staff affect positively the performance	DROPOUT GPA T2D	
Scholarships and equivalent economic status affect positively the performance	DROPOUT GPA T2D	
Students' job affects negatively the performance	DROPOUT GPA T2D	
Family's pressure affects negatively the performance	DROPOUT GPA T2D	

The sense of isolation affects negatively the performance	DROPOUT GPA T2D	
High score in entrance test affects positively the performance	DROPOUT GPA T2D	Early Factors
First term ETCS affect positively the performance	DROPOUT GPA T2D	
High first term GPA affects positively the GPA	GPA	
Average number of attempts per exam and years of taking the graduation affect positively the GPA	GPA	
High grade in a specific exam affects positively the GPA	GPA	
Student's age affects positively the GPA	GPA	
Nationality affects the GPA	GPA	
Course of study affects the GPA	GPA	
High first term GPA affects positively the dropout	DROPOUT	
High grade in a specific exam affects positively the dropout	DROPOUT	
Average number of attempts per exam affects negatively the dropout	DROPOUT	
Age affects negatively the dropout	DROPOUT	
Nationality affects the dropout	DROPOUT	
Course of study affects the dropout	DROPOUT	
High 1st term GPA affects negatively the time to graduation	T2D	
High grade in a specific exam affects negatively the time to graduation	T2D	
Average number of attempts per exams per exam affects negatively the time to graduation	T2D	
Nationality affects the time to graduation	T2D	
Course of study affects the time to graduation	T2D	
Age affects negatively the time to graduation	T2D	

Source: Author's release

Chapter 4. BACKGROUND, DATA AVAILABLE AND DATA PRE-PROCESSING

4.1. The project

This project was born thanks to a collaboration between Mathematical Engineering and Management Engineering in a technical university in Italy.

Through the analysis of the technical university's data in Italy, the aim of the project was try to find the reasons why students might have low performance, focusing on time to graduation, students' retention and GPA, and try to take actions over the ones who seem to need more help.

The data was given directly by the university its-self, and it owns data regarding bachelors and Master of Science's careers of students enrolled in engineering, architecture or design in the academic years from 2010 to 2018.

In particular, the team was composed by three professors, two of Mathematics and one of Management Engineering, one post-doctoral student from mathematics' department, one research assistant and one master thesis student, both from management engineering. The last team focused on Master of Science students, while Mathematical Engineering on Bachelors.

Regarding bachelor's students, the main objective was to understand the reasons why this university has had 30% of retention rate per year and try to make predictive analysis on the students who are more willing to leave in order to give them the proper help to make them succeed with their graduation.

Focusing on master's students, the main problem was due to the international students, as I will show later. The aims were to understand which entry barriers the university needs to put in order to filter the best students in the World in term of performance and to find out which actions the university's needs to take in order to help students to increase their outcomes.

This thesis will focus just on the university 's master of science students of engineering, after having reasoned why design and architecture do not need the same support.

Basing on demographic and academic data, I aim to prove that it is enough to look into their first-year careers in order to get insights about how good students will perform when they will graduate and if they will leave the university for a reason different from graduation.

Data about international students from 2013 to 2015 will be used, trying to find out which are the characteristics that do not allow them to have excellent outcomes, in order to take action for the following years.

This will be very important not only for this university, but also for other three main groups: potential university applicants, prospective employers of graduates, and public policymakers in government.

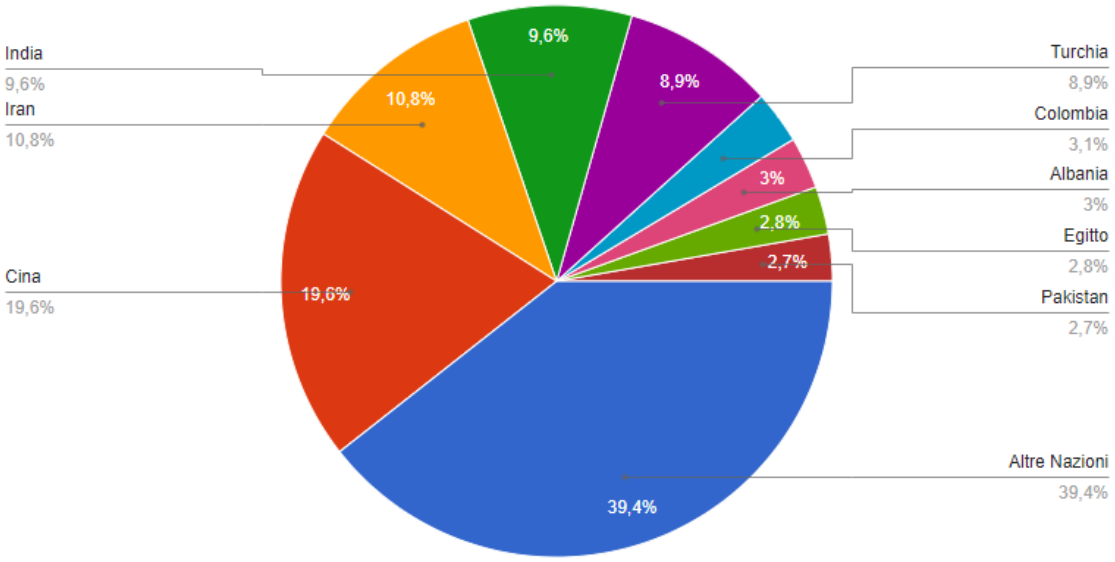
4.2. The evolution of international students in the university selected

Having understood that international students in Italy are becoming always more a hot topic from chapter 1, an analysis of data from the internal dataset of the university is provided in order to get insights about the reason why their performance are the most critical among the other group of students and why they need further attention.

Now, a focus on international university students enrolled in the university is given.

The figure below shows their nationalities in the university selected in 2016/2017. Here, the most common country of origin was China, followed by Iran and India.

Figure 4.2.1. Distribution of nationalities of international students in the selected university

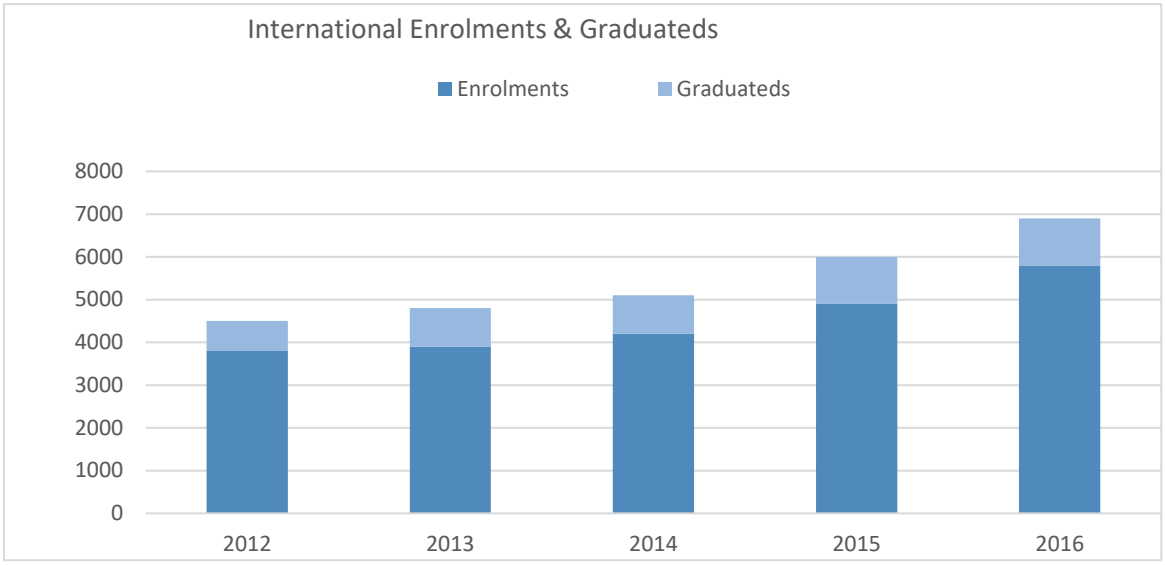


Source: Ustat

An analysis of the number of enrolments and graduation that the university had had during the years from 2012 to 2016 is provided.

The figure below shows their trend. As noticeable, it is positive: from the first year to the last one, the number of enrolments increases of about 2000 international students.

Figure 4.2.2. Number of enrolments and graduation of international students in the university selected



Source: Ustat

4.3. Data available and data pre-processing

4.3.1. Overview

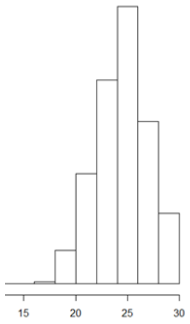
This section aims to let the reader understand the data I am dealing with.

After a clear description about the dataset I received, the data merging and the variables definitions, there will be a focus on the data preparation in which the reader will have the chance to understand the variables I create in order to hold the analysis.

Later on, data selection will be described, and after having divided the dataset according to engineering, design and architecture careers, I will focus on engineering students.

Before entering in the analysis, I have to specify the performance I aim to study, knowing that I choose to have binary performance for simplicity.

I choose to investigate the dropout's phenomenon because there was a high number of retentions belonging to the international students of the university (as the reader can understand by looking at chapter 6.2. I set a threshold of 23 of the final GPA since its mean value ranged between 24-26, as it is noticeable in the picture below, and I want to highlight the problems of those students who belong to the slice of lower GPA than the mean.

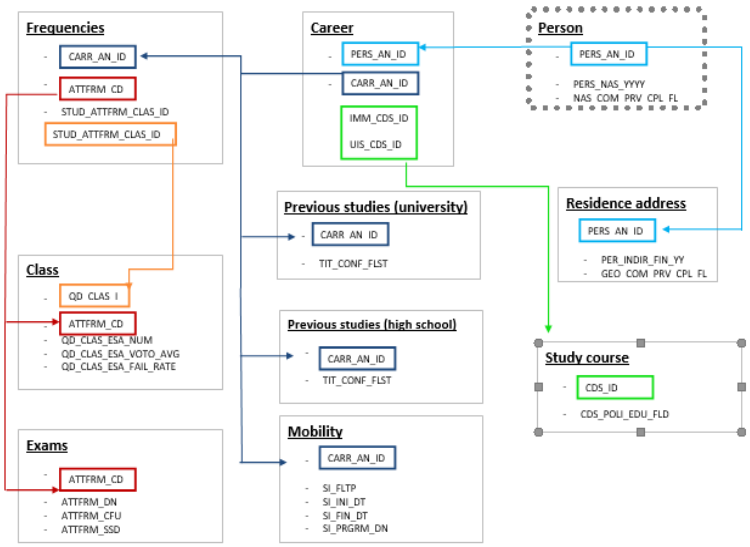


I set a threshold of years taken to graduate equal to three because the average time spent in graduation for international students, as noticeable in chapter 4.3.6., because I want to investigate and support the ones who have more troubles.

4.3.2. Raw data

In the figure, the relationships between the different datasets I have are shown.

Figure 4.3.1. Relational model



Source: Author's release

Ten datasets in CSV format were available to merge by the same key, which differs according to which dataset I am referring to. For example, Person and Career have Pers_An_ID as key, which is a number that indicates the same person, while Frequency and Class have the number of the course taken by the student which also represents the teacher who takes the class.

The most important dataset and variable given and used for the analysis where:

Person:

- PERS_AN_ID: a random number which identifies a student
- PERS_NASC_YYYY: year of birth
- NAS_COM_PRV_CPL: if s/he born in a county or not
- NAS_PRV_CD: the county where s/he born
- NAS_STT_ID: code of the country of birth
- PERS_GENERE: male or female

- PERS_CITT_STT_ID: code of the country of citizenship

Career:

- PERS_AN_ID: a random number which identifies a student
- CARR_AN_ID: a random number which identifies one career
- CARR_DETT_FLTP: type of career, if s/he is enrolled in the university or if s/he is an incoming student from another university
- IMM_CDS_ID: course of study where enrolment start
- UIS_CDS_ID: course of study where graduation happens
- STUD_AMM_VOTO: entry grade
- CARR_INI_AA: enrolment year
- CARR_INGR_AA: enrolment year at the university
- CARR_FIN_AA: graduation year
- CARR_INGR_FLTP: how s/he enters
- CARR_INI_ETA': year of enrolment
- CARR_FLST: status of the career

Frequency:

- CARR_AN_ID: a random number which identifies one career
- ATTFRM CD: number of the subject taken by the student
- STUD_ATTFRM_CLAS_ID: class and professor of the single course
- STUD_ATTFRM_FRQ_AA: year of the exam
- STUD_ATTFRM_FRQ_SEM: semester of the exam
- STUD_ACQSZ_CFU_VOTO: grade in the exam
- STUD_ACQSZ_CFU_LODE: if s/he gets a laude or not
- STUD_ACQSZ_CFU_MOD_FLTP: if the exam was oral or written
- ESA_VERB_NUM: number of times a student retakes the same exam

Class:

- QD CLASS ID: class and professor of a single course
- ATTFRM_CD: number of the subject taken by the student
- QD_CLASS_ESA_NUM: number of subscriptions for the exam
- QD_CLASS_ESA_VOTO_AVG: average grade in the class

- QD_CLASS_ESA_FAIL_RATE: percentage of failed people in the exam

Exams:

- ATTFRM_CD: number of the subject taken by the student
- ATTFRM_DN: name of the subject
- ATTFRM_CFU: subject ETCS
- ATTFRM_SSD: area of the subject

Previous study (university):

- CARR AN ID: a random number which identifies one career
- TIT_CONF_FLST: country of the title
- TIT_ATN_ID: university's code
- TIT_CDS_TIPO_CD: type of degree

Previous study (high school):

- CARR AN ID: a random number which identifies one career
- TIT_CONF_FLST: country of the title
- TIT_TP_CD: type of high school

Mobility:

- CARR AN ID: a random number which identifies one career
- SI_FLTP: type of mobility
- SI_INI_DT: starting date of the mobility
- SI_FIN_DT: ending date of the mobility
- SI_PRGRM_DN: name of mobility

Study course:

- CDS_ID: code of the course of study
- CDS_POLI_EDU_FLD: type of enrolment

After having integrated all the sub-dataset, I have variables, which included information about the student, his/her previous carrier in high school and university, his/her current career made of all the exams s/he took, the class the exams belong to and his/her mobility during the current carrier.

4.3.3. *Hypotheses' verification*

The table below shows which hypotheses will be tested with my model, and which ones are will not. This table helped the analysis in understanding which variables I need to create and which kind of data preparation I need to design in order to test them.

Table 4.3.1. Hypotheses' verification

HYPOTESIS	PERFORMANCE	AREA	VERIFIABLE
Cultural intelligence affects positively the performance	DROPOUT GPA<23 T2D>3	Cultural Shock	X
Different academic learning approach affects negatively the performance	DROPOUT GPA<23 T2D>3		X
Languages' problems affect negatively the performance	DROPOUT GPA<23 T2D>3		X
Extra curriculum activities affect positively the performance	DROPOUT GPA<23 T2D>3	Social Integration	X
Peer to peer relationship affects positively the performance	DROPOUT GPA<23 T2D>3		X
Teachers' willingness to promote learning affects positively the performance	DROPOUT GPA<23 T2D>3		X
Informal contacts with academic staff affect positively the performance	DROPOUT GPA<23 T2D>3		X
Scholarships and equivalent economic status affect positively the performance	DROPOUT GPA<23 T2D>3		X
Students' job affects negatively the performance	DROPOUT GPA<23 T2D>3		X
Family's pressure affects negatively the performance	DROPOUT GPA<23 T2D>3		X
The sense of isolation affects negatively the performance	DROPOUT GPA<23 T2D>3		X
High score in entrance test affects positively the performance	DROPOUT GPA<23 T2D>3	Early Factors	X
First term ETCS affect positively the performance	DROPOUT GPA<23 T2D>3		√
High first term GPA affects positively the GPA	GPA<23		X

Average number of attempts per exam and years of taking the graduation affect positively the GPA	GPA<23		√
High grade in a specific exam affects positively the GPA	GPA<23		X
Student's age affects positively the GPA	GPA<23		√
Nationality affects the GPA	GPA<23		√
Course of study affects the GPA	GPA<23		√
High first term GPA affects positively the dropout	DROPOUT		√
High grade in a specific exam affects positively the dropout	DROPOUT		X
Average number of attempts per exam affects negatively the dropout	DROPOUT		√
Age affects negatively the dropout	DROPOUT		√
Nationality affects the dropout	DROPOUT		√
Course of study affects the dropout	DROPOUT		√
High 1st term GPA affects negatively the time to graduation	T2D>3		√
High grade in a specific exam affects negatively the time to graduation	T2D>3		X
Average number of attempts per exams per exam affects negatively the time to graduation	T2D>3		√
Nationality affects the time to graduation	T2D> 3		√
Course of study affects the time to graduation	T2D>3		√
Age affects negatively the time to graduation	T2D>3		√

Source: Author's release

Notes: X means: not possible to be verified, √ means: verifiable with my data

4.3.4. Data preparation

Firstly, I count the number of years that a student took to graduate, deleting those who were still active. The formula used was (year of graduation – year of enrolment + 1), with the year of graduation and year of enrolment I mean the academic year. So that, if a

student started his/her carrier in September 2015, then it will record 2015, because it is the academic year 2015/2016, if it ended in July 2017, its academic year would be 2016/2017, so it will report 2016. For this reason, I add a "+1" in the formula.

Then I compute the average grade a student took in his carrier and his/her average grade in the first semester and the first of the carrier with the following code, not considering the exams unpassed (the code provided is relative just to the final GPA).

```
avg.evals = ddply(passed.exams, .(StudentID),
                  function(x) data.frame(WeiAvgEval = weighted.mean(x$Score[x$Score != 0], x$NumberECTS[x$Score != 0])))
careers.new = merge(careers, avg.evals, by="StudentID", all.x = T)
careers.new$WeiAvgEval[which(is.na(careers.new$WeiAvgEval))] <- 0
careers <- careers.new
```

Afterward, I take the exams of each student which had the number of attempts higher than zero, so the ones who had tried to pass the exam at least once. I calculate the average number of attempts for examination per each student, both in the whole carrier and in the first semester and the first year. The code below represents the calculation of the average number of attempts when students finish their career.

```
avg.att = ddply(exams, .(StudentID),
                function(x) data.frame(AvgAttempts = mean(x$NumberAttempts)))
careers.new = merge(careers, avg.att, by="StudentID", all.x = T)
careers.new$AvgAttempts[which(is.na(careers.new$AvgAttempts))] <- 0
careers <- careers.new
```

I compute the number of CFU that each student gets on his/her first semester and the first year and in the carrier.

```
tot.cred2 = ddply(passed.exams, .(StudentID),
                  function(x) data.frame(TotalCredits1.1 = sum(x$NumberECTS)))
unione = merge(tot.cred2, enrolls, by='StudentID')
tot.cred.11 = subset(unione[,c("StudentID","TotalCredits1.1")])
careers.new = merge(careers, tot.cred.11, by="StudentID", all.x = T)
careers.new$TotalCredits1.1[which(is.na(careers.new$TotalCredits1.1))] <- 0
careers <- careers.new
```

The reason why I choose to calculate the performance initially both in the first semester and in the first year was that my analysis aims to early spot the low performance of students. Unfortunately, if I analysed only the first semester, I would not be able to find information about the first semester of those students who enrolled in the same academic

year but in the second semester. So, I choose to select only the performance in the first year.

I create three dummies variables: dropout (1- yes, 0-no), GPA higher than 23 (1-no, 0-yes), and time to degree higher than three years (1-yes, 0-no). The code below represents the three steps, respectively. These variables will later be considered as the ones measuring the performance of the students.

```
df3<-df3[-(which(df3$status=="A")),]
df3$status<-ifelse(df3$status<-"D",1,0)
```

```
df3$GPAlower23<-ifelse(df3$WeiAvgEval<23,1,0)
df3$GPAlower23<-as.character(df3$GPAlower23)
```

```
YearsToFinishDegree = df3$YearEndCareer +1 - df3$YearStartCareer
df3$HighTimetoDegree<-ifelse(df3$Status=='L',ifelse(df3$YearsToFinishDegree>3,1,0), NA)
df3$HighTimetoDegree<-as.character(df3$HighTimetoDegree)
```

4.3.5. Data selection

Among all the 135,124 careers I have in my dataset, 57,335 belonged to Master of science students, in particular I have 37,130 of engineering, 14,005 of design and 5,009 of architecture.

From now on, the analysis will be held only on the 37,130 careers in engineering.

I select the ones who did not change their course of study during the current carrier in those courses of study which had in the eight years more than 300 carriers. I get: 34,196 for engineering.

Moreover, I remove students who took double degrees, in EU or extra EU, the mobilities of those carries had more than one observation and the incoming students from other universities who spent just a limited time of their master at the university.

I select only the international students and changed their university into the top 8 nationalities for engineering, writing "others" to the students who come from universities in different countries. I get: 5,002 for engineering.

Then, I select all the students belonging to those courses of study those which have more than 20 international people enrolled, and I uniformed them into a unique way of typing. I get: 4,937 for engineering.

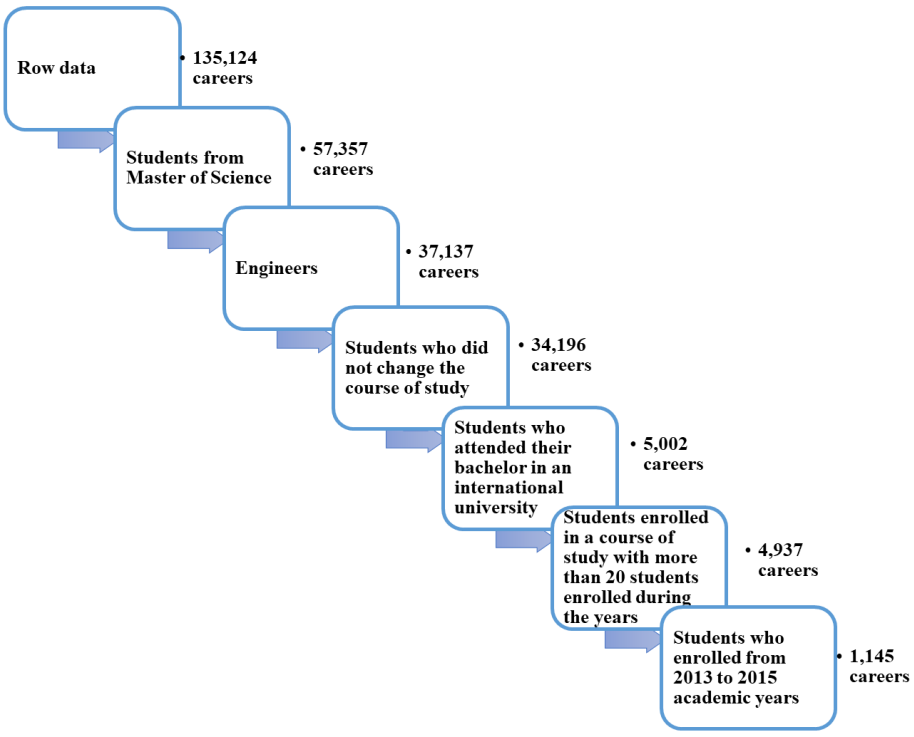
Finally, I select the years of my analysis (2013, 2014, 2015) and the carriers which were not active or suspended, obtaining a final dataset of: 1,145 students of engineering.

Three clusters of students were designed in order to understand which type of student is performing better, to answer in a more effective way to research question number 1. They were: University, students who previously attended the university in their bachelor; the ones who attended another Italian university in their bachelor, called Italians; the ones who attended an international university in their bachelor and who enrolled in the university selected for their MSc, called International.

I hold the analysis dividing the students enrolled in 2013 and 2014 academic years, as the training set, and students enrolled in 2015 as test set.

In the figure below, a representation of the data selection held is shown.

Figure 4.3.2. Data selection



Source: Author's release

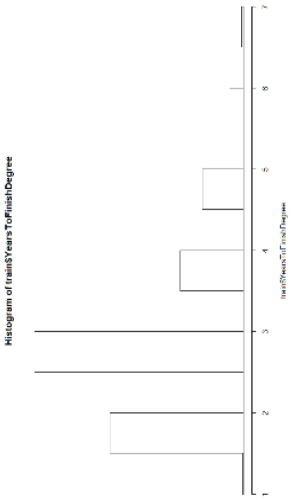
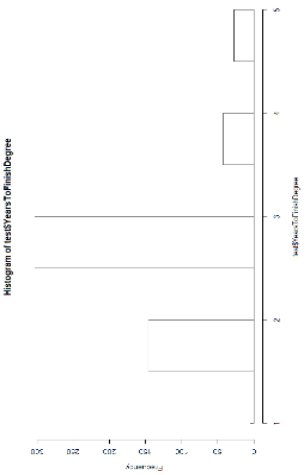
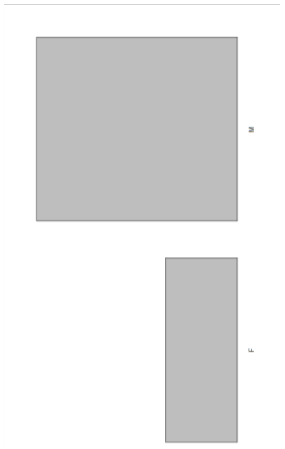
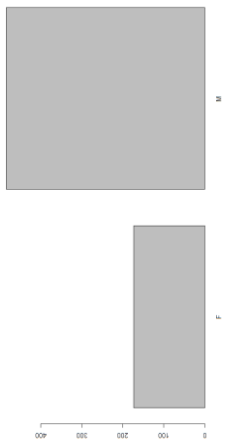
4.3.6. Attribute selection and description

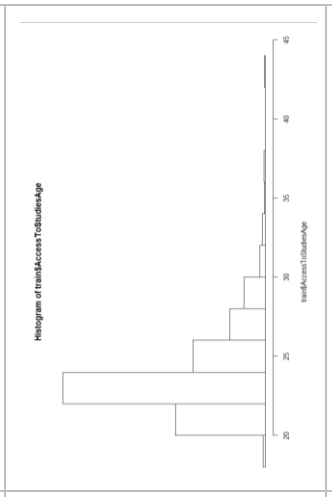
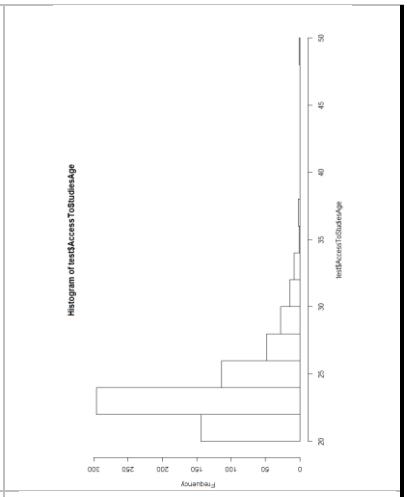
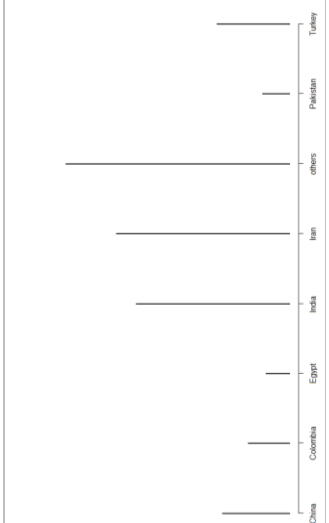
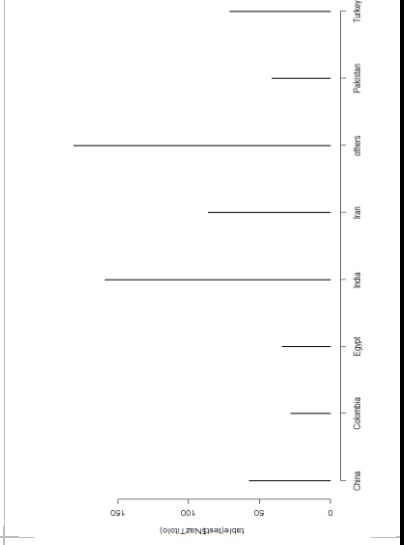
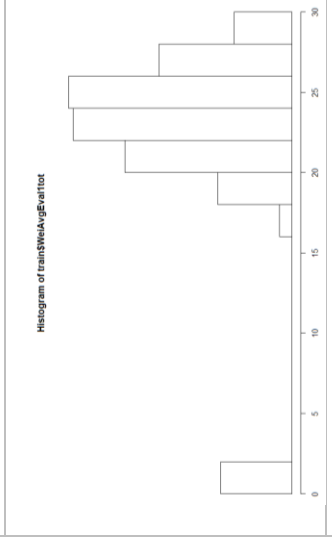
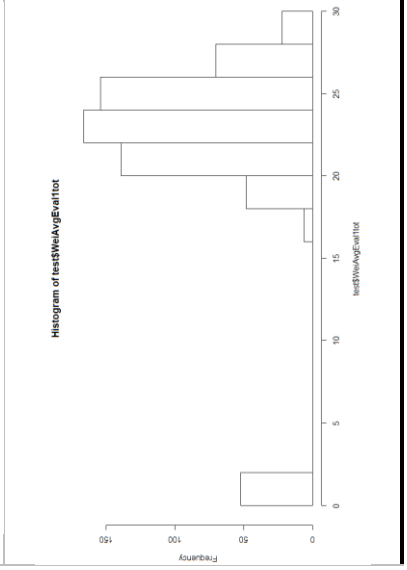
In this section, a more in-depth analysis of the attributes we used to build the machine learning algorithms is given to the reader.

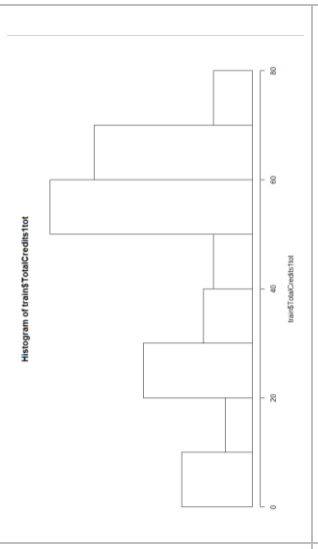
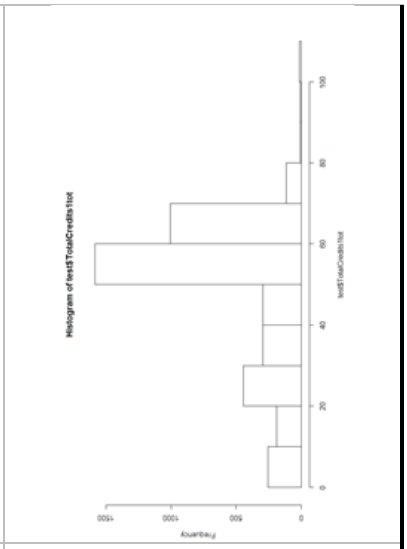
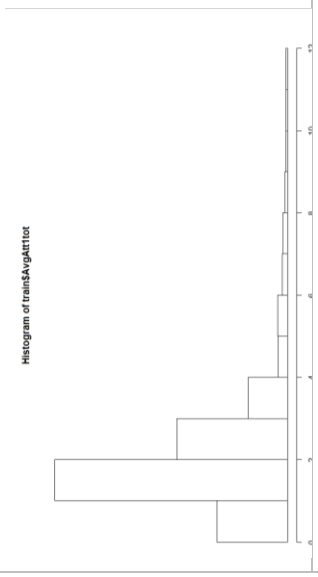
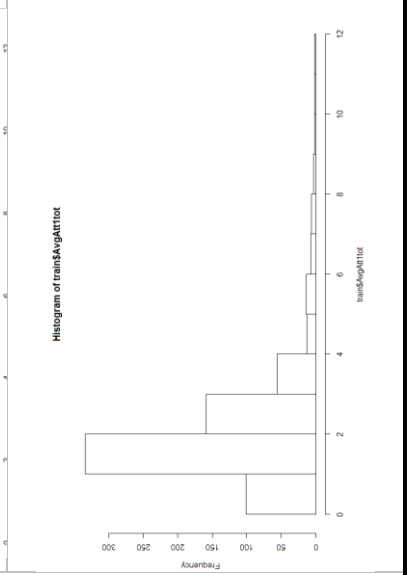
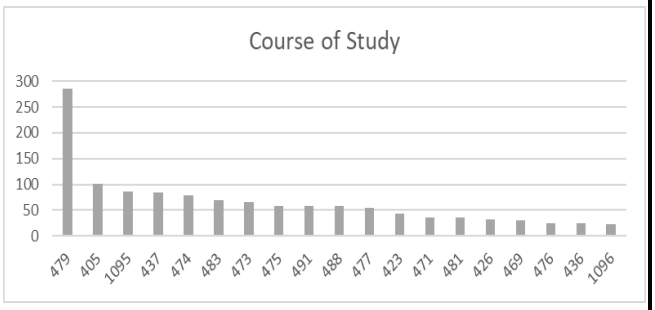
First of all, I select as training set the years 2013/2014 and 2014/2015, as a test set 2015/2016. The reason why I do not go into the previous year was that from 2013 many courses of study have been taking in English.

The following table shows the variables selected to hold the analysis for the international students enrolled in the university.

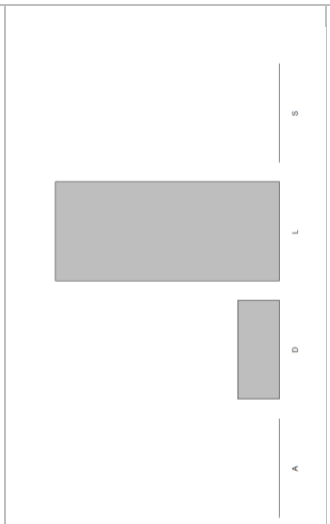
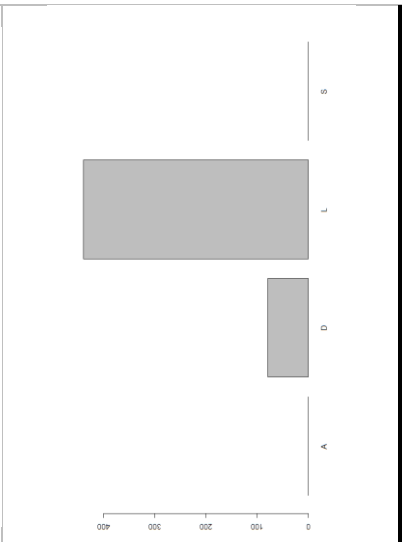
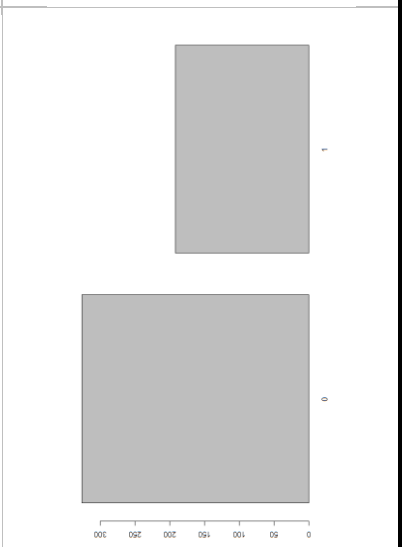
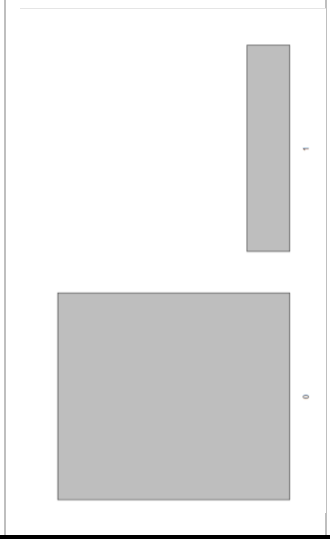
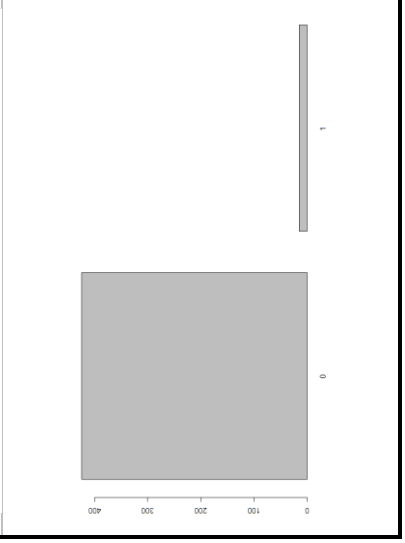
Table 4.3.2. Variables selected for the model

VARIABLE	VARIABLE DESCRIPTION	VARIABLE TYPE	VARIABLE RANGE	HISTOGRAM – BOXPLOT (TRAIN)	HISTOGRAM – BOXPLOT (TEST)
YearsToFinishDegree	# of years took by the students from the enrolment to the graduation Year of graduation – the year of enrolment + 1	numeric	1-7 (1 dropout)		
Sex	Of the student	Factor	Male / Female		

AccessToStudiesAge	Age in which a student enrolled	numeric	19 - 49		
NazTitolo	The country in which a student has taken his/her previous degree	factor	China Egypt Colombia Iran India Pakistan Turkey Others		
WeiAveEval1tot	The weighted average of the grade gets by a student, and the number of CFU per each exam obtained during the first year	Numeric	0 - 30		

TotalCredits1tot	Number of CFU a student got in the first year of his/her carrier	Numeric	0 – 50		
AvgAtt1tot	Average of the number of attempts made for each exam in the first year	Numeric	0 – 7.6		
DegreeProgramme.in	The course of study chosen by a student for his/her carrier	Management Engineering - Ingegneria Gestionale Electrical Engineering - Ingegneria Elettrica Materials Engineering And Nanotechnology Civil Engineering For Risk Mitigation Automation And Control Engineering - Ingegneria Dell' Building And Architectural Engineering Telecommunication Engineering - Ingegneria Delle Telec Mechanical Engineering - Ingegneria Meccanica Ingegneria Civile - Civil Engineering Energy Engineering - Ingegneria Energetica Biomedical Engineering - Ingegneria Biomedica Computer Science And Engineering - Ingegneria Inform. Aeronautical Engineering - Ingegneria Aeronautica Factor: 17 levels		

The courses in the cell of the left are written in descending order for the number of enrolments. Consequently, course 479 is relative to Management engineering and so on.

dropout	Students' identification when their carrier is closed for a reason different from graduation	Binary	D: student is a dropout L: a student is not a dropout		
GPAlower 23	If a student has the carrier's GPA < 23 is defined as "low performer"	Binary	1: GPA lower than 23 0: GPA higher than 23		
HighTime ToDegree	If the total carrier's time is > 3 years, the student is defined as "slow student" *	Binary	1: slow student 0: fast student		

Source: Author's release

Notes: in the columns, the reader can find the name, the explanation, the type of variables, the values it can assume and the histograms about the distribution of the variable in the train set (from 2013 to 2014) and the test set (2015)

Among the variables of the table, I describe their distributions and values in the following part, focusing on the training set and on the values that have more than 5 observations:

- YearsToFinishDegree

Value	2	3	4	5	6
#	168	257	65	45	15

As noticeable, the majority of the students took more than 3 years to graduate, with a peak on the third year from their enrolment.

- Sex

Value	M	F
#	408	144

The number of males over doubled the number of females in engineering course of study.

- AccessToStudyAge

Value	21	22	23	24	25	26	27	28	29	30	31
#	28	83	141	116	58	33	31	15	14	16	6

Here, the great majority enrols at 22-23-24 years old, but there is a significant portion of students who enrol later.

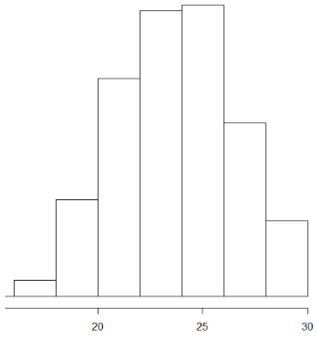
- NazTitolo

Value	China	Egypt	Colombia	Iran	India	Pakistan	Turkey	Others
#	51	17	29	124	103	18	58	152

As stated in subchapter 4.2. the most common nation from which students come from is Iran, followed by India. The percentage of nations coming from a European country is contained in the 152 belonged to others.

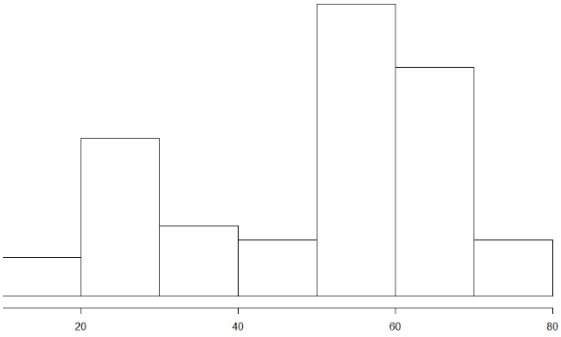
- WeiAveEval1tot

In a table it would be hard to be shown since it is a continuous variable. But according to the histogram shown in the first table and reported below, the great majority of the students seems to have taken a GPA between 24 and 26.



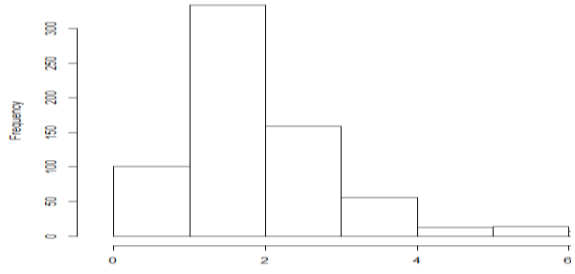
- TotalCredits1tot

In a table it would be hard to be shown since it is a continuous variable. But according to the histogram shown in the first table and reported below, the great majority of the students seems to have gained about 50 credits in the first years. There is also a higher number of students who gained 20/30 credits.



- AvgAtt1tot

In a table it would be hard to be shown since it is a continuous variable. But according to the histogram shown in the first table and reported below, the great majority of the students seems to have taken an average that ranges from 1 to 4 attempts.



- DegreeProgrammeIn

As shown in the figure of the first table, the most common chosen course of study is Management Engineering, followed by Electrical and Material engineering. They have respectively 287, 98, 85 international enrolments from academic year 2013 to 2015.

- Dropout

Value	D	L
#	82	466

As shown, the number of dropouts is not insignificant and for this reason I add to the performance to predict this variable.

- GPAlower23

Value	0	1
#	368	184

Even if the number of students who have the GPA over 23 is much higher than the others, having 184 who have GPA below 23 is significant, that is why I decide that this is going to be one variable I want to predict in the following analysis.

- HighTimeToDegree

Value	0	1
#	404	62

Here, it is crucial to remember that there are a lot of still active careers (a career is defined "active" when the student did not graduate, did not dropout, and still pay taxes and take exams). This table considers only the finished careers. Having other 86 careers which are still active, they would become either a dropout or slow students, so it is very important to predict this outcome.

The other variables were not considered for the construction of the model since sometimes they appear to be not significant among the international students.

For example, the entry grade was not available for a significant part of them; another example was that the type of high school for all of them was "estera", this means that they attended an international high school.

4.3.7. Why did I not consider International students of design and architecture?

Regarding design and architecture, the charts below show the performance of each school divided into the tree classes. The outcomes in design and architecture are almost insignificant compared to international engineering students, the analysis of these students will be held in chapter 7.2.

For this reason, I stop here with their analysis, since it is shown that they do not have critical signs of underperformance.

As stated before, the comparisons were made among International, Italians and University students, between the variables: dropout, GPA lower than 23 and time to degree higher than 3 years.

Focusing on architecture, international students' worst performance is GPA < 23 for international students, and their dropout. Overall, they are not very significant due to the lower number of enrolments and the lower percentage of difficulties.

Table 4.3.3. Comparison of the outcomes between Architecture's students

	DROPOUT			GPA<23			T2D> 3			TOTAL		
YEAR STARTCAREER	International	Italian	University	International	Italian	University	International	Italian	University	International	Italian	University
2010	24	5	37	17	3	1	34	10	100	200	59	909
2011	16	9	28	12	4	9	21	12	108	190	75	982
2012	23	6	52	17	2	12	22	9	86	244	100	956
2013	22	6	34	9	8	10	20	19	79	250	112	878
2014	15	8	18	13	5	14	25	9	42	300	100	840
2015	18	7	20	16	5	7	0	0	5	400	100	1000

Source: Author's release

Table 4.3.4. Comparison of the outcomes between Design's students

	DROPOUT			GPA<23			T2D> 3			TOTAL		
YEAR STARTCAREER	International	Italian	University	International	Italian	University	International	Italian	University	International	Italian	University
2010	3	4	11	1	1	1	12	4	30	64	67	250
2011	4	6	19	1	0	0	13	12	33	73	86	236
2012	6	3	11	0	1	5	7	11	30	64	79	250
2013	7	5	12	2	1	5	10	11	5	72	100	250
2014	15	4	16	5	2	4	12	9	13	173	90	217
2015	12	3	4	7	3	6	2	0	1	200	100	200

Source: Author's release

Moving now to design, table 4.3.4., it is evident that the worst percentage is related to the international students regarding the Time to Degree in the first years. I have also to point out that in the last years, the number of still active carriers is quite low (in 2014 they are 7, in 2015 they are 15). Consequently, this performance would not be higher than 10% overall. Dropout is minimal and also the GPA is quite high and it does not highlight any issue as it did with engineering students.

Chapter 5. STATISTICAL AND ECONOMETRIC METHODS

5.1. Overview

In this chapter the reader will have the possibility to understand the machine learning models used for this analysis.

Starting from a theoretical point of view, mixed models are discussed, focusing on the two main analysis I hold: multilevel logistic regression and multilevel classification tree.

The reason why I choose to opt for mixed model is that the dataset I am working with is a clear example of nested database: it is clear that students coming from different course of study or from different nations need to be considered separately, since I do not have enough data to build different algorithms, the multilevel algorithms seem to be the most appealing choice.

Moreover, having translated all the output variables in binary problems, the classification algorithms are preferred to the regression ones.

I choose to design a logistic regression, in comparison of neural network or support vector machines, which are usually more powerful in prediction, since my goal is also to understand the importance of the variables that affect the performance, and this is made possible by the odd ratios of the logit.

I choose to implement a classification tree since I want to let the reader to understand better and graphically the relationships between the variables that affect students' outcomes.

The practical implementation of these algorithms is shown, analysing in particular how I deal with the training set and with the test set. Moreover, the explanation of the algorithms' performance is given to the reader.

At the end, a guide about how to read the results of these algorithm is shown.

5.2. Mixed Effects

Multilevel models are often used in the literature of educational, family, developmental, and organizational psychology to analyse data in which there are sources of nesting, and for which assumptions of independence are likely to be violated. For example, students from the same school, members of the same family, and people in the same organization may be more similar in their responses than people from different schools, families, or organizations (Hoffman and Rovine, 2007).

One of the highlights of the multilevel model is the difference between fixed effects and random effects. Fixed effects are effects of variables that are known as constant among all individuals in the sample. On the other hand, random effects are effects of variables that are specified to vary among all individuals in the sample.

Different from the linear model, in which there is a just one error that has to be reduced, multilevel models are easier to understand because distinct errors are quantified at each level.

These errors are assumed to be normal distributions with mean zero and some unknown standard deviation (to be estimated from the samples).

A multilevel model can be resumed in the formula below.

$$Y_{ij} = [\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]X_{1j} + [\beta_2 + b_{2i}]X_{2j} + \epsilon_{ij}, \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, m,$$

$$b_{0i} \sim \mathcal{N}(0, \sigma_1),\ b_{1i} \sim \mathcal{N}(0, \sigma_2),\ b_{2i} \sim \mathcal{N}(0, \sigma_3),\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma),$$

5.2.1. *Logistic Mixed Models*

If the predictive variable is binary, which means it can only take 0, for not having the characteristic, or 1, for having the characteristic, a linear regression, which aims to predict values between -∞ and +∞, would not be the right choice.

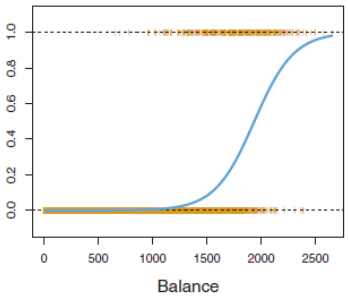
In this case, logistic regression analysis should be used.

Logistic regression gives the probability that an outcome variable gets a specific value, knowing the characteristic of one observation (e.g. the likelihood of dropping out having a GPA of 18).

The logistic function is used to predict this probability. This is an s-shaped curve: it is vertical in the middle, flatter at the beginning and at the end, figure 5.2.1.

The formula is reported below and indicates that the probability that an observation belongs to class 1 (p(X)), it is equal to the probability that the output variable gets the value of 1, knowing the values of the explanatory variables.

Figure 5.2.1. S-shape curve of logistic regression



$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Source: Author's release

Fortunately, the logit can be used to convert the s-shaped curve into a line and ease the reading of the results. So, it is possible to predict the logit of the probability that the outcome variable equals one (e.g. dropout) over the probability that it equals zero (e.g. not dropping out).

This is named as the log-odds and it corresponds to the possibility that something will happen rather than not. The formula is reported below.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Trying to understand β_1, assuming that β_1= -2 and X= GPA of a subject, this indicates that an increasement of one unit in GPA results in an expected decrease of 2 points in the log-odds of dropout, so that the probability of dropout decreases of 2 points.

To better interpret β_1, there is the need to raise it to the exponent to obtain an odds ratio, the number of times, that a characteristic increases/decreases the probability of dropping out. It is an approximation of the ratio of the dropout risks: being a male, rather than a female, increases by 1.1972 times the probability of dropout.

Odd ratio = $\frac{p(X)}{1-p(X)}$

If the odd ratio is higher than 1, this means that the variable (GPA=18) affects negatively the probability of dropout, while if it is lower than 1, this means that the variable affects positively the probability of dropout, reducing the probability. If it is almost equal to 1, this means that the variable does not affect the probability of dropout.

Now, assume that a study involves 5000 students from 20 different courses of study.

With such a data structure, it is not addressable to run a logistic regression because this goes against one of the most important assumptions in the linear model: observations need to be independent.

Students nested in the same cluster are more likely to behave in the same way than students nested in distinct clusters. There could be courses where students are more likely to dropout, maybe due to some difficult exams, and others where students do not dropout at all.

Multilevel logistic regression aims to separate the within-cluster effects from the between-cluster effects.

Multilevel logistic regression considers that the individual probability is also statistically dependent on the course of study.

In this case the log odd becomes:

$$\text{Logit}(p_i) = \log \text{ odds} = \log\left(\frac{p_i}{1-p_i}\right) = M + E_A$$

M: overall mean probability expressed on the logistic scale

E_A: course of study, its residuals are on the logistic scale and normally distributed with mean 0 and variance VA.

VA: variance of M

The equation of the log odds can be rewritten as:

$$p_1 = \frac{\exp(M + E_A)}{1 + \exp(M + E_A)}$$

Multilevel logistic regression needs strong assumptions, including:

- independent observations between clusters
- uncorrelated error terms at all levels with predictors
- no multicollinearity among predictors
- the predicting variable and predictors should be linearly correlated.
- The predicting variable should follow a Bernoulli distribution

Moreover, it is important to consider the Intraclass correlation (ICC) defined as:

$$ICC = \frac{var(u_{0j})}{var(u_{0j}) + (\pi^2/3)}$$

where $var(u_{0j})$ is the random intercept variance: the higher it is, the larger the variation of the average log-odds between clusters; $(\pi^2/3) = 3.29$ refers to the standard logistic distribution.

The ICC measured the homogeneity of the outcome within clusters, it represents the proportion of the between-cluster variation, ICC is also called VPC (variance partition coefficient) in its most general form.

5.2.2. Generalized Mixed-effects Trees

Decision trees can be used in regression and classification issues.

Tree-based methods are simple for interpretation, but they usually perform worse than logistic regression's algorithms.

Trees are methods that are based on constructing a set of decision rules on the predictor variables (Breiman et. Al, 1984). These rules are designed by recursively partitioning the observations into smaller groups with binary splits based on a single outcome. Splits for all of the predictors are examined by an exhaustive search procedure and the best split is chosen.
Its output is a tree with the branches determined by the splitting rules.

For regression trees, the selected split is the one that maximizes the homogeneity of the groups with respect to the predicting variable.
The criteria for choosing the split is the minimisation of the so-called Residual Sum of Squares, shown in the formula below. It represents the sum of difference in the real values of the predicted variables, and the values obtained after having applied the algorithm, to the power of 2. This allows to increase the power of the most significant errors and to decrease the power of the smaller ones.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

On the other hand, a classification tree is used to predict a qualitative variable rather than a quantitative and it is based on predicting that each observation belongs to the most frequent class of the training observations in the group to which it belongs.
Controversary to the decision trees, RSS cannot be used for making the binary splits.

In this case, 3 main methods are used:

- Classification error rate, it is the number of misclassified predicted observations over the number of the predicted observation;
  Where, p_{mk} represents the proportion of training observations in the region that are from the class;

$$E = 1 - \max_k(\hat{p}_{mk})$$

- Gini index, it represents a measure of total variance across the classes and it is a measure of purity, small value means that a node contains observations from a single class;

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- Entropy, a measure that takes on a small value if the node is pure and it is quite similar numerically to the Gini index.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

These methods are used to calculate the information gain in the descendent nodes, it is defined as: I(q) - $\sum \frac{Q_k}{Q} * I(q_k)$, where I(q) is one of the three methods explained before applied to the parent node (the upper one), the ratio is the percentage of the sample of the parent node (the upper one) that is placed in the corresponding descendant, and $I(q_k)$ is one of the three methods explained before applied to a descendant node a time.

As in the multilevel logistic regression, clustered data could be analysed through mixed models.

Sela and Simonoff (2012) designed a mixed-effects regression tree method for clustered data.

This tree method can appropriately deal with the possible random effects of observation-level covariates and can split observations within clusters.

Like the standard trees, this method is attractive because it provides easy to interpret models that can be graphically displayed and understood.

The model used for my project was the one designed by Fontana et. Al (2018), which intends to generalize the previous approach.

Here the fixed effect is estimated through a tree-based algorithm while the random component consists of a response variable Y from a distribution in the natural exponential family.

Theoretically, the steps to implement the algorithm are the following:

1. Initialize the estimated random effects b_i = zero.
2. Estimate the target variable probability with a logistic regression.
3. Estimate a regression tree on the probabilities of the previous step and the variables available.
4. Fit the mixed effects model, using a Bernoulli distributions' response variable and extract the random effects from the model.
5. Replace the predicted response at each terminal node of step three with the estimated population from the mixed-effects model fitted in step 4.

Each step has a different way to calculate the best model: step 2 is fitted through the maximum likelihood, it determines values for the parameters found such that it maximises the likelihood that the model produced the data that were actually observed; step 3 is fitted using the formula below, where the first term is the Residual Sum of Square, while in the second term T is the number of branches that the tree has and α can be optimized using the k-fold cross validation, it is a resampling procedure used to evaluate machine learning models and it has a unique parameter called k that refers to the number of groups data has to be split into, then for each group it takes a training and a test set, it fits the model with the train and then test and finally it summarize the skills of the model using the evaluation scores.

$$\sum_{\ell=1}^{|T|} \sum_{x_i \in R_\ell} (y_i - \hat{y}_{R_\ell})^2 + \alpha |T|.$$

5.3. Fitted Models

5.3.1. Methods

I use R programme language to code and implement the algorithms.

To train the model I use data from 2013/2014 and 2014/2015 academic years, since before these the offer of English courses of study in the university was different.

I use two classification algorithms in order to predict students' final performance: one is usually more accurate, and the other one is more visually understandable.

The first one is GLMer algorithm from "lme4" library. The GMLer function in this package can fit generic generalized linear mixed models, with different link functions included the logit. It has as optimizer the maximum likelihood, which can be expressed as an integral over the random effects space.

The second one is GMET algorithm, shown in appendix, and it can fit generalized mixed trees.

After having applied GMLER, I select which variables to insert according to the p-value: if the variable had a p-value higher than 0.05, I deleted it. This process was held for each p-value over than 0.05 in descendent order. Finally, I design the final model with the most important variables that affect students' performance and I train the model (see next sub-chapter to know how to select the best model and how to read the results).

On the other hand, the first model of GMET was already the final one, since the tree was already been pruned, with cp= 0.06 (complexity parameter: its aim is to save computing time without selecting splits that are not worthy), minimum bucket size= 15 (the minimum number of observations in any leaf node), and max depth 10 (set the maximum depth of all the nodes of the final tree, with the root node numbered as zero). Except for the minimum bucket size, I use the algorithm already implemented.

At this point, I calculate the total variance splitting it into different components due to the various clusters in the data. The intraclass correlation is equal to the percentage of the variation that is found at the higher level and it is generally called the Variance Partition Coefficient (VPC). It is constant across all individuals and its formula for binary problem is shown below.

$$VPC = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\pi^2}{3}}$$

Where σ^2_b is the variance of the model.

It measures how much variability is explained by the grouping in the clusters. In particular, a high VPC (close to 1) means that there are similarities in the groups and heterogeneity between groups, a low VPC (close to 0) means that there are individuals in the same group with different characteristics, and so it was not necessary to hold a multilevel analysis.

When it comes to predict the new performance, I use data of finished carriers of the academic year 2015/2016.

In particular, the predicted values of the models are probabilities of success and they are related to the linear predictor through the inverse-logit function.

This probability is used to build a binary classifier: if this probability is higher than p_0 then the outcomes will be 1 and if this probability is lower than p_0 then y = 0. The cut-off value p_0 is called decision threshold.
After the classification, test sample observations are labelled as TP (true positive: number of well classified observations which belongs to class 1), TN (true negative: number of well classified observations which belongs to class 0), FP (false positive: number of miss-classified observations which actually belong to class 1) or FN (false negative: number of miss-classified observations which actually belong to class 0).

Given these measures, it is possible to derive the following measurements of performance:

- Accuracy: proportion of correctly classified observation (both true positives and true negatives) among the total number of observations.

- Sensitivity (true positive rate): ratio between the true positive over the true positive observation plus the false negative ones.
- Specificity (true negative rate): ratio between the true negative over the true negative observations plus the false positive ones.

The choice of the decision threshold p_0 is made through ROC curve, it is a curve designed by plotting the sensitivity (on y axes) and the false positive rate (1 - specificity) for the distinct threshold values.

In my case, I give more importance to the false negative rate (1- sensitivity), since it measures the number of students who had lower performance, but the algorithm was not able to predict them, and so this misclassification cost could have been higher than the one related to the false positive rate, which indicated those students who did not have problems but were classified as having troubles with their performance.

Consequently, my aim is to maximise the sensitivity in order to minimize the false negative rate.

For my analysis, I consider more important to take prevention over the false negatives rather than taking later corrective actions on the false positives.

5.3.2. How to read the results of the model

Logistic regression

Figure 5.3.1. Summary of a logistic regression with mixed effect (GLMER) in my model

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: dropout ~ (1 | DegreeProgramme.in) + NazTitolo + WeiAvgEval1tot +
    TotalCredits1tot + AvgAtt1tot + Sex + AccessToStudiesAge
   Data: train
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   202.6    263.0    -87.3    174.6      538

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.7158 -0.1503 -0.0764 -0.0203 12.5279

Random effects:
 Groups              Name        Variance Std.Dev.
 DegreeProgramme.in (Intercept) 0.3117   0.5583
Number of obs: 552, groups:  DegreeProgramme.in, 13

Fixed effects:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.01061    2.41633  -0.004   0.9965
NazTitoloChina      -1.66167    1.51527  -1.097   0.2728
NazTitoloColombia   -2.96904    1.92065  -1.546   0.1221
NazTitoloEgypt     -16.40015 3977.69437  -0.004   0.9967
NazTitoloIndia       0.88451    0.60743   1.456   0.1454
NazTitoloIran       -0.10117    0.54686  -0.185   0.8532
NazTitoloPakistan  -16.96555 3468.91959  -0.005   0.9961
NazTitoloTurkey      -1.03693    1.07384  -0.966   0.3342
WeiAvgEval1tot       -0.12331    0.04810  -2.563   0.0104 *
TotalCredits1tot     -0.09935    0.01475  -6.737 1.62e-11 ***
AvgAtt1tot           -0.04420    0.14453  -0.306   0.7598
SexM                  0.09217    0.52909   0.174   0.8617
AccessToStudiesAge    0.18188    0.07782   2.337   0.0194 *
```

First quantile and third quantile have to be almost symmetric to the Median

Stars represent the most important variables: the p-value for each term tests the null hypothesis: the coefficient is equal to zero. A low p-value (< 0.05) indicates that the analyst can reject the null hypothesis.

A predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable.

Source: Author's release

The signs represent if the variable affects positively or negatively the performance in the opposite sign

In the final model estimate plus or minus 2*std. error should to be different from zero, it gives you the confidence interval

Figure 5.3.2. summary of a tree with mixed effect (GLMET) in my model for the identification of a dropout student

```
n= 552

node), split, n, deviance, yval
      * denotes terminal node

 1) root  552 44.11761000 1.485507e-01
   2) TotalCredits1tot>=23 479  4.37329800 4.977222e-02
     4) TotalCredits1tot>=33.5 392  0.40122970 2.003961e-02 *
     5) TotalCredits1tot< 33.5 87  1.63750300 1.979622e-01
      10) NazTitolo=China,Colombia,Egypt,Pakistan,Turkey 21  0.01649028 1.123543e-08 *
      11) NazTitolo=a.others,India,Iran 66  0.83863410 1.024985e-01 *
   3) TotalCredits1tot< 23 73  4.40349000 7.967001e-01
     6) TotalCredits1tot>=7.5 28  1.30305000 7.446556e-01 *
     7) TotalCredits1tot< 7.5 45  0.16097010 9.693192e-01 *
```

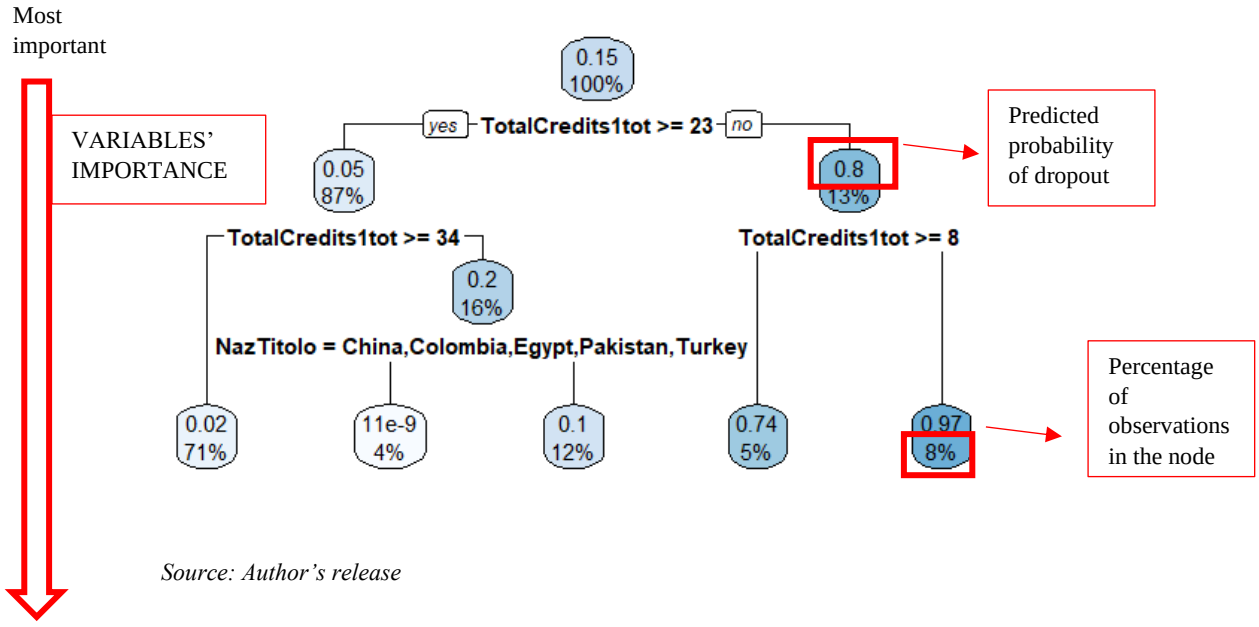
Predicted probability of dropout

Number of observations per branch

Leaf nodes

Source: Author's release

Figure 5.3.3. Plot of a tree with mixed effect (GLMET) in my model for the identification of a dropout student



Source: Author's release

Chapter 6. RESULTS

6.1. Overview

In this chapter, it is possible to receive clear answers to all the different research questions in three different way.

For answering the first research question a descriptive analysis was held to compare the performance of the international students with the performance of the others, Italians and University, in the university selected (the definition of them is given in chapter 1.4.). To better describe this, a focus on just the performance and the variables correlations of international students is provided and a cluster analysis, in order to understand better the impact of their lower performance, is shown.

To answer the second question a comparison about the four algorithms designed was provided focusing on each one most relevant variables and on their random effects.

Finally, the third question was answered, as for the second question, a comparison among the algorithms was held but, in this case, the comparison was for understanding and selecting the one that performed better in term of sensitivity.

6.2. Research question number 1: Which are the differences between international and home country students?

6.2.1. Comparison of outcomes among MSc's students in the university selected

In this section, the aim is to try to answer research question number 1 in my real case of study.

Firstly, I segment all the MSc's students into the three classes: University, Italians, and International.

I select three indicators to check the outcomes of those three categories: dropout, students' identification when their carrier is closed for a reason different from graduation; GPA inferior 23 (GPA <23), weighted average of the grades got by a student and the number of ETCS per each exam; time to degree superior 3 years (T2D>3), if a student's enrolment lasts for more than 3 academic years.

A focus on engineering's outcomes is now given.

Table 6.2.1. Comparison of engineering students' dropout, enrolments, and graduation among international, Italians, and university students.

	INTERNATIONAL		**ITALIANS**		**UNIVERSITY**	
YEAR STARTCAREER	graduated	dropout	graduated	dropout	graduated	dropout
2010	79,93%	17,92%	91,89%	7,43%	93,95%	5,53%
2011	84,55%	13,82%	95,11%	4,44%	95,71%	3,26%
2012	80,32%	16,06%	89,42%	10,58%	94,88%	3,98%
2013	83,23%	13,71%	90,76%	7,98%	94,21%	3,22%
2014	72,25%	15,25%	90,68%	4,04%	93,63%	2,48%
2015	67,12%	12,02%	79,39%	5,29%	88,82%	1,77%

Source: Author's release

As the reader can see, the percentages of dropout among the years belonging to the international students are the highest. In more recent years, it decreased because most of the students are still working on getting their master's degrees and so their careers were still active.

Moreover, as noticeable with the table 6.2.2. they also have the worst performance in term of GPA inferior to 23 and of time to degree superior to 3 years.

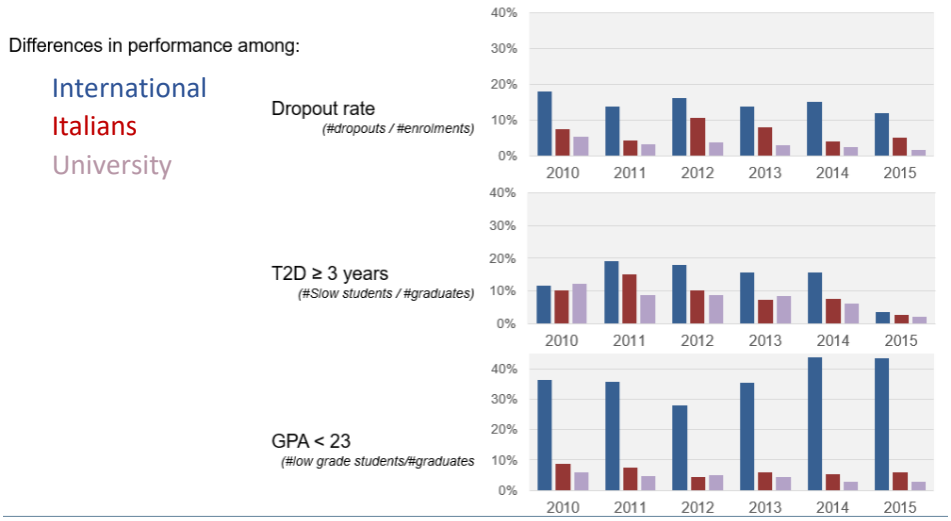
Table 6.2.2. Comparison of engineering students' outcomes, among international, Italians and university students

	INTERNATIONAL		ITALIANS		UNIVERSITY	
YEAR STARTCAREER	GPA<23	T2D>3	GPA<23	T2D>3	GPA<23	T2D>3
2010	36,32%	11,66%	8,82%	10,29%	6,01%	12,26%
2011	35,58%	19,63%	7,48%	14,95%	4,54%	8,86%
2012	28,00%	18,00%	4,30%	10,22%	5,19%	8,59%
2013	35,34%	15,66%	6,02%	7,41%	4,49%	8,53%
2014	43,60%	15,57%	5,48%	7,53%	2,88%	6,14%
2015	43,54%	3,40%	5,96%	2,81%	3,07%	1,99%

Source: Author's release

The figure below shows and confirms all their outcomes together.

Figure 6.2.3. Comparison of engineering students' outcomes during the years



Source: Author's release

Furthermore, the number of international students' active carriers for the last two years are respectively 17 and 55. This means that if all of them graduated, the percentage of students with time to degree higher than three years would increase respectively to 16% and 15%.

6.2.2. Focus on international students' performance

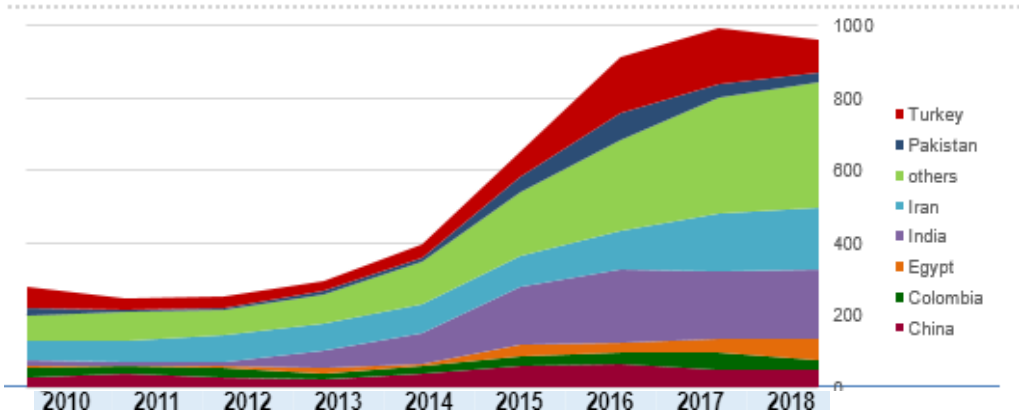
6.2.2.1. Descriptive analysis of the outcomes of the university's students

In this section, the readers will get to know an overview of the variables selected for the model and about how they interact between them-selves.

First of all, the aim is to understand the trends of enrolments along the years about the top countries for enrolments. Figure 6.2.4. shows it.

Except the "Others" which is formed by all other countries than the written ones, India, Turkey, and Iran are the most populated. Each of them is a positive trend, except for Colombia and Pakistan.

Figure 6.2.4. trends of international students by country in engineering

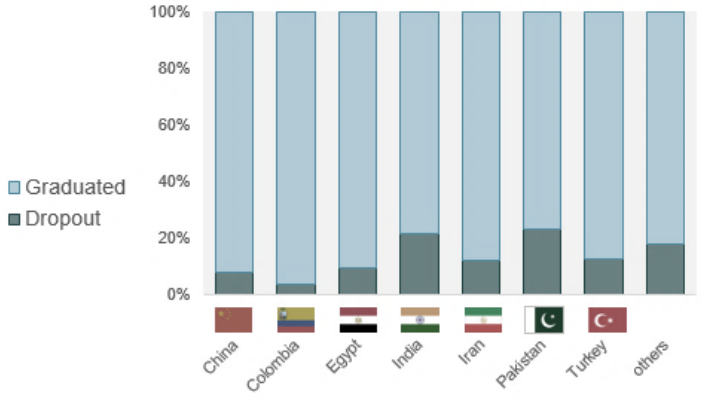


Source: Author's release

Once I have understood the positive trend of international engineering students in the university, a focus on their outcomes is needed, trying to check if there are some correlations among the variables, then I will check them in chapter 6.2. when I will develop the machine learning analysis.

Figure 6.2.5. shows their dropout rate during the years selected to train and test the model (A.A. 2013/2014, 2014/2015, 2015/2016). This highlights a higher percentage of dropout related to Indians and Pakistanis, while Colombians had the lowest percentage.

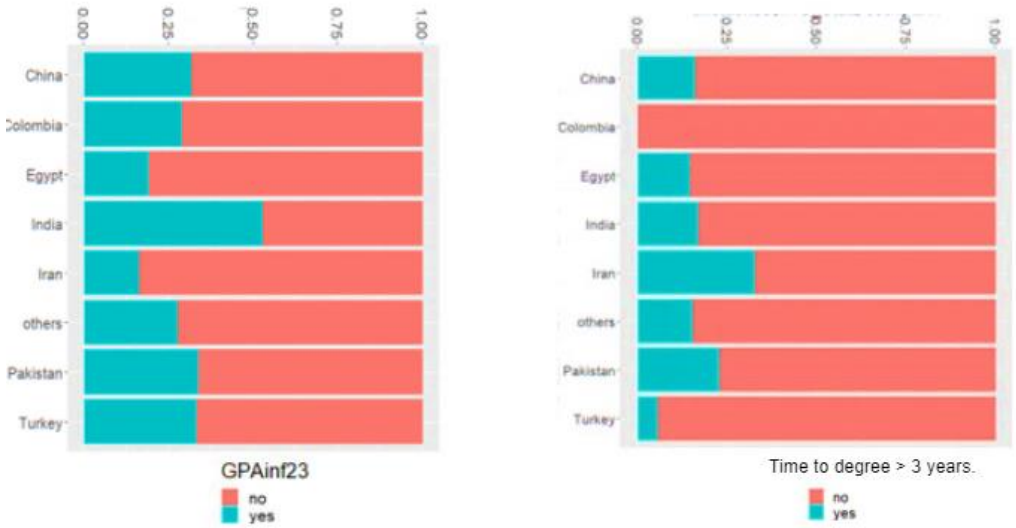
Figure 6.2.5. graduations vs dropout rate between different countries



Source: Author's release

The Figure below shows their performance in term of the other two outcomes. About the first image, it is easy to state that Indians have the lowest performance in term of GPA, while Iranians have the lowest in term of time to degree but the best in term of GPA higher than 23. Colombians perform quite well in term of time to degree.

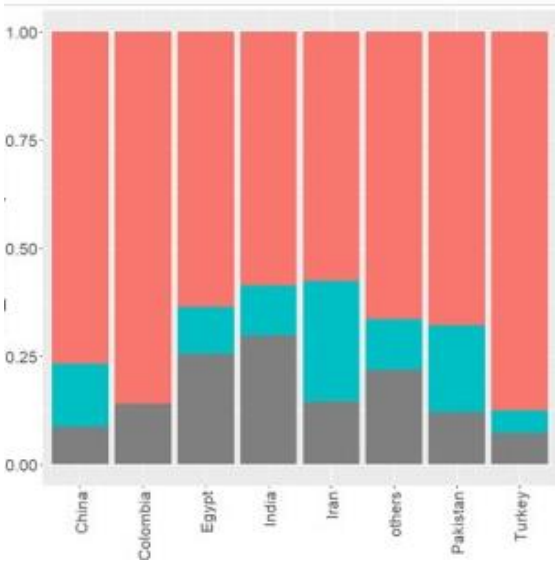
Figure 6.2.6. GPA<23 and T2D>3 years by countries



Source: Author's release

Figure 6.2.7. shows the percentage of active careers per country. Here, the red ones are the careers finished with graduation, the green ones are finished with dropout, and active careers are coloured in grey. People who studied in India and Egypt have the highest percentage of active carriers, which would enlarge the rate of time two degree higher than three years, or the dropout.

Figure 6.2.7. The ratio between active and dropout and graduated careers



Source: Author's release

Now, my aim is going into details about which universities most of these students attended before.

I select universities with more than 45 enrolments, from 2010 to 2018, in order to have a robust sample to consider.

The percentage of students who have time to degree higher than 3 years could be conditioned by the fact that many carriers were still active when I extract the data.

The percentage regarding the variable has been calculated among the number of students who attend that particular university, while the percentages "overall" have been calculating over the sum of the students who have that binary variable positive among those belonging to all the universities selected.

Table 6.2.8. Outcomes of engineering international students by biggest university

numbers are expressed in %	**GPA <23**	**OVERALL**	**DROPOUT**	**OVERALL**	**T2D>3**	**OVERALL**
INDIA						
Anna University	67,00	26,00	4,00	22,00	8,00	38,00
Jawaharlal Nehru Technological University, Hyderabad	80,00	31,00	21,00	53,00	31,00	50,00
Visvesvaraya Technological Univesity	51,00	20,00	6,00	14,00	2,00	6,00
Jawaharlal Nehru Technological University, Kakinada	60,00	23,00	9,00	11,00	5,00	6,00
IRAN						
Università Azad Islamica	35,00	35,00	6,00	32,00	22,00	56,00
University of Teheran	18,00	18,00	10,00	41,00	9,00	19,00
Università Tecnologica Amirkabir	23,00	16,00	7,00	18,00	10,00	14,00
Sharif University of Technology	25,00	25,00	4,00	9,00	10,00	11,00
PAKISTAN						
University of Engineering and Technology	18,00		10,00		-	
TURKEY						
Istanbul Teknik Universitesi	48,00	58,00	10,00	93,00	2,00	50,00
Orta Dogu Teknik Universitesi	34,00	42,00	3,00	7,00	2,00	50,00
COLOMBIA						
Universidad De Los Andes	27,00	51,00	3,00	50,00	2,00	50,00
Universidad Nacional De Colombia	26,00	49,00	4,00	50,00	2,00	50,00
CHINA						
Tongji University	19,00		2,00		4,00	

Source: Author's release

As noticeable in the chart, in India, the lowest performance university is Jawaharlal Nehru Technological University, Hyderabad.

In Iran, Università Azad Islamica is the lowest one in term of GPA and Time to Degree, while the University of Teheran is the lowest one for dropout.

Istanbul Teknik Universitesi is the lowest one in term of GPA and dropout. Colombian universities perform almost the same in all three categories.

Although these countries had a higher percentage of dropout if considered all universities, for example, Pakistan and India, looking at this chart, it seems that overall these universities perform quite good, and I have reasons to say that the most of dropout students come from smaller universities.

Analysing now a different point of view, there is a focus on the different courses of study and see for which of them this performance is the best and which is the worst.

For the percentage of dropout, the best performance is related to Automation and Control Engineering Ingegneria dell'Automazione, while the lowest is Computer Science and Engineering – Ingegneria Informatica.

In term of GPA higher than 23, the best is Building and Architectural Engineering, and the lowest one is Aeronautical Engineering.

For the time to degree, there is the need to say that many carriers are still active and so I do not have all the students who have already finished their master. However, the fastest students, according to the table below, are management engineers while the slowest are energy engineers.

Table 6.2.9. Outcomes of international students by course of study

COURSE OF STUDY	DROPOUT	GPA< 23	T2D>3
Building And Architectural Engineering	22%	5%	7%
Materials Engineering And Nanotechnology- Ingegneria Dei Materiali E Delle Nanotecnologie	27%	46%	9%
Telecommunication Engineering- Ingegneria Delle Telecomunicazioni	38%	37%	13%
Automation And Control Engineering- Ingengeria Dell'automazione	20%	29%	18%
Energy Engineering- Ingegneria Energetica	48%	43%	33%
Biomedical Engineering - Ingegneria Biomedica	35%	45%	6%
Management Engineering-Ingegneria Gestionale	21%	39%	5%
Electrical Engineering- Ingegneria Elettrica	32%	32%	11%
Civil Engineering For Risk Mitigation	42%	18%	7%
Computer Science And Engineering - Ingegneria Informatica	75%	26%	7%
Aereonautical Engineering - Ingegneria Aeronautica	32%	82%	18%
Mechanical Engineering - Ingegneria Meccanica	53%	43%	8%

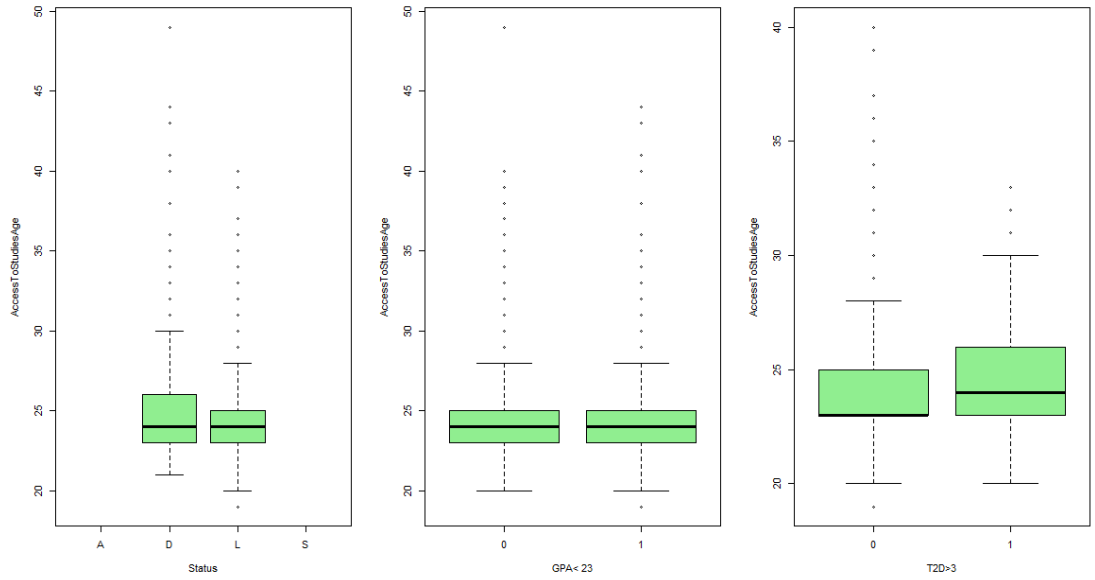
Source: Author's release

6.2.2.2. Descriptive analysis of the correlations among the variables

Now, the aim is finding the hidden correlations between variables and doing a purely descriptive analysis between variables in the school of engineering.

There is no evidence to state that the age is correlated with time to degree, GPA higher than 23 and dropout since these two boxplots in each figure are quite the same for class 1 and class 0, except the outliers. The graph that changes the most is the one about time to degree.

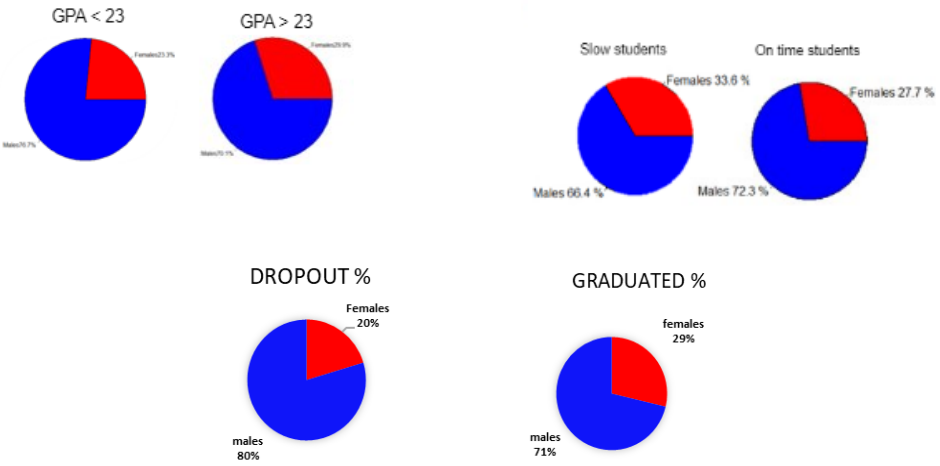
Figure 6.2.10. Boxplot correlating international students' outcomes with their age



Source: Author's release

Figure 6.2.11. shows the ratio between female and male students for each variable: there population of engineers is dominated by men. However, while males performed better in term of GPA higher than 23 (20% against the 16% of women), women are faster in concluding their academic path (26% against 33%), and they are more resilient in dropping out (11% against 17%).

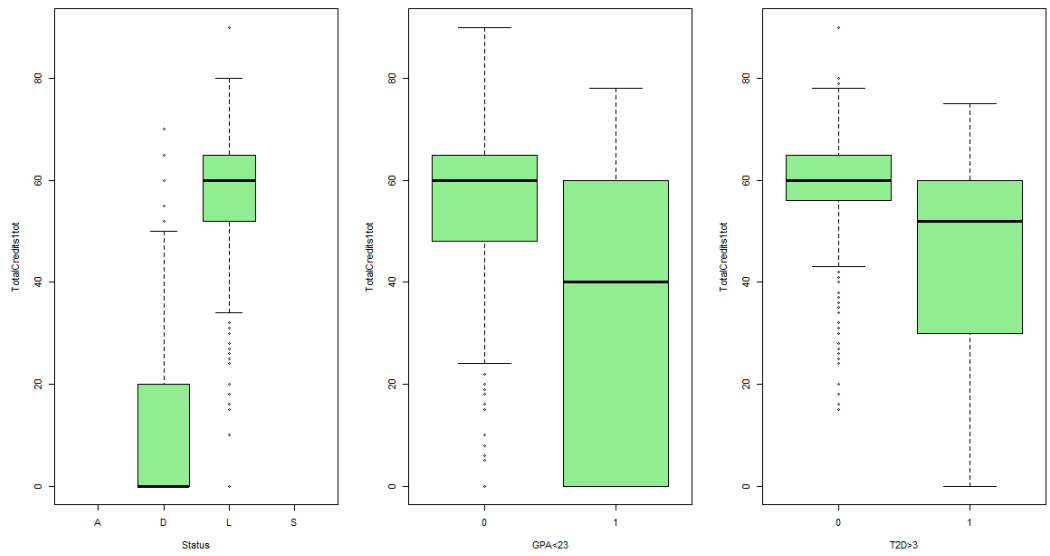
Figure 6.2.11. Cakegraphs correlating students' outcomes with their gender



Source: Author's release

CFU earned in the first year is strictly correlated with these variables, as shown in the figure 6.2.12.

Figure 6.2.12. Boxplot correlating international students' outcomes with their acquired CFU in the first year

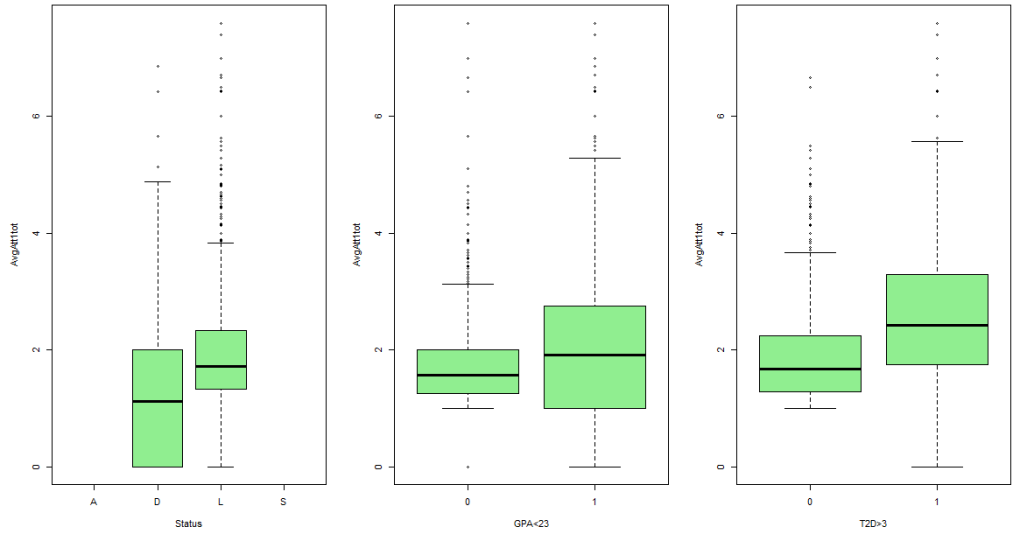


Source: Author's release

The ones who dropped had very few credits in the first year; also, the ones with lower performance in term of GPA and time to degree seem to behave the same.

We will also analyse the relation between these variables and the average number of attempts made in the first year to pass an exam, shown in the figure below.

Figure 6.2.13. Boxplot correlating international students' outcomes with their number of attempts per exam



Source: Author's release

The ones who dropped have very few attempts for an exam during the first year.

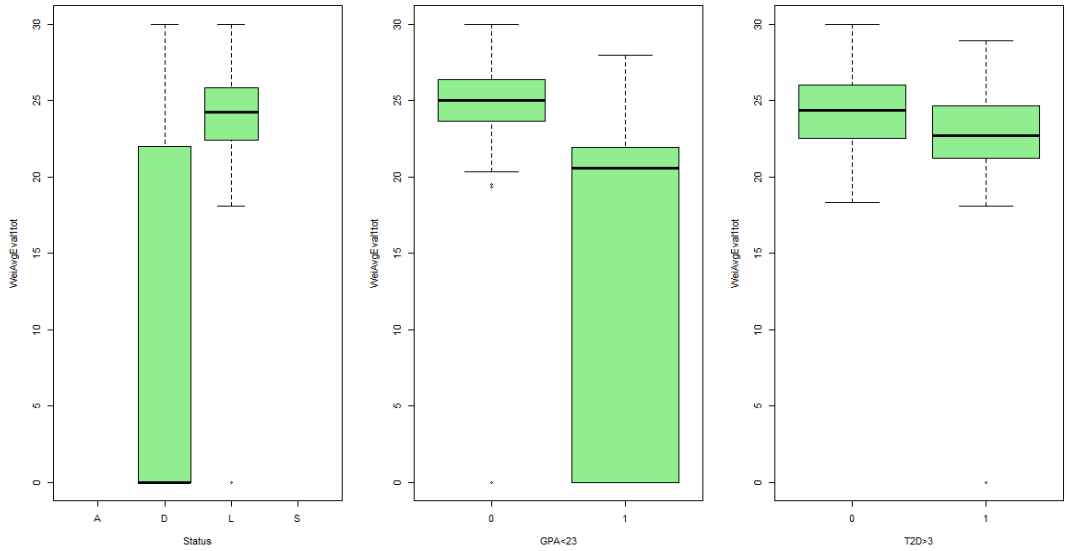
The difference between the students who had taken more than three years to graduate and the ones who took less is that the first ones take the exams several times before accepting it or pass it.

While the ones with a GPA lower than 23, took the exams more times than the ones with a higher GPA.

Finally, the analysis of the correlation between the GPA during the first year and the three variables is held.

In this case: the dropout students have low GPA or equal to zero, when they drop with zero CFU did; the lower GPA performance at the end have also lower GPA the first year, this happened because they are correlated, and the slow students have fewer GPA points during the first year than the faster ones.

Figure 6.2.14. Boxplot correlating international students' outcomes with their acquired GPA in the first year



Source: Author's release

6.2.2.3. Conclusion

There is significant evidence to say that there are some countries that have lower performance than the others. India and Pakistan have the highest number of dropouts in the considered years, while Colombia has the lowest one. India has the greatest number of lower performer students in term of GPA, while Iran and Egypt have the greatest number of students with a GPA higher than 23. Colombia and Turkey have a very small number of slow students, while Iran has the greatest amount. Considering this, it is noticeable that different countries have different cultures and their students tend to behave almost the same in performing. Egypt and Pakistan are the only two divisions in which the students seem to behave the same respectively to the GPA and time to degree.

According to the division by course of study, computer science and engineering has the greatest number of dropouts, aeronautical engineering has the highest number of students with lower GPA, and energy engineers are the students who took more time to graduate.

Regarding the correlations of the variables, the most significant differences can be noticed on:

- Average evaluation in the first year that is different from the students who dropout and not, and the students who have GPA higher than 23 from the ones who have not,
- Total credits achieved in the first year divide the students who dropout from the ones who do not, and also the ones with GPA higher than 23 from the ones who have not,
- The students who usually have the lowest number of attempts per exam tend to have a higher probability of dropout.

6.2.3. Cluster Analysis

To further explore visually the performance of the international students, I decide to design a cluster analysis among two dimensions: time to degree and GPA.

To design the clusters, I do not consider the students who dropout and who had their careers still active. Active carriers are the ones not finished with graduation and not finished as dropouts.

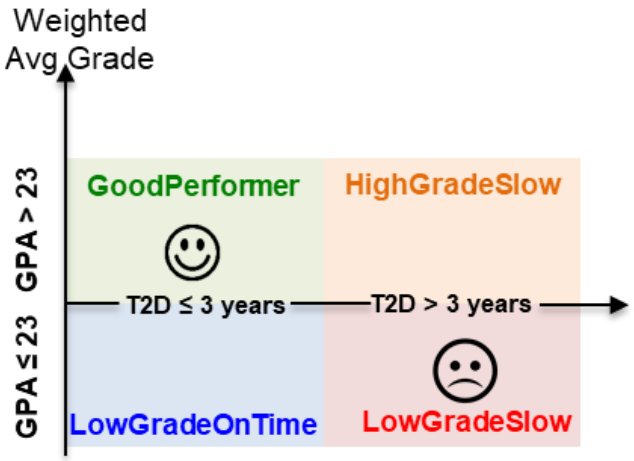
I divide the dimensions of the cluster analysis, time to degree and GPA, following the criteria: GPA higher than 23, GPA lower than 23, time to degree higher than 3 years, time to degree lower than 3 years.

According to this division, the students are divided into 4 clusters:

1. GoodPerformer: Best international students in term of GPA and time to degree
2. LowGradeonTime: Best international students in term of time to degree but worst students in term of GPA
3. HighGradeSlow: Best international students in term of GPA but worst students in term of time to degree
4. LowGradeSlow: Worst international students in term of GPA and time to degree.

Figure 6.2.15. shows the clusters. The preferences of the clusters followed the list just written since I suppose that for the policymakers is more critical that a student finishes on time than the grade with which s/he graduates.

Figure 6.2.15. Clusters of students

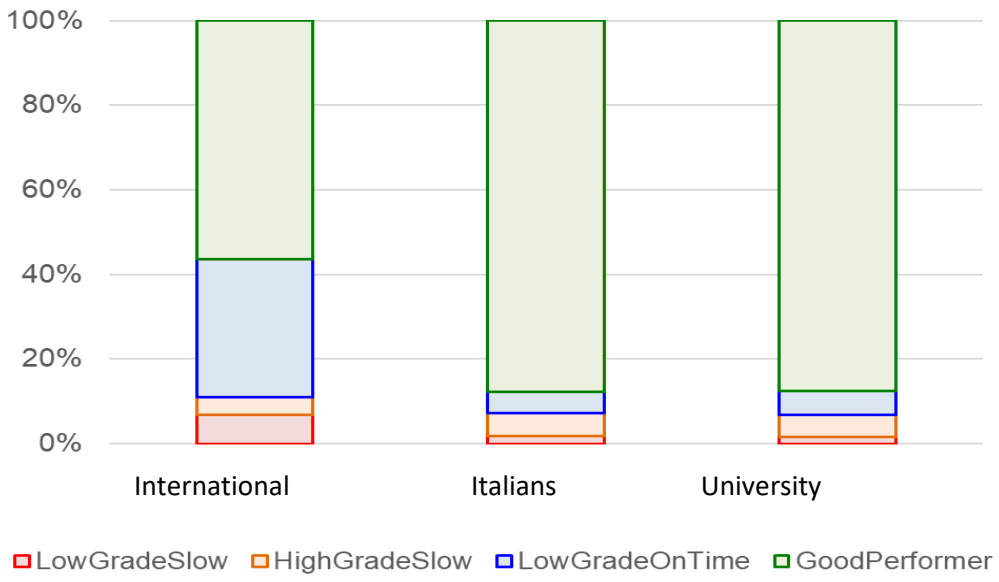


Source: Author's release

The aim now is assessing how the students of these clusters are spread in term of years, and countries. Moreover, a general comparison about how the other classes of students, Italian and University, are spread will be given in order to assess, another time, that international students perform worse than others.

Figure 6.2.16. shows this last topic. As evident, university and Italian students performed quite the same during the years, with a very significant part of good performers. International students have a big part of LowerGradeOnTime and a higher part of LowGradeSlow. On the other hand, it is noticeable that non international students also have a higher slice of HighGradeSlow, although it is not very significantly different.

Figure 6.2.16. Comparison of clusters between the three group of students



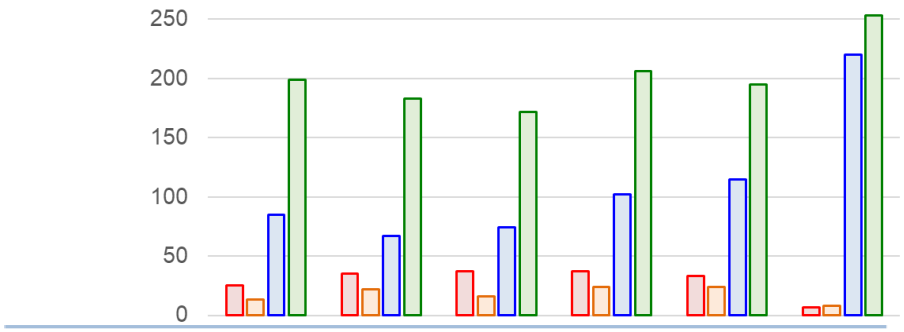
Source: Author's release

In the figure below, there are represented the number of international students per cluster per year from 2010 to 2018.

As noticeable, GoodPerformers are decreasing, giving their places to LowGradeOnTime.

LowGradeSlows are constant over the years, considering that in 2014-2015 there are still a lot of active careers, HighGradeSlows behave the same.

Figure 6.2.17. comparison of clusters over the years



	2010	2011	2012	2013	2014	2015
GoodPerformer	61.80%	59.61%	57.53%	55.83%	53.13%	51.84%
LowGradeOnTime	26.40%	21.82%	24.75%	27.64%	31.34%	45.08%
HighGradeSlow	4.04%	7.17%	5.35%	6.50%	6.54%	1.64%
LowGradeSlow	7.76%	11.40%	12.37%	10.03%	8.99%	1.43%
	100%	100%	100%	100%	100%	100.00%

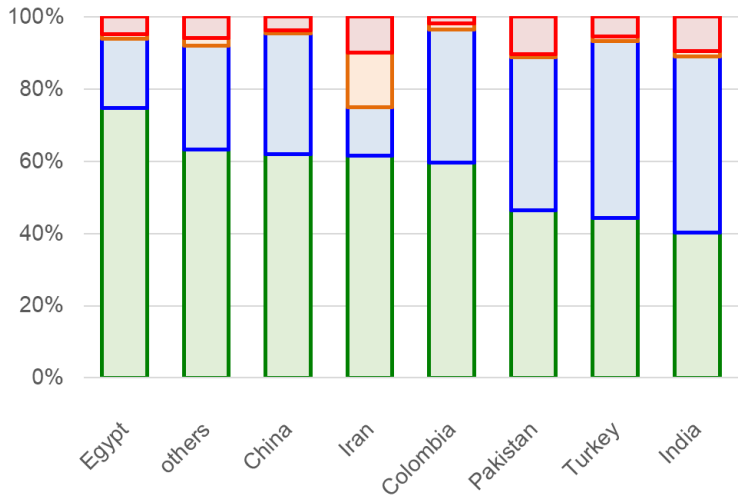
Source: Author's release

Regarding the differences of the belonging clusters across the countries, it is easy to understand that India, Iran, and Pakistan are the lower performers.

Colombia, China, and Egypt have the highest percentage of GoodPerformer along the years and the remaining significant part of LowGradeOnTime.

Turkish are split between GoodPerformers and LowGradeOnTime.

Figure 6.2.18. comparison of clusters among countries



Source: Author's release

6.3. Research question number 2: Which factors are the most relevant in predicting the performance of international students?

In this section, the answer of question 2 will be provided according to the different performance; there will be a part for the most relevant factors for dropout students, one for slow students and the last one for low GPA students. The instructions to read the charts are explained in chapter 5.3. The reader will find out that these results are coherent with the descriptive analysis held in the chapter 6.2., especially focusing the attention on the figures of the random effects.

6.3.1. Dropout

Table 6.3.1. Comparison of algorithms' outputs and summaries for identification of dropout students

<table>
<tr><th>NATIONALITIES</th><th>COURSE OF STUDY</th><th></th></tr>
<tr><td>dropout ~ (1|NazTitolo)
+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+
Sex + AccessToStudiesAge</td><td>dropout ~ (1 | DegreeProgramme.in) + NazTitolo +
WeiAvgEval1tot +
TotalCredits1tot + AvgAtt1tot + Sex +
AccessToStudiesAge</td><td>GMLER- 1st model formula</td></tr>
<tr><td>Scaled residuals:
Min 1Q Median 3Q Max
-6.5278 -0.1617 -0.0864 -0.0486 13.0312

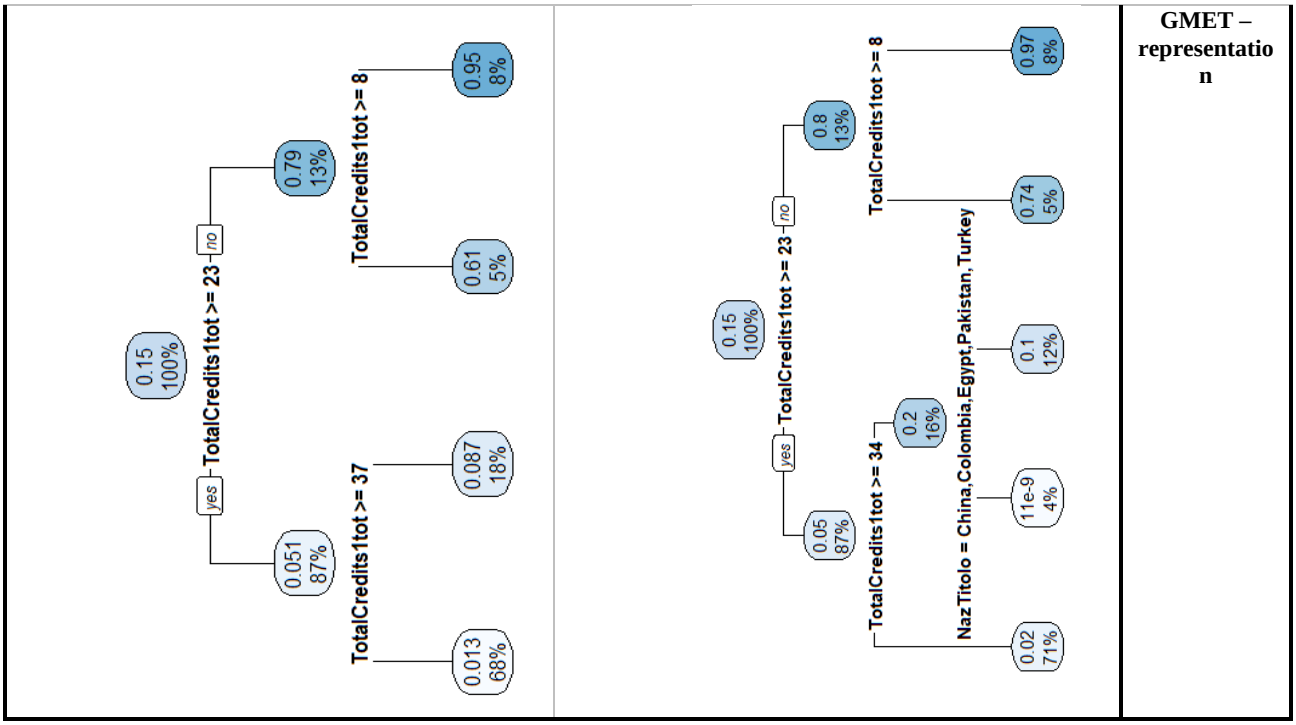
Random effects:
Groups Name Variance Std.Dev.
NazTitolo (Intercept) 0.5214 0.722
Number of obs: 552, groups: NazTitolo, 8

Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.667982 2.065786 -0.323 0.7464
WeiAvgEval1tot -0.098929 0.041750 -2.370 0.0178 *
TotalCredits1tot -0.097623 0.014022 -6.962 3.35e-12 ***
AvgAtt1tot 0.002602 0.147532 0.018 0.9859
SexM 0.140576 0.492836 0.285 0.7755
AccessToStudiesAge 0.157571 0.071515 2.203 0.0276 *</td><td>Scaled residuals:
Min 1Q Median 3Q Max
-7.7158 -0.1503 -0.0764 -0.0203 12.5279

Random effects:
Groups Name Variance Std.Dev.
DegreeProgramme.in (Intercept) 0.3117 0.5583
Number of obs: 552, groups: DegreeProgramme.in, 13

Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.01061 2.41633 -0.004 0.9965
NazTitoloChina -1.66167 1.51527 -1.097 0.2728
NazTitoloColombia -2.96904 1.92065 -1.546 0.1221
NazTitoloEgypt -16.40015 3977.69437 -0.004 0.9967
NazTitoloIndia 0.88451 0.60743 1.456 0.1454
NazTitoloIran -0.10117 0.54686 -0.185 0.8532
NazTitoloPakistan -16.96555 3468.91959 -0.005 0.9961
NazTitoloTurkey -1.03693 1.07384 -0.966 0.3342
WeiAvgEval1tot -0.12331 0.04810 -2.563 0.0104 *
TotalCredits1tot -0.09935 0.01475 -6.737 1.62e-11 ***
AvgAtt1tot -0.04420 0.14453 -0.306 0.7598
SexM 0.09217 0.52909 0.174 0.8617
AccessToStudiesAge 0.18188 0.07782 2.337 0.0194 *</td><td>GLMER – 1st model variables importance</td></tr>
<tr><td>A
MIN 1Q MEDIAN 3Q MAX
-6.4 -0.2 -0.1 0 13.3</td><td>B
MIN 1Q MEDIAN 3Q MAX
6.5 -0.2 -0.1 0 11.5</td><td>GLMER-final model Scaled residuals</td></tr>
<tr><td>VARIANCE STD DEV
0.5447 0.738</td><td>VARIANCE STD DEV
0.3198 0.5655</td><td>GLMER-final model Random effects</td></tr>
<tr><td>ESTIMATE STD. ERROR P-VALUE
Intercept -0.5 1.98 0.8
WeiAvgEval1tot -0.1 0.04 0.02</td><td>ESTIMATE STD. ERROR P-VALUE
Intercept 1.02 1.86 0.58
WeiAvgEval1tot -0.11 0.05 0.01</td><td>GLMER-final model Parameters</td></tr>
</table>

TotalCredit1tot	-0.1	0.01	$2.56*10^{-12}$	TotalCredit1tot	-0.1	0.01	$3.76*10^{-12}$	
AccessToStudiesAGe	0.15	0.07	0.03	AccessToStudiesAGe	0.13	0.06	0.04	
AccessToStudiesAge		**1.166**		**AccessToStudiesAge**		**1.13**		**ODD RATIOS**
WeiAvgEval1tot		**0.84**		**WeiAvgEval1tot**		**0.89**		
TotalCredits1tot		**0.96**		**TottalCredits1tot**		**0.90**		
0.142 0.125							0.0886 0.19	**VPC GMLER/** **GMET**
dropout ~ (1|NazTitolo)+ WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge				dropout ~ (1|DegreeProgramme.in)+ NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt 1tot+ Sex + AccessToStudiesAge				**GMET-formula**
C 1) root 552 41.7960200 0.14855070 2) TotalCredits1tot>=23 479 3.3495050 0.05101963 4) TotalCredits1tot>=37 378 0.1305971 0.01258572 * 5) TotalCredits1tot< 37 101 0.6717888 0.08673304 * 3) TotalCredits1tot< 23 73 3.9926540 0.78851500 6) TotalCredits1tot>=7.5 28 0.5953877 0.60547510 * 7) TotalCredits1tot< 7.5 45 0.1600158 0.94558040 *				D 1) root 552 44.11761000 1.485507e-01 2) TotalCredits1tot>=23 479 4.3732980 4.977222e-02 4) TotalCredits1tot>=33.5 392 0.4012297O 2.003961e-02 * 5) TotalCredits1tot< 33.5 87 1.6375030O 1.979622e-01 10) NazTitolo=China,Colombia,Egypt,Pakistan,Turkey 21 0.01649028 1.123543e-08 * 11) NazTitolo=a.others,India,Iran 66 0.8386341O 1.024985e-01 * 3) TotalCredits1tot< 23 73 4.4034900O 7.967001e-01 6) TotalCredits1tot>=7.5 28 1.3030500O 7.446556e-01 * 7) TotalCredits1tot< 7.5 45 0.1609701O 9.693192e-01 *				**GMET-model**



Source: Author's release

Notes: the outputs of every model can be found in the appendix as well as the codes

Firstly, the VPC shows that the model which has the maximum homogeneity in the group and maximum heterogeneity between groups is the one with the random effects on the courses of study that belong to the GMET algorithm. Followed by the GMET and GMLER with random effect on the nationalities.

In the GMLER_course of study, there are the least differences between group, but, being far from 0, it is significant.

Basing on this, I can understand from the plot of the GMET_courses of study that the essential variables that affect the retention of the students are the total credits earned in the first year and the student's nationality. In particular, if s/he has less than 23 credits at the end of the first year, s/he has the 80% of probability of dropout. The GMET_nationality confirms this fact. So, according to them, the number of credits in the first year affects the dropout positively. Regarding the nationality, it shows that if a student is Indian or Iranian has a major probability of dropping out.

Looking now at the odds ratios and the summary of the GMLER models, based on the nationalities, it can be found that the other two critical variables are the age of the students when s/he enrolled, and the GPA gained in the first year. In particular, age hurts the dropout while the GPA affects it positively.

Focusing now in details on the different GMLER summary, they confirm what odd ratios revealed. In particular, the most significant variable for model A is the total credits achieved in the first year, and it is the same for model B. Both the models seem to be robust and acceptable by the instructions of chapter 5.3.2.

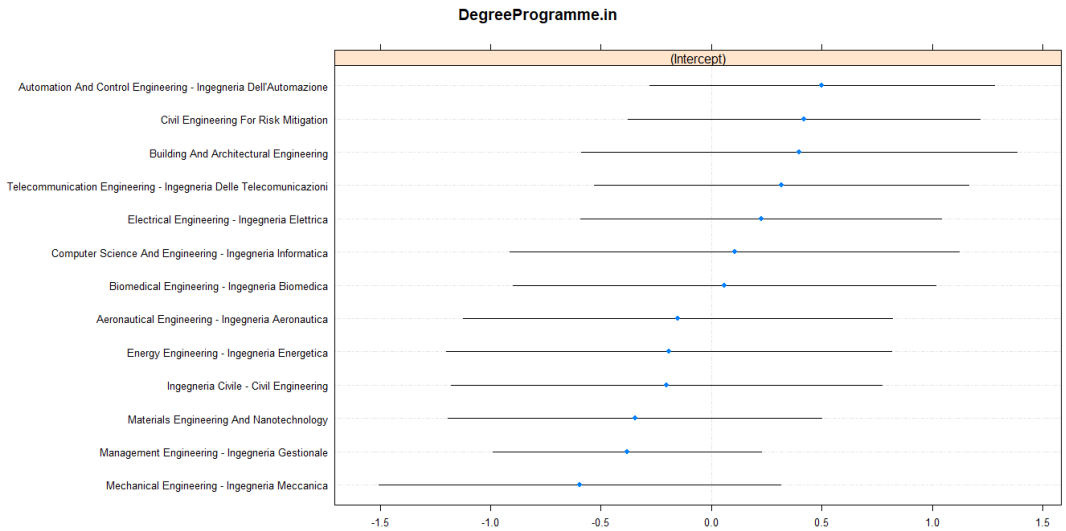
Regarding the trees, the conditions of the splitting are the same except from the number of total credits in the second branch: model C has a number of credits of the first year below 37 as threshold, while model D has 34 credits as threshold. Moreover, from model D, it is possible to notice the nationalities that have more probability of dropout (10%) that, as already said, are Indians and Iranians.

- Conditioned variances

➢ COURSE OF STUDY

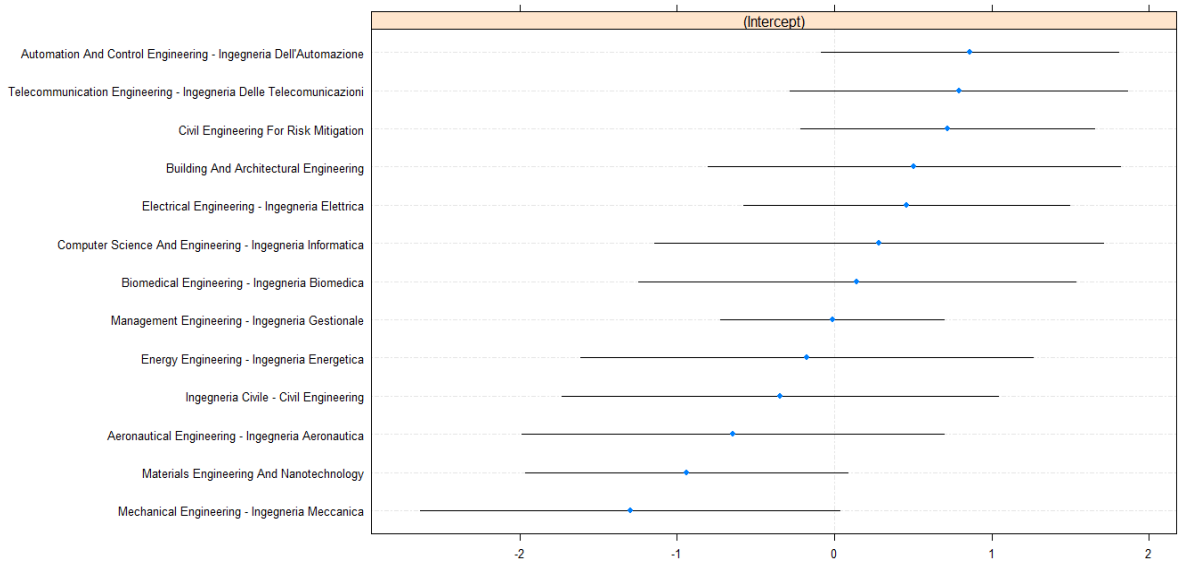
GLMER

Figure 6.3.1. random effect on the course of study for dropout using GMLER



Source: Author's release

Figure 6.3.2. random effect on the course of study for dropout using GMET



Source: Author's release

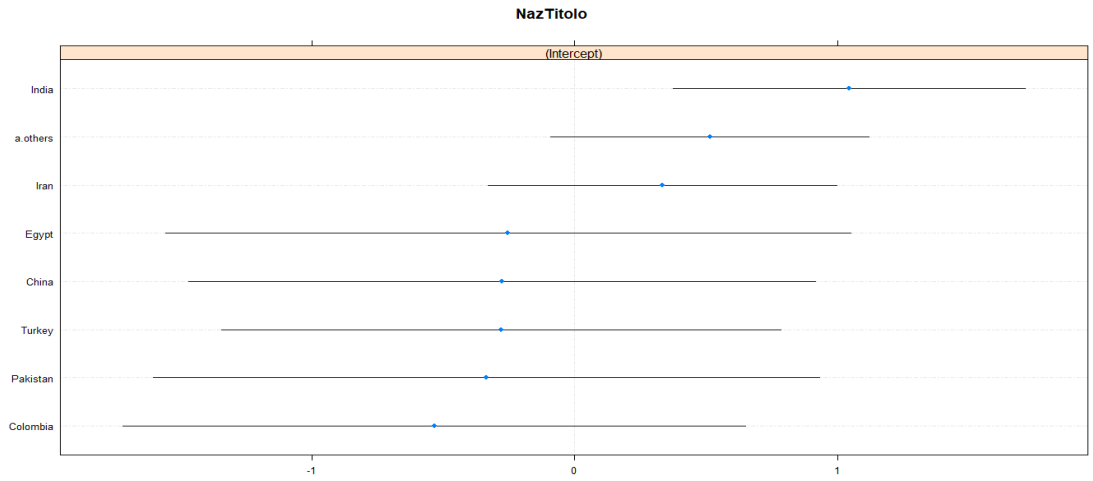
These two figures show the same first course of study and the last one: Automation and Control Engineering is the course where international students have the most significant probability of dropout, while the one with the minor is Mechanical Engineering.

In both of the models the nesting of the courses of study seem to be relevant.

- NATIONALITY

GLMER

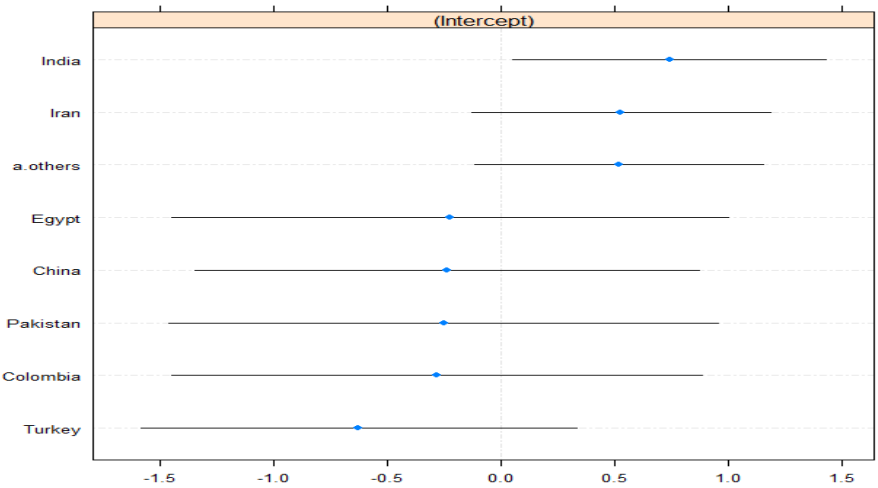
Figure 6.3.3. random effect on nationality for dropout using GMLER



Source: Author's release

GMET

Figure 6.3.4. random effect on nationality for dropout using GMET



Source: Author's release

These two figures show a result that confirms the GMET_course of study results; Indians and Iranians are the most likely students to dropout. The students from universities in all the other countries are less likely to drop out.

In both of the models, the nesting does not seem to be very relevant.

- Hypothesis verification:

Table 6.3.2. Verification of the hypothesis related to the dropout

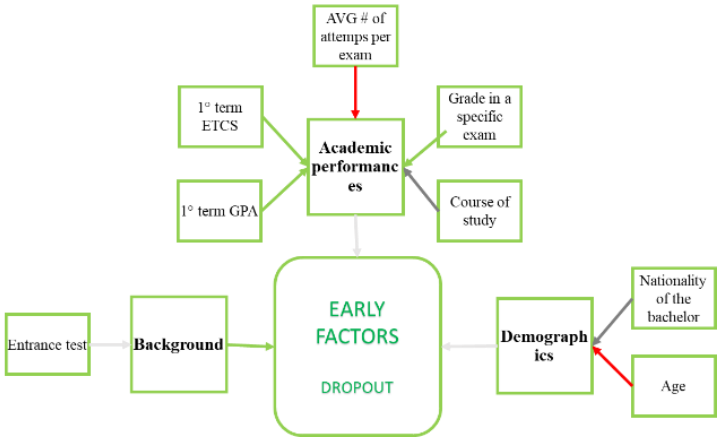
H13	first term ETCS affect the performance positively	TRUE _ GLMER course of study	TRUE_ GLMER Nationality	TRUE _ GMET course of study	TRUE _ GMET Nationality
H18	high 1st term GPA affects the dropout positively	TRUE _ GLMER course of study	TRUE _ GLMER Nationality	FALSE _ GMET course of study: 1st term GPA does not appear in the model	FALSE _ GMET Nationality: 1st term GPA does not appear in the model
H20	the average number of attempts per exam affects negatively the dropout	FALSE _ GLMER course of study: 1st term attempts do not appear in the model	FALSE _ GLMER Nationality: 1st term attempts do not appear in the model	FALSE _ GMET course of study: 1st term attempts do not appear in the model	FALSE _ GMET Nationality: 1st term attempts do not appear in the model
H21	age affects negatively the dropout	TRUE _ GLMER course of study	TRUE _ GLMER Nationality	FALSE _ GMET course of study: age does not appear in the model	FALSE _ GMET Nationality: age does not appear in the model
	Nationality of the bachelor	FALSE_ GLMER course of study: Nationality is not in the final model	GLMER Nationality: Not a lot	TRUE _ GMET course of study: Nationality is in the final model	GMET Nationality: Not a lot
	Course of study	GLMER course of study: Not a lot	X	GMET course of study: Not a lot	X

Source: Author's release

Notes: If the answer is TRUE or FALSE without any explanation means that the result of the testing is the same or the opposite, if it is FALSE : "it does not appear in the model" it means that the model does not retain that that variable is important in determining if a student is or not a dropout. The X is for the absence of the variable in the initial model

The table shows which hypotheses belonging to chapter 3.3. are confirmed and which are not. Here, the only available variables are considered. The figure below shows the initial hypotheses.

Figure 6.3.5. Framework about the hypothesis of the early factors on the dropout



Source: Author's release

6.3.2. Time to degree higher than 3 years

Table 6.3.3. Comparison of algorithms' outputs for identification of slow students, the outputs of every model can be found in the appendix as well as the codes

NATIONALITIES	COURSE OF STUDY	
HighTimetoDegree ~ (1|NazTitolo) +WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge	HighTimetoDegree ~ (1|DegreeProgramme.in)+ NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge	**GMLER - formula**
Scaled residuals: Min 1Q Median 3Q Max -2.4410 -0.3478 -0.2062 -0.1157 5.1763 Random effects: Groups Name Variance Std.Dev. NazTitolo (Intercept) 0.6958 0.8342 Number of obs: 466, groups: NazTitolo, 8 Fixed effects: Estimate Std. Error z value Pr(>|z|) (Intercept) 2.79466 2.34730 1.191 0.23382 WeiAvgEval1tot -0.16731 0.06325 -2.645 0.00817 TotalCredits1tot -0.04853 0.01057 -4.592 4.39e-06 AvgAtt1tot 0.85913 0.16560 5.188 2.13e-07 SexM -0.34944 0.36757 -0.951 0.34177 AccessToStudiesAge -0.01140 0.07246 -0.157 0.87498	Scaled residuals: Min 1Q Median 3Q Max -2.4250 -0.3339 -0.1874 -0.0797 7.2426 Random effects: Groups Name Variance Std.Dev. DegreeProgramme.in (Intercept) 0.1011 0.3179 Number of obs: 466, groups: DegreeProgramme.in, 13 Fixed effects: Estimate Std. Error z value Pr(>|z|) (Intercept) 3.718e+00 2.542e+00 1.463 0.14351 NazTitoloChina -1.549e-01 6.581e-01 -0.235 0.81387 NazTitoloColombia -1.401e+00 1.127e+00 -1.243 0.21384 NazTitoloEgypt -1.614e+01 2.308e+03 -0.007 0.99442 NazTitoloIndia -4.730e-01 5.840e-01 -0.810 0.41798 NazTitoloIran 1.073e+00 4.519e-01 2.374 0.01760 * NazTitoloPakistan 1.412e+00 7.375e-01 1.914 0.05556 . NazTitoloTurkey -1.658e+00 1.016e+00 -1.632 0.10267 WeiAvgEval1tot -1.958e-01 6.871e-02 -2.849 0.00438 ** TotalCredits1tot -4.952e-02 1.123e-02 -4.408 1.04e-05 *** AvgAtt1tot 8.722e-01 1.753e-01 4.976 6.48e-07 *** SexM -4.309e-01 3.879e-01 -1.111 0.26665 AccessToStudiesAge -8.592e-03 7.643e-02 -0.112 0.91049 ---	**GLMER – 1ST model**

MIN (A)	1Q	MEDIAN	3Q	MAX	MIN (B)	1Q	MEDIAN	3Q	MAX	**GLMER-final model Scaled residuals**
-2.2	-0.3	-0.2	-0.1	5	-2.7	-0.3	-0.1	0.08	6.45	

VARIANCE	DEV STD	VARIANCE	DEV STD	GLMER-final model Random effects
0.6938	0.8329	0.08208	0.2865	

	ESTIMATE	STD. ERROR	P-VALUE		ESTIMATE	STD. ERROR	P-VALUE	GLMER-final model Parameters
Intercept	2.4	1.56	0.12	Intercept	3.37	1.78	0.06	
WeiAvgEval1tot	-0.17	0.06	0.006	WeiAvgEval1tot	-0.21	0.07	0.0048	
TotalCredit1tot	-0.05	0.01	$5.3*10^{-6}$	TotalCredit1tot	-0.05	0.011	$1.14*10^{-5}$	
AvgAtt1tot	0.84	0.16	$3.08*10^{-7}$	AvgAtt1tot	0.85	0.17	$8.88*10^{-7}$	
				NazTitoloIran	1.06	0.44	0.018	
				NazTitoloPakistan	1.26	0.72	0.08	

				ODD RATIOS
WeiAvgEval1tot	0.84	WeiAvgEval1tot	0.81	
TotalCredit1tot	0.95	TotalCredit1tot	0.94	
AvgAtt1tot	2.32	AvgAtt1tot	2.33	
		NazTitoloIran	2.88	

		VPC GLMER/ GMET
0.174	0.024	
0.169	0.015	

		GMET-formula
HighTimetoDegree~(1|NazTitolo)+ WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot + Sex + AccessToStudiesAge	HighTimetoDegree~(1|DegreeProgramme.in)+ NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge	

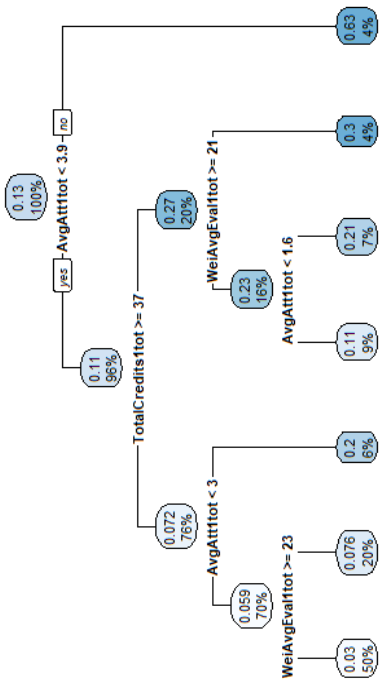
GMET-model

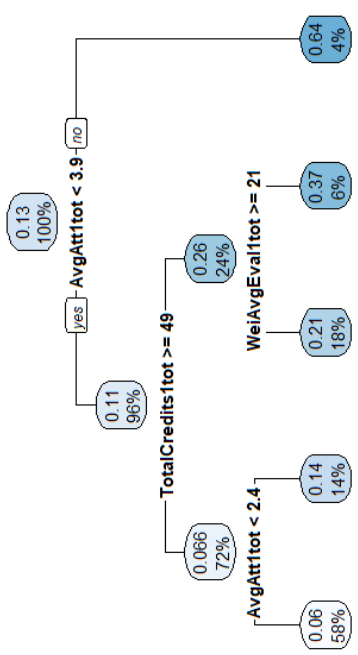
C

```
1) root 466 11.4802900 0.13304720
  2) AvgAtt1tot< 3.944444 448  6.3988890 0.11412840
    4) TotalCredits1tot>=37 354  1.3489330 0.07206247
      8) AvgAtt1tot< 2.95 326  0.4855585 0.05933885
       16) WeiAvgEval1tot>=23.26336 232  0.1606707 0.03044372 *
       17) WeiAvgEval1tot< 23.26336 94  0.1684350 0.07627372 *
      9) AvgAtt1tot>=2.95 28  0.1961310 0.20249680 *
    5) TotalCredits1tot< 37 94  2.0644730 0.27254690
     10) WeiAvgEval1tot>=20.63333 76  0.8905403 0.23207610
       20) AvgAtt1tot< 1.633333 44  0.1534094 0.10642100 *
       21) AvgAtt1tot>=1.633333 32  0.3722138 0.21042560 *
     11) WeiAvgEval1tot< 20.63333 18  0.5238735 0.30311920 *
  3) AvgAtt1tot>=3.944444 18  0.9301535 0.63084830 *
```

D

```
1) root 466 15.30268000 0.13304720
  2) AvgAtt1tot< 3.944444 448  9.89538600 0.11468470
    4) TotalCredits1tot>=48.5 335  2.18673600 0.06617441
      8) AvgAtt1tot< 2.414286 271  0.49238480 0.04158263
       16) NazTitolo=a.others,China,Colombia,Egypt,India,Turkey 213  0.14501040 0.03530
709 *
       17) NazTitolo=Iran,Pakistan 58  0.15006610 0.14112350 *
      9) AvgAtt1tot>=2.414286 64  0.83649640 0.17030520
       18) NazTitolo=a.others,China,Colombia,Egypt,India,Turkey 43  0.20442840 0.094889
54 *
       19) NazTitolo=Iran,Pakistan 21  0.30748260 0.24156140 *
    5) TotalCredits1tot< 48.5 113  4.58320500 0.25849830
     10) WeiAvgEval1tot>=21.225 83  1.67061300 0.19545870
       20) NazTitolo=a.others,China,Colombia,Egypt,India,Turkey 47  0.29189180 0.109614
20
       40) AvgAtt1tot< 1.791667 30  0.05371227 0.10506650 *
       41) AvgAtt1tot>=1.791667 17  0.09580486 0.17816200 *
     21) NazTitolo=Iran,Pakistan 36  0.58017860 0.30990450 *
   11) WeiAvgEval1tot< 21.225 30  1.67018700 0.35643080 *
  3) AvgAtt1tot>=3.944444 18  1.49657400 0.62653560 *
```





This model has been pruned with maxdepth=4 for reasons of representation

GMET-representation

Source: Author's release

Firstly, the VPC shows that the model which has the maximum homogeneity in the group is the one with the random effects on nationality that belong to the GMLER algorithm. Followed by the GMET_nationality. The other two algorithms show that there is not significant homogeneity in the sub clustered designed, so they are not shown any more in this section because the models seem to be worthless: both of them are very closed to zero, a fixed-effect model should have been designed.

Looking now at the odds ratios and the summary of the GMLER model, based on the nationalities, it can be found that there are three critical variables: the GPA gained in the first year, the number of credits earned in the first year and the average number of attempts taken per exam. In particular, the last one hurts the time to degree while the other two have a positive effect on the measured performance.

This was wholly confirmed by the GMET, which shows that the most crucial variable is the average number of attempts, according to which if students did more than an average of 3.9 attempts, s/he would be more likely to graduate in more than three years with a 63% of probability. Moreover, if one student did less than 3.9 attempts but s/he has less than 39 credits in the first year and a GPA in the first year lower than 21, s/he will be more likely to graduate in more than three years.

Focusing better on the two GMLER summaries, it can be found that the most important variables in the models are the total credits in the first year and the average number of attempts in the first years, as also highlighted by the odd ratios. The two model are acceptable according to the conditions expressed in chapter 5.3.2. In model B, Iran is significant with a p-value of 0.018. This confirms the descriptive analysis held in the previous section.

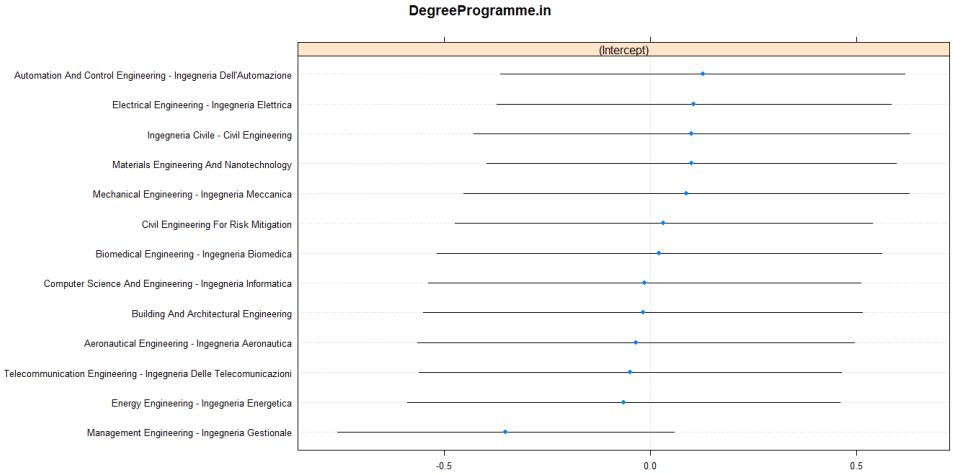
Regarding the trees, the root node's splitting condition is the average number of attempts, the second split's threshold is on the total credits achieved in the first year and it is 37 for model C while 49 for model D. Following the right side of the tree, the following condition is the GPA the first year, which in both of the models appears to be over 21. In model C, there is another condition to be respected: average number of attempts in the first year below 1.6. The left side of the trees are different. In model C, the following branching is made by the average attempts in the first year, as well as for model D, but the threshold is respectively 3 and 2.4. While model D stops there with the branching, model C has its last one according to the GPA over 23. It is crucial to notice that in the tree D there is not branching on the nationality.

- Conditioned variances

  - ➢ COURSE OF STUDY

GMLER

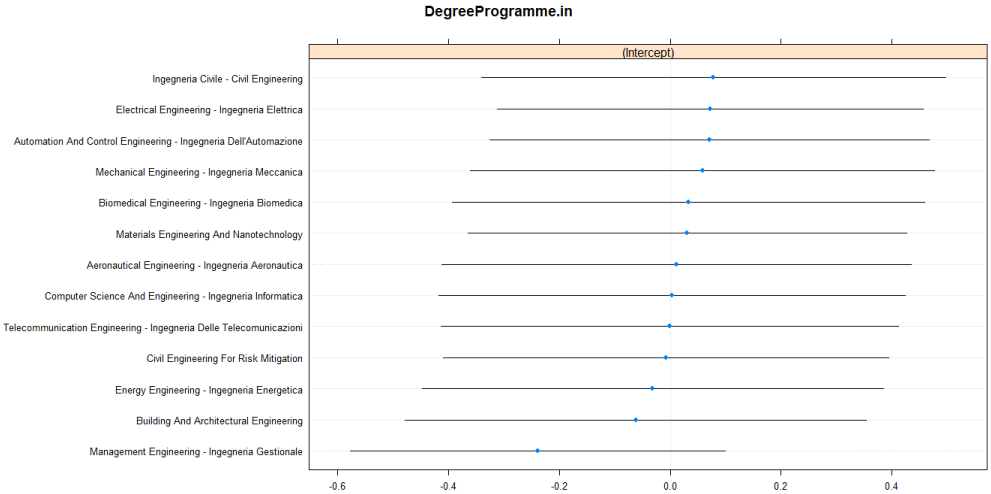
Figure 6.3.6. Random effect on course of study for slow students using GMLER



Source: Author's release

GMET

Figure 6.3.7. random effect on course of study for slow students using GMET



Source: Author's release

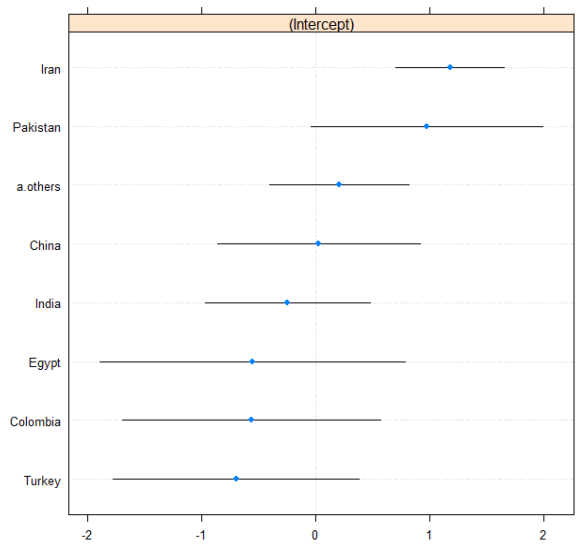
Here, the decision would be to not go into detail with these two figures, for a reason stated above related to the VPC. However, Management engineering seems to be the course with the lowest probability of having time to a degree higher than three years.

Overall it seems that the nesting on the courses of study is not very significant.

- NATIONALITY

GMLER

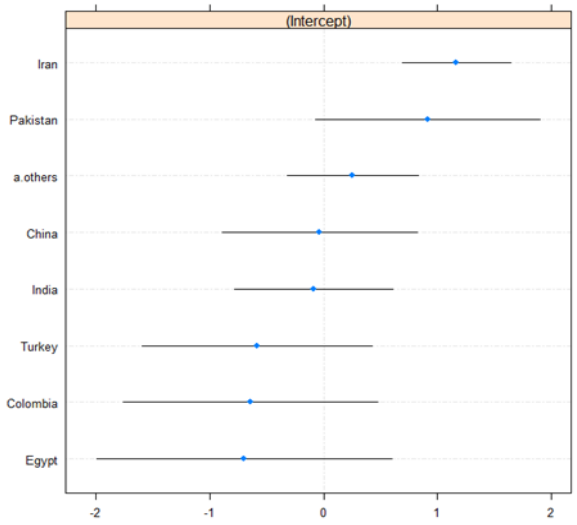
Figure 6.3.8. random effect on nationality for slow students using GMLER



Source: Author's release

GMET

Figure 6.3.9. random effect on nationality for slow students using GMET



Source: Author's release

These two figures show almost the same results, except for an inversion of Turkey and Egypt, but there are insignificant differences among the last three countries. Iranians and Pakistanis seem to be the students who have the highest probability of finishing after the third year.

Overall, it seems that the nesting on the nations is significant.

- Hypothesis verification:

Table 6.3.4. verification of the hypothesis related to the slow students

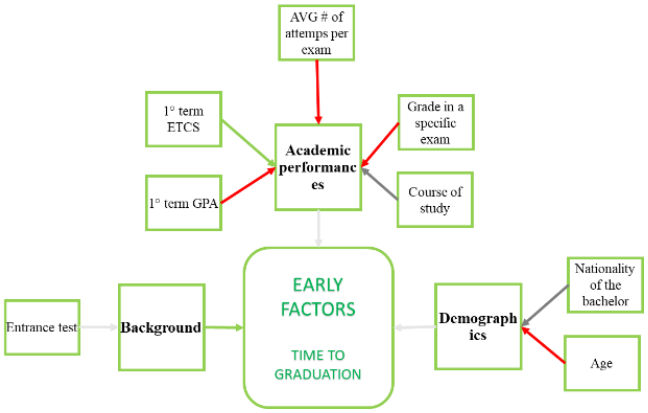
H13	first term ETCS affect positively the performance	TRUE _ GLMER course of study	TRUE_ GLMER Nationality	TRUE _ GMET course of study	TRUE _ GMET Nationality
H22	high 1st term GPA affects negatively the time to graduation	FALSE _ GLMER course of study	FALSE_ GLMER Nationality	FALSE _ GMET course of study	FALSE _ GMET Nationality
H24	average number of attempts per exam affects negatively the time to graduation	TRUE _ GLMER course of study	TRUE_ GLMER Nationality	TRUE _ GMET course of study	TRUE _ GMET Nationality
H25	age affects negatively the time to graduation	FALSE _ GLMER course of study: age does not appear in the model	FALSE _ GLMER Nationality: age does not appear in the model	FALSE _ GMET course of study: age does not appear in the model	FALSE _ GMET Nationality: age does not appear in the model
	Nationality of the bachelor	TRUE _ GLMER course of study	GLMER Nationality: Not a lot	TRUE _ GMET course of study	GMET Nationality: Not a lot
	Course of study	GLMER course of study: Not a lot	X	GMET course of study: Not a lot	X

Source: Author's release

Notes: If the answer is TRUE or FALSE without any explanation means that the result of the testing is the same or the opposite, if it is FALSE : "it does not appear in the model" it means that the model does not retain that that variable is important in determining if a student is or not a slow student. The X is for the absence of the variable in the initial model.

The table shows which hypothesis belonging to chapter 3.3. are confirmed and which are not. Here, the only available variables are considered. The figure shows the initial hypothesis.

Figure 6.3.10. framework about the hypothesis of the early factors on the time to graduation



Source: Author's release

6.3.3. GPA inferior than 23

Table 6.3.5. comparison of algorithms' outputs for identification of low GPA

<table>
<tr><th>NATIONALITIES</th><th>COURSE OF STUDY</th><th></th></tr>
<tr><td>GPAlower23 ~ (1|NazTitolo)+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge+YearsToFinishDegree</td><td>GPAlower23 ~ (1|DegreeProgramme.in)+ NazTitolo+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge+ YearsToFinishDegree</td><td>GMLER-formula</td></tr>
<tr><td><pre>Scaled residuals:
    Min   1Q Median   3Q   Max
-20.9874 -0.4724 -0.2948 0.0869 3.8853

Random effects:
 Groups   Name   Variance Std.Dev.
 NazTitolo (Intercept) 0.5182   0.7199
Number of obs: 466, groups: NazTitolo, 8

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.059582   1.839627   0.576 0.56463
TotalCredits1tot -0.028076   0.009224 -3.044 0.00233 **
AvgAtt1tot   0.943594   0.164286   5.744 9.27e-09 ***
YearsToFinishDegree 0.677582 0.218187   3.106 0.00190 **
SexM   -0.070202   0.300605 -0.234 0.81535
AccessToStudiesAge -0.185865 0.064383 -2.887 0.00389 **</pre></td><td><pre>Scaled residuals:
    Min   1Q Median   3Q   Max
-6.5057 -0.4205 -0.2192 0.0443 4.1365

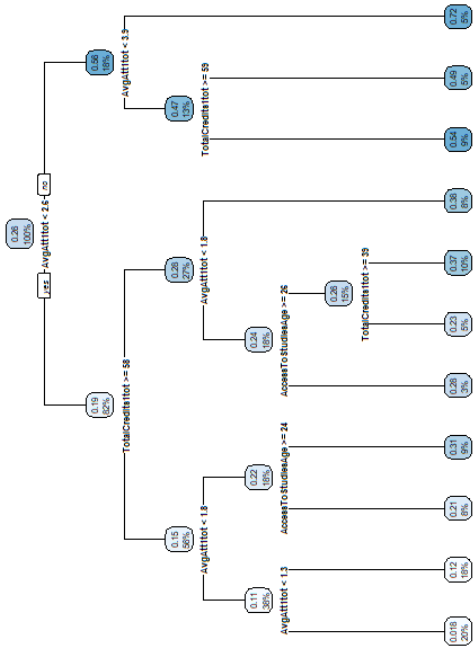
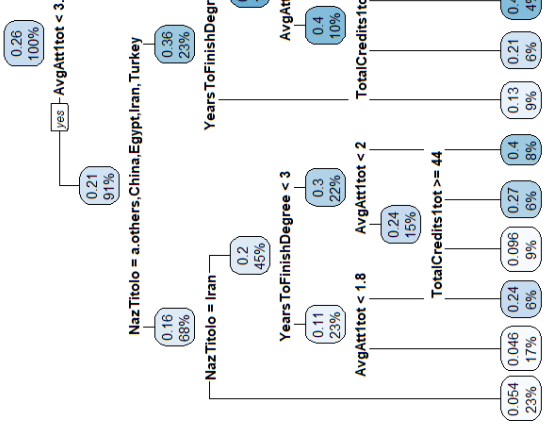
Random effects:
 Groups   Name   Variance Std.Dev.
 DegreeProgramme.in (Intercept) 2.49   1.578
Number of obs: 466, groups: DegreeProgramme.in, 13

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.30477   2.09737 -0.145 0.884467
NazTitoloChina   -0.12229   0.50990 -0.240 0.810457
NazTitoloColombia 0.59010   0.58122   1.015 0.309972
NazTitoloEgypt   -0.49915   0.77308 -0.646 0.518495
NazTitoloIndia   0.42199   0.42258   0.999 0.317989
NazTitoloIran   -2.33184   0.49339 -4.726 2.29e-06 ***
NazTitoloPakistan -0.05304   0.71941 -0.074 0.941224
NazTitoloTurkey   -0.77711   0.53582 -1.450 0.146968
TotalCredits1tot -0.02482   0.01019 -2.436 0.014848 *
AvgAtt1tot   1.26063   0.22291   5.655 1.56e-08 ***
SexM   -0.24068   0.33699 -0.714 0.475103
YearsToFinishDegree 0.86338   0.24642   3.504 0.000459 ***
AccessToStudiesAge -0.18307   0.06877 -2.662 0.007766 **</pre></td><td>GLMER – 1ST model</td></tr>
<tr><td>
<table>
<tr><th>MIN</th><th>1Q</th><th>MEDIAN</th><th>3Q</th><th>MAX</th></tr>
<tr><td>-21.22 A</td><td>-0.5</td><td>-0.3</td><td>-0.09</td><td>3,8</td></tr>
</table>
</td><td>
<table>
<tr><th>MIN</th><th>1Q</th><th>MEDIAN</th><th>3Q</th><th>MAX</th></tr>
<tr><td>-6.9 B</td><td>-0.4</td><td>-0.2</td><td>0.04</td><td>3.9</td></tr>
</table>
</td><td>GLMER-final model Scaled residuals</td></tr>
<tr><td>
<table>
<tr><th>VARIANCE</th><th>STD DEV</th></tr>
<tr><td>0.51</td><td>0.7171</td></tr>
</table>
</td><td>
<table>
<tr><th>VARIANCE</th><th>STD DEV</th></tr>
<tr><td>2.417</td><td>1.555</td></tr>
</table>
</td><td>GLMER-final model Random effects</td></tr>
<tr><td>
<table>
<tr><th></th><th>ESTIMATE</th><th>STD. ERROR</th><th>P-VALUE</th></tr>
<tr><td>Intercept</td><td>0.97</td><td>1.81</td><td>0.59</td></tr>
<tr><td>TotalCredit1tot</td><td>-0.03</td><td>0.009</td><td>0.00237</td></tr>
<tr><td>AvgAtt1tot</td><td>0.94</td><td>0.16</td><td>$9.62*10^{-9}$</td></tr>
<tr><td>AccessToStudiesAge</td><td>-0.18</td><td>0.06</td><td>0.004</td></tr>
<tr><td>YearsToFinishDegree</td><td>0.68</td><td>0.22</td><td>0.0019</td></tr>
</table>
</td><td>
<table>
<tr><th></th><th>ESTIMATE</th><th>STD. ERROR</th><th>P-VALUE</th></tr>
<tr><td>Intercept</td><td>-0.53</td><td>2.07</td><td>0.79</td></tr>
<tr><td>TotalCredit1tot</td><td>-0.02</td><td>0.01</td><td>0.0177</td></tr>
<tr><td>AvgAtt1tot</td><td>1.24</td><td>0.22</td><td>$1.81*10^{-8}$</td></tr>
<tr><td>AccessToStudiesAGe</td><td>-0.18</td><td>0.07</td><td>0.008</td></tr>
<tr><td>YearsToFinishDegree</td><td>0.87</td><td>0.25</td><td>0.0004</td></tr>
<tr><td>NazTitoloIran</td><td>-2.33</td><td>0.49</td><td>$2.21*10^{-6}$</td></tr>
</table>
</td><td>GLMER-initial model Parameters</td></tr>
<tr><td>
<table>
<tr><td>TotalCredits1tot</td><td>0.97</td></tr>
<tr><td>AccessToStudyAge</td><td>0.83</td></tr>
<tr><td>AvgAtt1tot</td><td>2.55</td></tr>
<tr><td>YearsToFinishDegree</td><td>1.97</td></tr>
</table>
</td><td>
<table>
<tr><td>TotalCredits1tot</td><td>0.98</td></tr>
<tr><td>AccessToStudyAge</td><td>0.84</td></tr>
<tr><td>AvgAtt1tot</td><td>3.45</td></tr>
<tr><td>YearsToFinishDegree</td><td>2.45</td></tr>
<tr><td>Iran</td><td>0.09</td></tr>
</table>
</td><td>ODD RATIO</td></tr>
<tr><td>0.135

0.126</td><td>0.4235

0.334</td><td>VPC GMLER/

GMET</td></tr>
<tr><td>GPAlower23 ~ (1|NazTitolo) +TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge+ YearsToFinishDegree</td><td>GPAlower23 ~ (1|DegreeProgramme.in) +NazTitolo+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge+ YearsToFinishDegree</td><td>GMET-formula</td></tr>
</table>

C 1) root 466 18.98317000 0.25536480 2) AvgAttItot< 2.645833 383 5.34137700 0.18897210 4) TotalCreditsItot>=57.5 259 1.92912200 0.14500020 8) AvgAttItot< 1.845238 177 0.53859150 0.10868280 16) AvgAttItot< 1.292857 92 0.11762000 0.01824602 * 17) AvgAttItot>=1.292857 85 0.25244350 0.11751200 * 9) AvgAttItot>=1.845238 82 0.65315490 0.22339260 18) AccessToStudiesAge>=23.5 39 0.14812230 0.20516070 * 19) AccessToStudiesAge< 23.5 43 0.24661220 0.31326100 * 5) TotalCreditsItot< 57.5 124 1.86547700 0.28081670 10) AvgAttItot< 1.775 86 0.79526070 0.23634180 20) AccessToStudiesAge>=25.5 16 0.10679690 0.27772640 * 21) AccessToStudiesAge< 25.5 70 0.48548700 0.25956840 42) TotalCreditsItot>=39 24 0.04569541 0.23017050 * 43) TotalCreditsItot< 39 46 0.31208990 0.37167800 * 11) AvgAttItot>=1.775 38 0.51512310 0.38327400 * 3) AvgAttItot>=2.645833 83 4.16312500 0.56173110 6) AvgAttItot< 3.866071 61 1.86053300 0.47441330 12) TotalCreditsItot>=59 40 0.70275990 0.54126340 * 13) TotalCreditsItot< 59 21 0.65890600 0.48688080 * 7) AvgAttItot>=3.866071 22 0.54794000 0.71654600 *	D 1) root 466 27.18077000 0.25536480 2) AvgAttItot< 3.183333 425 15.16731000 0.21000290 4) NazTitolo=a.others,China,Egypt,Iran,Turkey 316 6.32598300 0.15793980 8) NazTitolo=Iran 107 0.40497550 0.05426248 * 9) NazTitolo=a.others,China,Egypt,Turkey 209 4.76130900 0.20128570 18) YearsToFinishDegree< 2.5 105 0.48552370 0.10680420 36) AvgAttItot< 1.805556 79 0.09606323 0.04612174 * 37) AvgAttItot>=1.805556 26 0.11037210 0.23758790 * 19) YearsToFinishDegree>=2.5 104 2.39215600 0.29667560 38) AvgAttItot< 1.95 68 0.96705130 0.23764520 76) TotalCreditsItot>=43.5 42 0.29224570 0.09561744 * 77) TotalCreditsItot< 43.5 26 0.40190420 0.27293750 * 39) AvgAttItot>=1.95 36 0.74057670 0.39949220 * 5) NazTitolo=Colombia,India,Pakistan 109 5.50161900 0.36093790 10) YearsToFinishDegree< 2.5 41 0.28354670 0.12844560 * 11) YearsToFinishDegree>=2.5 68 2.90607400 0.47402630 22) AvgAttItot< 2.225 47 1.50381700 0.40227510 44) TotalCreditsItot>=39.5 29 0.48833250 0.20576770 * 45) TotalCreditsItot< 39.5 18 0.53894050 0.47356340 * 23) AvgAttItot>=2.225 21 0.61874490 0.70838720 * 3) AvgAttItot>=3.183333 41 2.07373900 0.77579210 *	GMET
 	 	GMET RAPPRESENTATION

Source: Author's release

Firstly, the VPC shows that the model which has the maximum homogeneity in the group is the one with the random effects on the course of study that belong to the GMLER algorithm. Followed by the GMET_course of study. The other two appear to be ok, but they are highly surpassed by the ones that focus on the course of study.

Looking at the models, it is possible to notice that the most critical variable is the average attempts per exam held in the first year. Another critical factor is the nationality of the previous university: Indians and Pakistanis are more likely to have a low final GPA. On the other hand, they show that Iranian students have a higher probability to have a high GPA.

In particular, looking at the GMET on nationalities it can be easily found that if a student have taken an average number of times per exams higher than 3.2, s/he has the 78% of having a low GPA. If s/he took less than 3.2 average attempts for exams, but s/he comes from India, Colombia or Pakistan and s/he took more than 3 years to graduate and the average number of attempts s/he took was higher than 2.2 then s/he has the 71% of having a low GPA.

Regarding the number of years taken for getting the degree, it is relevant in all of the models except the one of GMET_nationality: it seems that people who took more years to graduate are the ones with lower GPA.

Focusing now on the summaries of the GMLERs, the most significant variables for model A is average number of attempts in the first year, while for model B, the Iranian nationality and average number of attempts in the first year. This confirms the descriptive analysis in chapter 6.2. The two models appear to be acceptable according to the conditions expressed in chapter 5.3.2.

Regarding the trees, the root node condition is average number of attempts in the first year, which is below 2.6 in model C and 3.2 in model D, from there one the variables of the splitting change. Model C has total credits in the first year at the same level as again average number of attempts. In the second level of the tree the variables are the average number of attempts and the total credits achieved in the first year. On the third level, the average attempts and the age of enrolments appear to be significant. The fourth level has only a branching which belong to total credits in the first year.

Model D, after the root node, has a spitting in the nationality, the following step is spitted by Iranian nationality and of years to finish the degree. If a student is Iranian, s/he has the

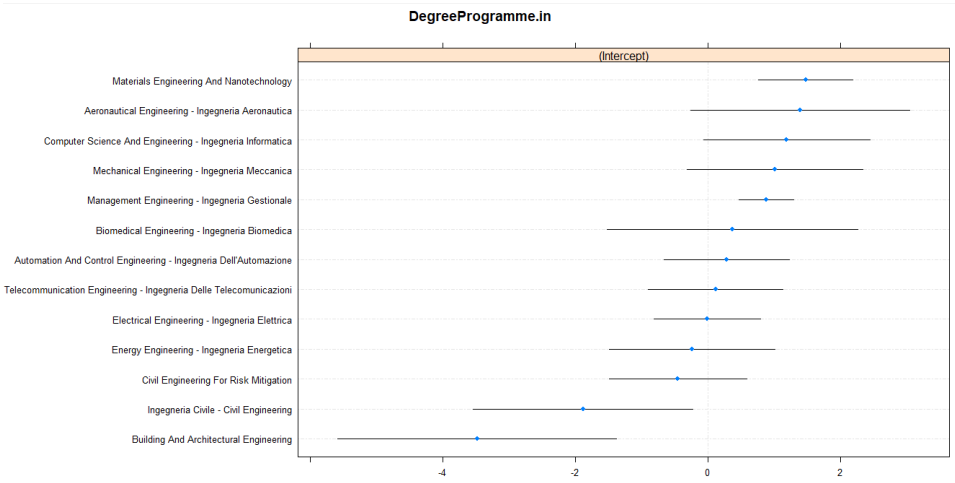
5% of probability of having a GPA lower than 23, this confirms the descriptive analysis of session 6.2.2. the third level is characterized by the presence of years to finish the degree and of average number of attempts. The following level is characterized by average number of attempts and total credits. The final level is composed just by a split on the total number of credits.

- CONDITIONED VARIANCES

  - COURSE OF STUDY

GMLER

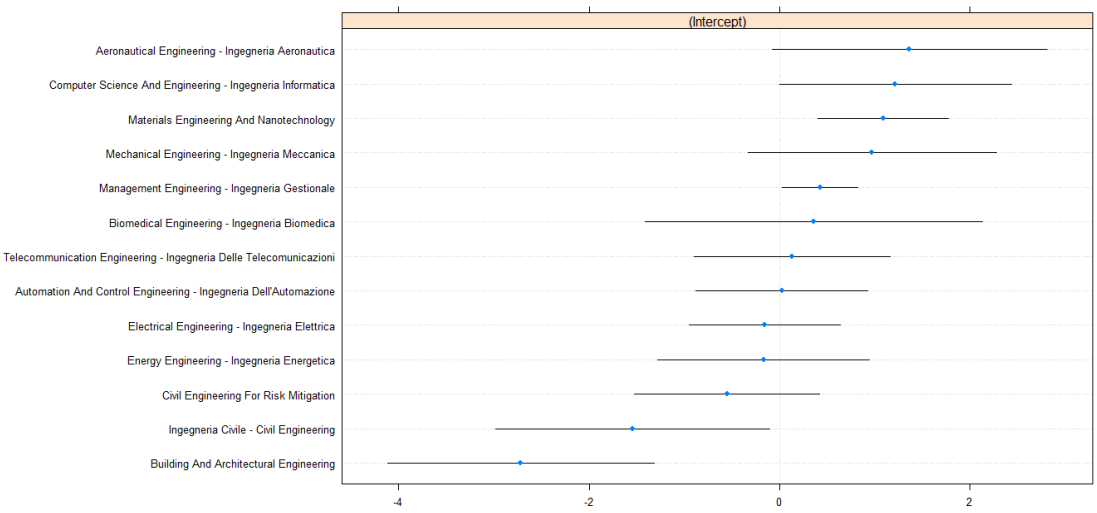
Figure 6.3.11. random effect on course of study for low GPA students using GMLER



Source: Author's release

GMET

Figure 6.3.12. random effect on course of study for low GPA students using GMET



Source: Author's release

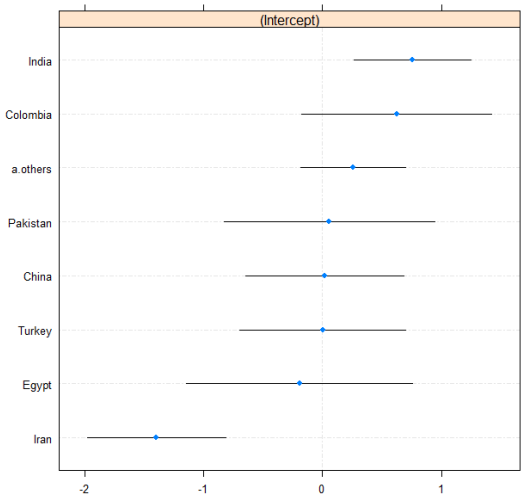
These two figures show the same first courses of study which have a higher probability of having a low GPA at the end of the master, which are Aeronautical Engineering, Material Engineering and Nanotechnology Engineering and Computer Science (although they are positioned in different orders). Regarding the courses of study with the lowest probability of having a low GPA, they are Building and Architectural Engineering, Civil Engineering and Civil Engineering for Risk Mitigation in both of the models.

It is possible to see how much the course of study impacts on the final GPA.

- NATIONALITY

GMLER

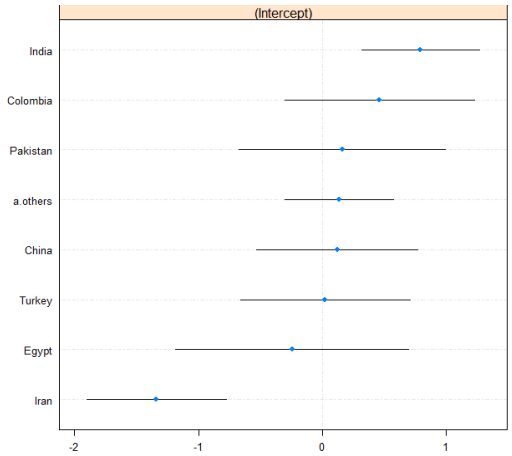
Figure 6.3.13. random effect on nationality for low GPA students using GMLER



Source: Author's realise

GMET

Figure 6.3.14. Random effect on nationality for low GPA students using GMET



Source: Author's realise

The two figures show the exact same results of the random effect on the nationality for the two models, and confirm the result of the previous research question.

According to the models, and the discussion held in the previous section about the variables' importance, the bachelor's nationality, which impacts the most on the students' final GPA are India and Pakistan.

Here, in both of the model, it is shown that also Colombians tend to be lower performers in term of GPA.

To confirm the analysis at the previous chapter, Iranians are the ones most likely to have a low probability of having a lower GPA.

It is possible to notice that, although the course of study's random effect impact on the final GPA in a greater way, it is noticeable that also nationality, in particular Iranian, is very significant to the classification I aim to do.

- Hypothesis verification:

Table 6.3.6. verification of the hypothesis related to the low GPA students

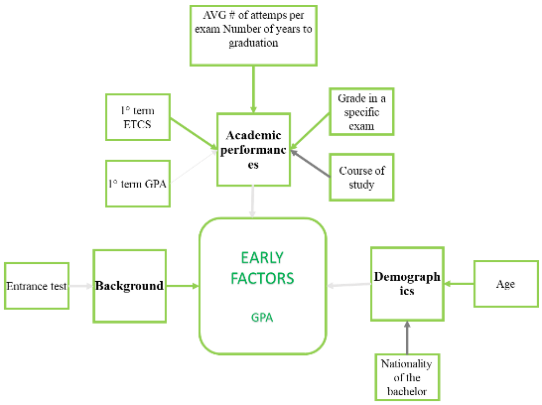
H13	first term ETCS affect positively the performance	TRUE _ GLMER course of study	TRUE_ GLMER Nationality	TRUE _ GMET course of study	TRUE _ GMET Nationality
H15	average number of attempts per exam and the number of years taken to finish the degree affects positively the GPA	FALSE _ GLMER course of study	FALSE_ GLMER Nationality:	FALSE _ GMET course of study	FALSE _ GMET Nationality # of attempts: false Years to finish degree does not appear in the model
H17	student's age affects positively the GPA	TRUE _ GLMER course of study	TRUE_ GLMER Nationality	TRUE _ GMET course of study	TRUE _ GMET Nationality
	Nationality of the bachelor	TRUE _ GLMER course of study	GLMER Nationality: Not a lot	TRUE _ GMET course of study	GMET Nationality: Not a lot
	Course of study	GLMER course of study: a lot	X	GMET course of study: a lot	X

Source: Author's release

Notes: If the answer is TRUE or FALSE without any explanation means that the result of the testing is the same or the opposite, if it is FALSE : "it does not appear in the model" it means that the model does not retain that that variable is important in determining if a student is or not a good GPA student. The X is for the absence of the variable in the initial model.

The table shows which hypothesis belonging to chapter 3.3. are confirmed and which are not. Here, the only available variables are considered. The figure shows these hypotheses.

Figure 6.3.15. Framework about the hypothesis of the early factors on the GPA

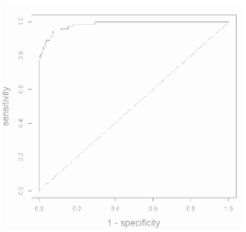
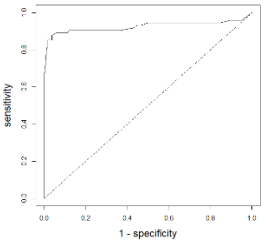
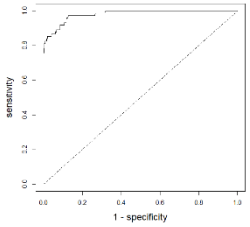
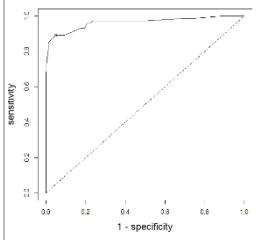


Source: Author's realise

6.4. Research question number 3: Will I be able to predict international students' performance based on machine learning algorithms?

6.4.1. Dropout

Table 6.4.1. Algorithm comparison for the dropout

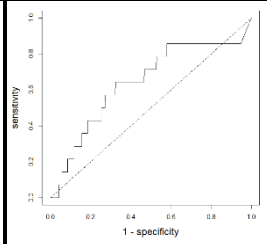
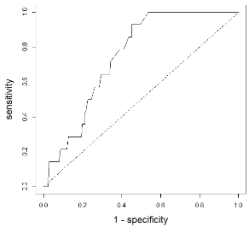
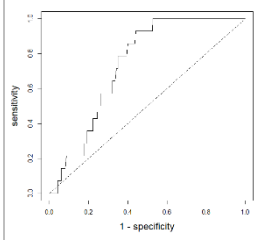
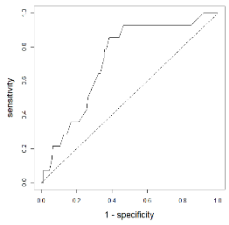
	GLMER- course of study			**GMET- course of study**			**GMLER-nationality**			**GMET-nationality**		
ROC												
Threshold p_0	0.209			0.144			0.199			0.14		
Misclassification table	Obs/ pred	Dropout	Graduated	Obs/ pred	Dropout	Graduated	Obs/ pred	Dropout	Graduated	Obs/ pred	Dropout	Graduated
	Dropout	70	30	Dropout	66	24	Dropout	68	36	Dropout	66	20
	Graduated	4	378	Graduated	8	384	Graduated	6	372	Graduated	8	388
performance	Accuracy	93%		Accuracy	93%		Accuracy	91%		Accuracy	94%	
	Sensitivity	95%		Sensitivity	89%		Sensitivity	92%		Sensitivity	89%	
	Specificity	93%		Specificity	94%		Specificity	91%		Specificity	95%	

Source: Author's realise
Notes: After having tested the algorithm, the aim is trying to understand which one is the best in term of future prediction on the dropout of the international students. The steps are described in chapter 5.3. in the section " Methods."

In the table, the four algorithms designed are showed. The best model in term of accuracy (percentage of well-classified observation over the overall) is the GMET-nationality, in term of sensitivity (TP/(TP+FN)) the GMLER-course of study, while in term of specificity (TN/(TN+FP)) the GMET_course of study. Having written in chapter 5.3. that I aim to minimize the false-negative rate (1-sensitivity), I select the model GMLER_course of study, which can classify 70 dropouts over 74.

6.4.2. *Time to degree higher than 3 years*

Table 6.4.2. Algorithm comparison for the time to degree

	GLMER- course of study			**GMET- course of study**			**GMLER- nationality**			**GMET- nationality**		
ROC												
Threshold p_0	0.129			0.114			0.114			0.129		
Misclassification table	Obs /pred	T2D > 3	T2D< 3	obs /pred	T2D > 3	T2D< 3	obs /pred	T2D > 3	T2D< 3	obs /pred	T2D > 3	T2D< 3
	T2D > 3	8	112	T2D > 3	8	103	T2D > 3	9	128	T2D > 3	9	131
	T2D < 3	6	280	T2D < 3	6	289	T2D < 3	5	264	T2D < 3	5	261
performance	Accuracy		75%	Accuracy		77%	Accuracy		72%	Accuracy		72%
	Sensitivity		57%	Sensitivity		57%	Sensitivity		64%	Sensitivity		64%
	Specificity		71%	Specificity		71%	Specificity		67%	Specificity		66%

Source: Author's realise
Notes: after having test the algorithm, the aim is trying to understand which one is the best in term of future prediction on the time to graduation of the international students. The steps are described in chapter 5.3. in the section " Methods".

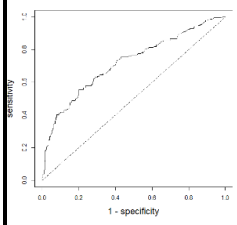
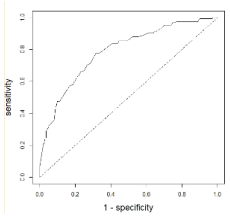
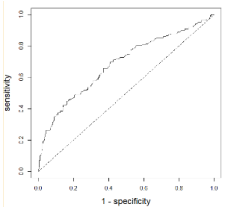
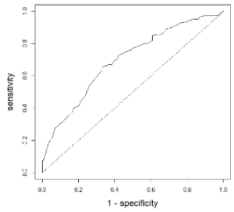
In the table, the four algorithms designed are showed.

It is crucial to notice that for running the algorithm, all the active careers had been deleted since they are not yet concluded, and I do not know if the student will dropout or finish his/her career in over than 3 years. Also for testing, I do not consider the active careers.

The best model in term of accuracy (percentage of well-classified observation over the overall) is the GMET-course of study, in term of sensitivity (TP/(TP+FN)) the GMLER-nationality, while in term of specificity (TN/(TN+FP)) the algorithms with the random effects on the course of study. Having written in chapter 5.3. that I aim to minimize the false-negative rate (1-sensitivity), I select the model GMLER_nationality, which can classify 9 slow students over 14. If the reader have a look to the ROC curve, s/he can notice how worse than the previous case they are: the reason why is that the number of slower students is low due to the detection from the dataset of the still active careers, which just in the training set would have been 86 more observations of late graduation.

6.4.3. GPA inferior than 23

Table 6.4.3. Algorithm comparison for GPA

	GLMER- course of study			**GMET- course of study**			**GMLER- nationality**			**GMET- nationality**		
ROC												
Threshold p_0	0.274			0.337			0.334			0.35		
Misclassification table	obs /pred	GPA< 23	GPA > 23	Obs /pred	GPA< 23	GPA > 23	obs /pred	GPA< 23	GPA > 23	obs /pred	GPA< 23	GPA > 23
	GPA < 23	113	94	GPA < 23	120	104	GPA < 23	109	95	GPA < 23	120	104
	GPA > 23	74	201	GPA > 23	67	191	GPA > 23	78	200	GPA > 23	67	191
performance	Accuracy		65%	Accuracy		65%	Accuracy		65%	Accuracy		64%
	Sensitivity		60%	Sensitivity		64%	Sensitivity		60%	Sensitivity		66%
	Specificity		68%	specificity		66%	Specificity		68%	Specificity		62%

Source: Author's realise
Notes: after having test the algorithm, the aim is trying to understand which one is the best in term of future prediction on the low GPA of the international students. The steps are described in chapter 5.3. in the section " Methods".

In the table, the four algorithms designed are showed. The best model in term of accuracy (percentage of well-classified observation over the overall) is the GMLER-course of study, in term of sensitivity (TP/(TP+FN)) the GMET-nationality, while in term of specificity (TN/(TN+FP)) the GMLER-course of study. Having written in chapter 5.3. that my aim is to minimize the false-negative rate (1-sensitivity), I select the model GMET_nationality, which is able to classify 123 low GPA student over 187.

6.5. Conclusion

In this chapter I deep into the answers to the research questions. In particular, the hypothesis about the first research question is confirmed: the international students do differ from Italian ones, at least in engineering.

There is significant evidence to say that there are some countries that have lower performance than the others. India and Pakistan have the highest number of dropouts in the considered years, while Colombia has the lowest one. India has the greatest number of lower performer students in term of GPA, while Iran and Egypt have the greatest number of students with a GPA higher than 23. Colombia and Turkey have a very small number of slow students, while Iran has the greatest amount. Considering this, it is noticeable that different countries have different cultures and their students tend to behave almost the same in performing. Egypt and Pakistan are the only two divisions in which the students seem to behave the same respectively to the GPA and time to degree.

According to the division by course of study, computer science and engineering has the greatest number of dropouts, aeronautical engineering has the highest number of students with lower GPA, and energy engineers are the students who took more time to graduate.

Regarding the correlations of the variables, the most significant differences can be noticed on the average evaluation in the first year that is different from the students who dropout and not, and the students who have GPA higher than 23 from the ones who have not, the total credits achieved in the first year divide the students who dropout from the ones who do not, and also the ones with GPA higher than 23 from the ones who have not.

The cluster analysis confirms the descriptive analysis and pointed out that comparing to Italian students, the international ones have a higher percentage of low grade on time and of low grade slow. Among the countries, Iranians are the students with the highest percentage of high grade slow, Egyptians have the highest percentage of good performer, Turkey and India have the greatest number of low grades on time, while Iran and Pakistan have the highest percentage of low grade slow.

Secondly, many hypotheses related to the second research question were destroyed and others confirm.

In general, the most significant findings were:

- The GPA in the first year affects positively both the probability of dropout and the probability of being a slow student,
- The total number of credits affect positively all the performance,
- The age of enrolment affects positively the dropout but negatively the probability of having a GPA below 23,
- The average number of attempts per exam in the first year affects negatively the slow students and the students who have a GPA below 23,
- Being an Iranian affect positively the probability of having a GPA lower than 23, but negatively the time to degree,
- Access to study age affects positively only the GPA lower than 23.

Regarding the random effects, the most significant nesting belongs to the students who have GPA lower than 23, in which the course of study appears to be very relevant in determining students' GPA. Less significant were the nationality for the time to degree and for the GPA lower than 23.

Finally, about the predictions held to answer research question number 3, the dropouts' GLMER with random effects on the course of study was the best performing, with an accuracy of 93%, sensitivity of 95% and a specificity of 93%, while for the GPA lower than 23 and the time to degree higher over 3 years, they seemed to need a broader data-range in term of observations and variables since the performance of the implemented algorithms seem to be not very satisfactory. The algorithms chosen in these cases were GMLER for the time to graduation and GMET for the GPA lower than 23, both with random effects on the nationality.

Chapter 7. DISCUSSION, POLICY AND MANAGERIAL IMPLICATIONS & CONCLUSION

7.1. Overview

In this chapter a discussion on the takes away of the analysis I hold is developed, according to the three different research questions.

Moreover, there is a relevant highlight, in each of the sections, about the future development of the project, in order to increase the international students' performance, to increase the number of relevant variables belonging to more multi-dimensional factors, and to increase the power of the algorithms in predicting the final performance of the students.

Finally, a conclusion of what the managerial and policy makers should apply in order to support international students is presented.

7.2. Research question number 1: how to deal with the misalignment in the performance between international and Italian students?

After having understood that the gap between international and Italian students does exist and that it is relevant in the university selected, the focus will be on the supports that the university should give to its students in order to improve their performance.

It seems from the analysis that there is not a problem of internationalization of students, but of their different culture and of the course of study they chose to enrol in.

Among the selected nationalities, the most common in engineering programs in the university, the behaviour of the students appear to be different on average:

- Indians and Pakistanis are the ones who dropout more,
- Iranians belong to the highest percentage of students who took more than 3 years to graduate but they have the lowest percentage of students who have a GPA lower than 23,
- Pakistanis tend to have a lower GPA and to take more than 3 years to graduate,
- Indians have the highest percentage of people who have GPA lower than 23,
- Colombians and Turkish have almost no students who have a high time to graduation,
- Egypt has a low rate of slow students and of people who have low GPA.

In this thesis, the readers could have the chance of understanding that not always lower performance are symptoms of a worse bachelor's preparation or a lower willingness of the students to perform better. There could be significant reasons belonging to no-academic sphere of why they perform worse, as shown in the framework in chapter 3.3.. A low social integration, the sense of responsibility for the lack of money in the family and the cultural shock are three of the most common causes of lower performance in international students. Unfortunately, they are not as easy to measure as the academic parameters.

Regarding the courses of study chosen by the international students, before enrolling they should be aware of the percentages of students who had lower performance studying in the selected courses:

- Computer Science is the course where the 75% of the international students drops out (24 students),
- Energy engineering has the 49% of students who decided to stop their careers (22 students), but also the highest percentage of students who took more than 3 years to graduate 33% (14 students),
- Aeronautical engineering is the course with the highest percentage of students who have a final GPA lower than 23 (13 students).

In the following sections, the corrective actions that the university could take to succeed in its objectives are explained.

These will be useful for university's staff, as well as for policy makers, who intend to study the careers of international students in the world.

7.2.1. Academic support

The most common supports international students usually receive are: orientation programmes, such as the welcome week, that is a week dedicated to students' welcoming and orientation programs which aims to give students the information needed to study and live in a new country and to provide them an initial network of friends; advisors and counsellors available for students in order to get all the information they need and to support them psychologically and in choosing the best options for their study's plan; organizations, with the aim to promote the integration of the students in the culture of the visiting country; workshops and webinars about the academic life. Sometimes the programs are mandatory, sometimes not.

Workshops and webinars on the university academic life, customized per each course of study, included the effort needed to pass the exams, the methods of study the particular subjects will need, all the information about the city where they will live and about the culture, should be provided.

Some schools offer "Academic Tips," which are small narrative videos about these topics.

Tutoring is another service available in many universities, both in online and face-to-face forms. In the first form, it takes the name of online courses, as already explained in chapter 2.6., it is a new learning tool in higher education which aim to give lessons about different

topics to people who are willing to learn. They are usually offered in English, and they are attended by different students. International online courses aim to promote knowledge freely and to benefit more people. One of their primary problems is that sometimes they do not have the possibility to test the knowledge of the students in an effective way or to create their commitments in forums or other activities.

Reich (2014) focused on the relationship between the learning goals and completion rates of the students who attended different MOOCs. He showed that their learning goals had a critical influence on their completion rate: even if students' goal was not to complete the course, and they registered to the course just as trial or watched some lecture's videos, they performed better than the ones who did not even enrol in the course.

7.2.2. Language's problems

Most of the English language programs offered by the university have various levels of fluency and aimed to develop skills in listening, speaking, and academic writing and reading.

Other language development included: language exchanges, programme designed to mix the culture and language of the country where the university is and the countries where the students come from. This usually also leads to a high form of social integration. For example, in Indiana University, a project was offered: a student was paired with a native English speaker to practice English, but this program was also useful for home students who were learning foreign languages. For example, if a native English speaker was learning Chinese and wanted to speak to a Chinese speaker, then s/he could sign up paired with a Chinese international student (Özütürk, 2013).

Moreover, to ease the understandability of English for the mandatory course, an online class per country could be design: the most popular countries in the university are seven, as shown in chapter 4.3., so that, implementing an online class per each spoken language, these students belong to, in English could be a solution. Consequently, they will have less problem with understanding the pronunciation, and they will understand the topics in a more precise way.

International students should furthermore receive a list of books in their native language that covers the pre-requisites needed, in order to better understand the topics and be ready for the tests, without retaking them multiple times.

Boston University designed different duration professional English programs available to its international students: English courses for students in a specific subject are offered, such as business, engineering, or architecture (Matsuda,2012).

At UT Dallas, English conversations are held and focused on improving English through conversation, written test, and games in small interactive groups (Benton et al., 2007).

Another university provided speech therapy to all students, including accent reduction sections; these lessons were led by speech-language pathologists (American Speech-Language-Hearing Association, 2016).

7.2.3. Cultural and social support

Multicultural exchange and social support are usually accomplished through the organization of global festivals, world fairs, and cultural celebrations. These are useful for international students to develop cross-cultural networks and enlarge their cultural intelligence. For example, it could be organized an Indian festival in which many Indians could prepare Indian food or dances, and this will lead to their integration among the other international students, but also among their home students.

Sightseeing trips are implemented by worldwide universities to give international students the chances to visit places and learn about their university's country's culture.

At Northeastern University, the global student's mentors' program deals with international and home students to offer "social, academic, and educational support" through a variety of activities such as mentoring, orientations, workshops, or networking activities (Wadia-Fascetti and Leventman, 2000).

7.3. Research question number 2: hypotheses verification and comparison with the analysis held for the Italian students.

In the first part of this section, a table with all the hypotheses is designed, in order to have an overview on the ones that have been verified, the ones that have not, and the ones that were verified and confirmed. The hypotheses of the model early factor are presented in the table, with their result. The hypotheses were confirmed if the models chosen in chapter 6.6. agreed with them or not.

In the following part, there is a comparison among Italian and International students to verify if the hypotheses, considered wrong or insignificant in the model on the international students, were confirmed in the model on the Italian students. After the table, the reader can read the differences in the models built.

Table 7.3.1. Hypotheses' overview

HYPOTESIS	PERFORMANCE	AREA	VERIFIABLE	CONFIRMED
Cultural intelligence affects positively the performance	DROPOUT GPA<23 T2D>3	Cultural Shock	X	-
Different academic learning approach affects negatively the performance	DROPOUT GPA<23 T2D>3		X	-
Languages' problems affect negatively the performance	DROPOUT GPA<23 T2D>3		X	-
Extra curriculum activities affect positively the performance	DROPOUT GPA<23 T2D>3	Social Integration	X	-
Peer to peer relationship affects positively the performance	DROPOUT GPA<23 T2D>3		X	-
Teachers' willingness to promote learning affects positively the performance	DROPOUT GPA<23 T2D>3		X	-
Informal contacts with academic staff affect positively the performance	DROPOUT GPA<23 T2D>3		X	-
Scholarships and equivalent economic status affect positively the performance	DROPOUT GPA<23 T2D>3		X	-
Students' job affects negatively the performance	DROPOUT GPA<23 T2D>3		X	-
Family's pressure affects negatively the performance	DROPOUT GPA<23 T2D>3		X	-
The sense of isolation affects negatively the performance	DROPOUT GPA<23 T2D>3		X	-

High score in entrance test affects positively the performance	DROPOUT GPA<23 T2D>3	Early Factors	X	-
First term ETCS affect positively the performance	DROPOUT GPA<23 T2D>3		√	√
High first term GPA affects positively the GPA	GPA<23		X	-
Average number of attempts per exam and years of taking the graduation affect positively the GPA	GPA<23		√	X
High grade in a specific exam affects positively the GPA	GPA<23		X	-
Student's age affects positively the GPA	GPA<23		√	√
Nationality affects the GPA	GPA<23		√	√
Course of study affects the GPA	GPA<23		√	√
High first term GPA affects positively the dropout	DROPOUT		√	√
High grade in a specific exam affects positively the dropout	DROPOUT		X	-
Average number of attempts per exam affects negatively the dropout	DROPOUT		√	-
Age affects negatively the dropout	DROPOUT		√	√
Nationality affects the dropout	DROPOUT		√	-
Course of study affects the dropout	DROPOUT		√	√
High 1st term GPA affects negatively the time to graduation	T2D>3		√	X
High grade in a specific exam affects negatively the time to graduation	T2D>3		X	-
Average number of attempts per exams per exam affects negatively the time to graduation	T2D>3		√	√
Nationality affects the time to graduation	T2D> 3		√	√
Course of study affects the time to graduation	T2D>3		√	√
Age affects negatively the time to graduation	T2D>3		√	-

Source: Author's realise

Notes: first column: the hypothesis explained in chapter3.3 , second column: the performance affected, third column: group of variable the hypothesis belong to explained in chapter 2.5 and 3.2, fourth column: the hypothesis is verifiable with the data we have? X: no, √: yes, fifth column: the hypothesis is confirmed by the model selected to be the best one in chapter 6.6? X: no, √:yes, -: the variable does not appear in the model as relevant (if the verifiable answer was yes, otherwise it cannot be determined).

A resume of the hypotheses' verification and of the variables' importance is presented below:

- GPA in the first year affects positively both the probability of dropout and the probability of being a slow student,
- The total number of credits affect positively all the performance,
- The age of enrolment affects positively the dropout but negatively the probability of having a GPA below 23,
- The average number of attempts per exam in the first year affects negatively the slow students and the students who have a GPA below 23,
- Being an Iranian affect positively the probability of having a GPA lower than 23, but negatively the time to degree,
- Access to study age affects positively only the GPA lower than 23.

The most significant nesting belongs to GPA lower than 23, in which the course of study appears to be very relevant in determining students' GPA. Less significant were the nationality for the time to degree and for the GPA lower than 23.

The variables which were supposed to be significant in the models, but which were not, were:

- Average number of attempts per exam affects negatively the dropout,
- Nationality affects the dropout,
- Age affects negatively the time to graduation.

The hypotheses not verified because of a wrong interpretation of their impact were:

- Average number of attempts per exam and years of taking the graduation affect positively the GPA,
- High first term GPA affects negatively the time to graduation.

For the hypotheses that I expected to be true, but actually were not or did not appear in the model, in the following part there is the test on these hypotheses using the Italian students of the university ("Italians" in this case means students who had their bachelor in the university or students who had their bachelor in another Italian university).

To do so, there is an increase of the threshold for the GPA, which for them will be 25, since the average of the final GPA of these Italian students was 26.36, and there is a decrease of the threshold of the time to graduation of these students to two years, because the average time to degree of these students was 2.43 years.

There is no hypothesis double verification on the nationality, since all these Italian students came from an Italian university.

Furthermore, just for the GPA, I test the hypotheses also on the international students who have the GPA higher than 27, in order to see if their academic dynamics change if they belong to a different cluster of students.

In the following sections the comparisons are shown.

7.3.1. Dropout

In the figure below the summary of the GLMER function, the logistic regression used in chapter 6 to run the model, with random effects of the course of study is shown. It is based on the performance of the Italian students, with data from 2013 to 2014.

Table 7.3.1. summary of GMLER with the random effects on the course of study for Italian students for dropout

Formula: Dropout = (1|DegreeProgramme.in)+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+AccessToStudiesAge				
Min	*1Q*	*Median*	*3Q*	*Max*
-7.543	-0.056	-0.033	-0.023	37.010
Variance		*Std. Dev.*		
0.4515		9.672		
	Estimate	*Std. error*	*P-value*	
WeiAvgEval1tot	-0.11	0.002	< 2e-16	
TotalCredits1tot	-0.14	0.002	< 2e-16	
AvgAtt1tot	0.39	0.002	< 2e-16	
AccessToStudiesAge	0.07	0.002	< 2e-16	

Source: Author's realise

As noticeable in the picture, all the variables are very significant in predicting Italian's dropout, having very low p-values.

Odd ratios:

Table 7.3.2. Comparison of the odd ratios of the most relevant variables in the model for dropout

	Italians	**International**
WeiAvgEval1tot	0.89	0.89
TotalCredits1tot	0.86	0.9
AvgAtt1tot	1.47	-
AccessToStudyAge	1.07	1.13

Source: Author's realise

The only hypothesis verifiable here is that the average number of attempts per exam is relevant for the model and it has a negative effect on the dropout, as in the hypothesis.

7.3.2. Time to degree higher than 2 years

In the figure below the summary of the GLMER function with random effects of the course of study is shown. It is based on the performance of the Italian students, with data from 2013 to 2014.

Table 7.3.3. Summary of GMLER with the random effects on the course of study for Italian students for time to degree

Formula: HighTimetoDegree = (1|DegreeProgramme.in)+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+AccessToStudiesAge				
Min	*1Q*	*Median*	*3Q*	*Max*
-2.4349	-0.2535	-0.1743	-0.1231	11.6502
Variance		*Std. Dev.*		
0.2153		0.464		
	Estimate	*Std. error*	*P-value*	
WeiAvgEval1tot	-0.09	0.032	0.003	
TotalCredits1tot	-0.05	0.004	< 2e-16	
AvgAtt1tot	1.04	0.07	< 2e-16	
AccessToStudiesAge	0.12	0.039	0.0019	

Source: Author's realise

It seems that the most significant variables are the total credits and the average number of attempts per exam in the first year, with a p-value which is very low.

Odd ratios:

Table 7.3.4. Comparison of the odd ratios of the most relevant variables in the model for time to degree

	Italians(T2D>2 years)	**International(T2D> 3 years)**
WeiAvgEval1tot	0.9	0.84
TotalCredits1tot	0.95	0.95
AvgAtt1tot	2.82	2.32
AccessToStudyAge	1.12	-

Source: Author's realise

The only hypothesis verifiable here is that the access to study age: for the Italian students, it is relevant, and it affected negatively the time to graduation. While for the GPA at the first term, it confirmed the wrongness of the hypothesis: a high first term GPA affects positively the time to graduation.

7.3.3. GPA inferior than 25

For this performance, on one hand I wanted to show how similar the variables that affect the performance of Italian people are to the ones of International students, although they have overall different performance. On the other hand, I want to show that if I take the sample of students who have the highest group of students who have GPA over 27, the sign of the variables that affect the performance changes. In the first table, the significant Italian students' variables are shown, while in the second one the good performer GPA international students' variables are reported.

Table 7.3.5. Summary of GMLER with the random effects on the course of study for Italian students for GPA<25

Formula: GPAlower25 = (1|DegreeProgramme.in)+TotalCredits1tot+Sex+AccessToStudiesAge				
Min	*1Q*	*Median*	*3Q*	*Max*
-24.9890	-0.4190	-0.2720	-0.1718	6.9514
Variance		*Std. Dev.*		
0.4751		0.6893		
	Estimate	*Std. error*	*P-value*	
TotalCredits1tot	-0.023	0.004	2e-9	
AvgAtt1tot	0.94	0.06	< 2e-16	
SexM	0.45	0.12	0.00001	
AccessToStudiesAge	0.38	0.037	< 2e-16	

Source: Author's realise

Figure 7.3.6. Summary of GMLER with the random effects on the course of study for international students for GPA>27

Formula: GPAhigher27 = (1|DegreeProgramme.in)+YearsToFinishDegree+NazTitolo+AvgAtt1tot+Sex				
Min	*1Q*	*Median*	*3Q*	*Max*
-1.5555	-0.4429	-0.2645	-0.1231	6.9804
Variance		*Std. Dev.*		
0.7965		0.8925		
	Estimate	*Std. error*	*P-value*	
YearsToFinishDegree	1.13	0.298	0.001	
NazTitoloIndia	-1.45	0.59	0.01	
NazTitoloIran	1.21	0.42	0.004	
AvgAtt1tot	-0.55	0.25	0.027	
SexM	0.76	0.37	0.039	

Source: Author's realise

The variables of the first model, related to the Italian students, seem to be very significant, with very low p-values. This means that for the predictions, all the variables in the models are critical and it would give great results. In the second model, none of the variables appears to be very relevant, the only two lower p-values belonged to years to finish the degree and to Iranian's students.

Odd ratios:

Table 7.3.7. Comparison of the odd ratios of the most relevant variables in the model for GPA, according to the GMET output

	Italians	**International (GPA < 23)**	**International (GPA > 27)**
SexM	1.56	-	2.13
TotalCredits1tot	0.98	Positively	-
AvgAtt1tot	2.96	Negatively	0.57
AccessToStudyAge	1.46	Positively	-
YearsToFinishDegree	-	-	2.67

Source: Author's release

According to these results, the age of the enrolment of the Italian students affected negatively their final GPA. The hypothesis for the international students is confirmed.

However, the average number of the exams taken by a student in his/her first year affected negatively his/her final GPA, so, the model on the Italian students confirmed what the model on international students.

Focusing now on the group of good GPA international students, it appears that the average of the attempts per exam taken in the first year affects positively the probability of having

a high GPA: it means that the more attempts a student does in the first year, the higher his/her probability of having a good GPA will be. This confirm the second part of the hypothesis 15.

7.3.4. Conclusion

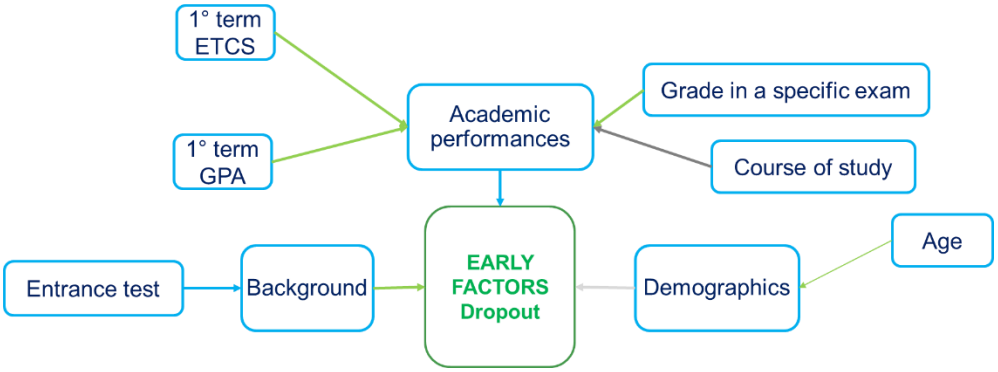
After having held these analyses, different issues were pointed out in my case of study:

1. There is no difference among models and nationalities in the number of credits a student gets in his/her first year and his/her performance. The higher the number of credits, the better will be his/her career.
2. An increase of the average number of attempts per exam taken during the first year always, when it is present as significant variable, means a decrease in the student's performance, no matter the nationality or the models. This happens when there is a focus on the lower grade students: the ones which may struggle to pass the exams or who do not care about the GPA. On the other hand, when there is a focus on the students with higher GPA, this variable affects the GPA positively.
3. The GPA a student has in his/her first year has always a positive impact on his/her final performance, no matter his/her nationality or the model used.
4. The age of the student, as already pointed out from the literature in chapter 2.3., is critical: according to the nationality and the models could appear as relevant or not, could increase student's performance or decrease them.
5. The nationality affects the GPA and the time to degree but does not affect the dropout. This could be related to the theories regarding the cultural shock in chapter 2.5.2., in fact for example an Iranian could be used to not taking care of good marks but pass all the exams on time and get his/her degree as soon as possible.
6. The sex of the student seems to impact only on Italian students' GPA, males have more probability of having a GPA lower than 25.
7. The number of years in which a student takes his/her graduation is not relevant for the best model, but being significant for all the other three models, I consider it as a critical variable determining the GPA. In all the models, including the ones

detecting the higher GPA and the lower, it appears that the higher the number of years a student took to graduate, the lower GPA a student will have.

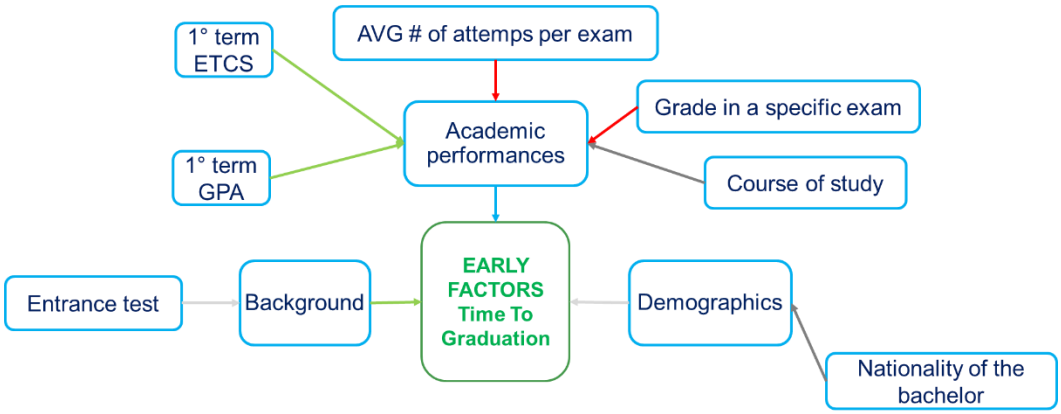
In the figures below the new three models regarding the early factors are shown, according to the results of my analysis, as well as the rules on how to read the different colours.

Figure 7.3.1. New and tested model about the variables that affects the dropout of international students



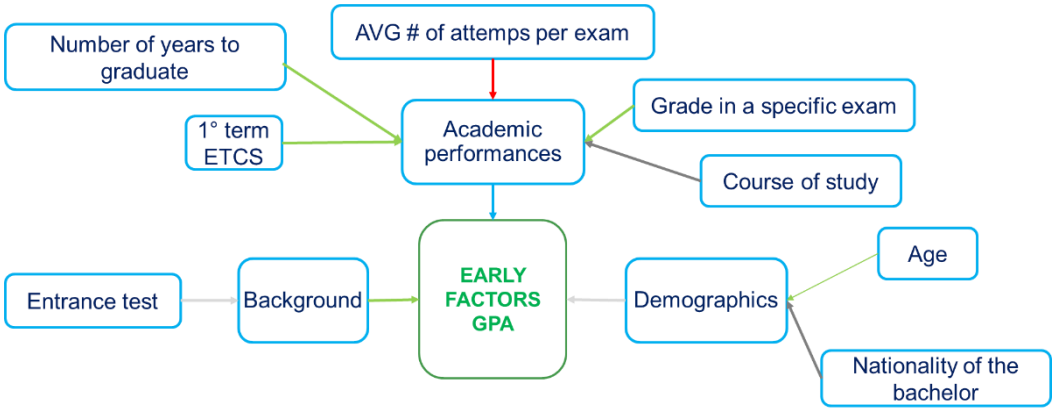
Source: Author's release

Figure 7.3.2. New and tested model about the variables that affects the time for graduation of international students



Source: Author's release

Figure 7.3.3. New and tested model about the variables that affects the GPA of international students



Source: Author's release

- Green arrows: the variable affects positively the performance
  - It increases the GPA
  - It decreases time to graduation
  - It decreases the probability of dropout
- Red arrows: the variable affects negatively the performance
  - It decreases the GPA
  - It increases time to graduation
  - It increases the probability of dropout
- Dark Grey arrows: it depends on the value of the variable if the performance is affected positively or negatively
- Light Grey arrows: the performance is affected in the way explained by the previous or following arrows

7.4. Research question number 3: analysis of the results of the prediction

As noticeable in chapter 7.4. the best performance' prediction is the one which predicts the dropout. Indeed, I was able to have excellent performance on this variable, thanks to GMLER on the course of study. Thanks to this algorithm, supports for the students with the characteristics shown in chapter 6 should be provided. Unfortunately, these students have to enrol in order to be predicted of dropping out or not, since I do not have information about their previous career. An entrance test that compares the results obtained by all the international students could be the solution for understanding which students are more likely to succeed and which are not. Furthermore, in order to understand a different point of view, a questionnaire about their previous experiences with a different culture (inheritance/ international friends/ international experiences) could be helpful to understand if this variable is or not useful in order to predict students' performance in the early stages of their careers.

In addition to this, since I am still not able to predict well the other two performance, for which I chose GMLER for the time to graduation and GMET for the GPA lower than 23, both with random effects on the nationality. I would need other variables and more data in order to take customized actions.

With the increase of the number of international students at the university, I will have more data of closed careers, and I will be able to train the algorithm both on the course of study and on the nationalities.

Doing so, I would be able to design a random effect model on the most common universities from which the students arrive. This will allow us to customize the intervention and supports needed by these students.

Regarding the increment of the variables, a questionnaire about students social integration could be held in the first semester of their study, so that, for the ones that appear to be less integrated into the culture and in the society, could be supported by many initiatives, as mentioned in section 7.2. These could be analysed with the other variables and it could be checked if they are relevant or not in the prediction of the performance.

Consequently, adding to the critical variables the dimension of the cultural shock, the dimension on the social integration, as well as the result of the entrance test, of the English test, more information about his/her previous studies and his/her career path, could be

enough to better and sooner predict the lower performers in term of GPA and time to degree.

A variable on the students' economic status or scholarship should be provided, in order to go deep also in this topic and check if these variables affect the final performance.

This multi-dimensional information and analysis could be used to select those students who are more likely to perform better, and support the ones who would need more help from the beginning of their careers.

Finally, a latent class analysis should be held in order to find out hidden patterns.

7.5. Managerial and Policy implications

The most important issues that arose in this discussion are the need of a centralized selection process in the university, the need of teachers who understand and are ready to work with always an increasing number of international students and the need of support staff who helps international students with their daily issues.

Focusing on the first issue, a centralized selection process will manage all the information about international students and will homologized it with the same criteria. It will help the international students, as well as the professors, with statistical evidences about the past performance of similar peers, coming from the same country and same school. This will give the student an idea about which subject s/he will need to take more attention to or to the teachers, some points of reflections about the difficulties their students might have in understanding their subject. Some highlights about the future performance, basing on the grades the student will get in each semester should be provided in order to let him/her think about if s/he is in line with the desired results. Customized suggestions will be given to each student according to the different previous paths. To allow this, data will need to be constantly updated in the database, so that, new algorithms could be done easily and faster. Finally, the selection process should be done in the same way for each student, thanks to the usage of machine learning algorithms that will allow the university to give the same scores to the most similar students.

Regarding the second issue, teachers need to support the academic path of their students, as already discussed in the chapter 7.2., by giving them the prerequisites needed for the courses and the support they need to succeed. Moreover, they can support the "globalization" of the teamwork stating that the projects need to be done at least with 4 international students per group, so that they can integrate with the domestic students, as well as with the other international ones.

Finally, the support staff is needed to organize welcome activities and events during the all year to let the students create a community. Furthermore, it needs to advertise psychological supports for those students who more feel homesickness and who feel alone in the new context of their studies.

7.6. Conclusion

This project contributes in the field of learning analytics in higher education, focusing on the performance of the international students.

With the lack of studies in this particular subject, the reader had the chance to understand that there is a significant gap in the performance of international and domestic students, and that distinct cultures behave differently regarding the outcomes. Indeed, from the analysis it appears that there are countries in which the GPA is very crucial but the time to graduation is not considered, others that care more on finishing on time and about the GPA.

The next step in this area could explore more the cultural significant variables which affect the performance, and try to find a way to measure them, as well as the social integration's ones, in order to implement the algorithms that consider these issues.

Regarding the academic variables, the study found out that: the higher the number of credits a student gets in his/her first year, the better will be his/her career; the higher the average number of attempts per exam in the first year means a decrease in the student's outcomes, if the lower performance's students are considered, otherwise, if the focus is on the high GPA' ones; the GPA a student has in his/her first year has always a positive impact on his/her final performance; the age of the student is critical, sometimes it appears that older students perform better than younger, sometimes it is the opposite; the gender of the student seems to impact only on Italian students' GPA; the number of years in which a student takes his/her graduation is relevant in predicting students' outcomes, it appears that the higher the number of years a student took to graduate, the lower GPA a student will have.

To improve the performance of the algorithms more careers are needed, so that they could be implemented with the random effects on the universities per each country for example, and give a clearer view on the incoming students about the outcomes their similar peers had got. Moreover, in order to apply the multidimensional model, new variables should be created and measured to keep track of every factor that could affect international students' performance. But on the top of that, an entrance test that homogenize every single student coming from every nation is needed, both to predict their outcomes and take actions on them, and to try to find a better entrance requirement that would be used as a filter of students on the enrolment in the university.

The knowledge about international students in higher education should be spread around in order to make domestic students, teachers and academic staff understand their issues when attending a university far from their country of origin. This awareness would be critical to make people realize that it is not easy as it seems to deal with a new nation.

To allow this, a centralized department in the university should be created to select and to support international students in their stays with the help of the results of the machine learning algorithms: psychological supports, welcoming activities, courses for teachers who deal with foreign students, courses for academic staff to support the students should be provided in order to create a more cosmopolitan and open environment in which they would be able to study better and to be more integrated.

To conclude, according to the article "Train 'em up. Kick 'em out" of The Economist: "[…] The question of whether to welcome foreign students ought to be much easier. They more than pay their way. They add to the host country's collective brainpower. And they are easy to assimilate. Indeed, for ageing rich countries seeking to import young workers to plug skills gaps and prop up wobbly pension systems, they are ideal. A foreign graduate from a local university is likely to be well-qualified, fluent in the local lingo and at ease with local customs. Countries should be vying to attract such people. […]

Welcoming foreign students is a policy that costs less than nothing in the short term and brings huge rewards in the long term. Hence the bafflement of James Dyson, a billionaire inventor, who summed up Britain's policy thus: "Train 'em up. Kick 'em out. It's a bit shortsighted, isn't it?" (The Economist, 2016).

According to this, universities and policy makers should not only invest in better their performance, but also try to find them an attractive career in the country.

To conclude, this thesis contribuetes in the field of learning analytics in solving the following issues:

1. The identification of criteria to support the selection of the incoming students,
2. The identification of criteria to recognise the more needing students and to support them with adequate methods,
3. The identification of effective machine learning models to predict students' performance.

References

Ackers, J. (1997) Evaluating Uk Courses: The Perspective Of The Overseas Student, In: D. Mcnamara and R. Harris (Eds) Overseas Students In Higher Education: Issues In Teaching And Learning (London, Routledge).

Al-Barrak, M. A., and Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. International Journal of Information and Education Technology, 6(7), 528.

American Speech-Language-Hearing Association. (2016). Scope of practice in speech-language pathology.

Ang, Soon, and Linn Van Dyne, (2008), "Conceptualization of cultural intelligence", in Ang, Soon, and Van Dyne, Linn,(eds.), Handbook of Cultural Intelligence: Theory, Measurement, and Applications, (pp. 3–15; M. E. Sharpe; Armonk,NY).

Araque, F., Salguero, A., Martínez, L., Navarro, E., and Calero, M. D. (2007). Data warehousing for improving web based learning sites. International Journal of Emerging Technologies in Learning, 2, 1–8.

Bandura, A. (2010). Self-efficacy. The Corsini encyclopedia of psychology, 1-3.

Benton, M., Dockendorf, L., Jin, W., Liu, Y., & Edmondson, J. A. (2007, August). The continuum of speech rhythm: computational testing of speech rhythm of large corpora from natural Chinese and English speech. In Proceedings of the 16th International Congress of Phonetic Sciences (pp. 1269-1272).

Bers, T.H. And Smith, K.E. (1991). 'Persistence Of Community College Students: The Influence Of Student Intent And Academic And Social Integration', Research In Higher Education 32(5), 539–556.

Bie, K. N. (1976) Norwegian Students At British Universities: A Case Study Of The Academic Performance Of Foreign Students, Scandinavian Journal Of Educational Research, 20(1), 1–24.

Black, J. Stewart, Mark Mendenhall, and Gary Oddou, (1991), "Toward a comprehensive model of international adjustment: An integration of multiple theoretical perspectives", Academy of Management Review 16, 291–317.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. Wadsworth Int. Group, 37(15), 237-251.

Bridgeman, B., Cho, Y., and DiPietro, S. (2015). Predicting grades from an English language assessment: The importance of peeling the onion. Language Testing, 33(3), 307–318. doi:10.1177/0265532215583066

Briggs, P., & Ammigan, R. (2017). A collaborative programming and outreach model for international student support offices. Journal of International Students, 7(4), 1080-1095.

Burns, R. (1991). Study And Stress Among ® Rst Year Overseas Students In An Australian University. Higher Education Research and Development, 10(1), 61± 77.

Cambruzzi, W. L., Rigo, S. J., and Barbosa, J. L. (2015). Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach. Journal of Universal Computer Science, 21(1),23–47.

Carvajal, C. M., González, J. A., Tassara, C. A., and Álvarez, M. S. (2018). Sobre-duración: una Aproximación Cuantitativa. Formación universitaria, 11(3), 19-28.

Chalmers, D. and Volet, S. (1997) 'Common Misconceptions About Students From South-East Asia Studying In Australia', Higher Education Research And Development 16 (1): 87–98.

Chen, H., Chiang, R. H. and Storey, V. C. 2012. Business Intelligence And Analytics: From Big Data To Big Impact. *Mis Quarterly*, 36**,** 1165-1188.

Coleman, J. S. (1961). The adolescent society: Academic achievement and the structure of competition. New York: Free Press of Glencoe.

Cooper, A. 2012. A Brief History Of Analytics, Jisc Cetis Analytics Series 1 (9). *University Of Bolton*.

Cope, R. G. "Are Students More Likely to Drop Out of Large Colleges?" College Student journal, 1968, 92-97.

Cox, P.L., E.D. Schmitt, P.E. Bobrowski, and G. Graham. 2005. Enhancing the first-year experience for business students: Student retention and academic success. Journal of Behavioural and Applied Management 7: 40–68.

Crawford, I. and Wang, Z. (2014), "Why are first year accounting studies inclusive?", Accounting and Finance, Vol. 54 No. 2, pp. 419-439.

Crowne, Kerri Anne, (2008), "What leads to cultural intelligence?", Business Horizons 51, 391–399.

Daniel, B. 2015. Big Data And Analytics In Higher Education: Opportunities And Challenges. *British Journal Of Educational Technology*.

Davenport, T. H. and Harris, J. G. 2007. *Competing On Analytics: The New Science Of Winning*, Harvard Business Press.

Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura. (Eds), *Proceedings of the 2nd International Conference on Educational Data Mining (pp. 41–50)*. Retrieved from http://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf

Del Blanco, Á., Serrano, Á., Freire, M., Martínez-Ortiz, I. and Fernández-Manjón, B. E-Learning Standards And Learning Analytics. Can Data Collection Be Improved By Using Standard Data Models? Global Engineering Education Conference (Educon), 2013 Ieee, 2013. Ieee, 1255-1261.

DeLoach, Steven B., Mark R. Kurt, and Neal H. Olitsky, (2015), "Does content matter? Analyzing the change in global awareness between business and non-business-focused short-term study abroad courses", Journal of Teaching in International Business 26, 4–31.

De Vita, G. (2002) Cultural Equivalence In The Assessment Of Home And International Business Management Students: A Uk Exploratory Study, Studies In Higher Education, 27, 221–231.

Durkheim, E. (1951). Suicide, Translated By J.A. Spaulding And G. Simpson, Glencoe: The Free Press. Originally Published As Le Suicide: Etude De Sociology, Paris: Felix Alcan, 1897.

Earley, P. C., and Ang, S. (2003). Cultural Intelligence: Individual Interactions Across Cultures. Palo Alto, Ca: Stanford University Press.

Eduventures. (2013). Predictive analytics in higher education: Data driven decision-making for the student life cycle.

Eisenberg, J., Lee, H.-J., Bruck, F., Brenne, B., Claes, M.-T., Mironski, J., and Bell, R. (2013). Can Business schools make students culturally competent? Effects of cross-cultural management courses on cultural intelligence. Learning and Education, 12(4), 603–621.

Elias, S. M., and Loomis, R. J. (2002). Utilizing Need For Cognition And Perceived Self-Efficacy To Predict Academic Performance 1. Journal Of Applied Social Psychology, 32(8), 1687-1702.

Engle, R., and Crowne, K. A. (2014). The impact of international experience on cultural intelligence: An application of contact theory in a structured short-term programme. Human Resource Development International, 17(1), 30–46.

Engle, R., and Nash, B. (2016). Foreign travel experiences and cultural intelligence: Does country choice matter? Journal of Teaching in International Business, 27(1), 23–40.

Ferguson, R., and Clow, D. (2017). Where Is The Evidence? A Call To Action For Learning Analytics. Proceedings Of The Seventh International Learning Analytics and Knowledge Conference (Pp. 56–65). Acm.

Ferguson, R. and Shum, S. B. Social Learning Analytics: Five Approaches. Proceedings Of The 2nd International Conference On Learning Analytics And Knowledge, 2012. Acm, 23-33.

Fleming, L., Engerman, K., and Williams, D. (2006). Why students leave engineering: The unexpected bond. In Proceedings of the 2006 American Society for Engineering Education Conferenceand Exposition.

Fontana, L., Masci, C., Ieva, F., and Paganoni, A. M. (2018). Performing Learning Analytics via Generalized Mixed-E ects Trees.

Fozdar, F., & Volet, S. (2012). Intercultural learning among community development students: positive attitudes, ambivalent experiences. Community Development, 43(3), 361–378.

Fritz, J. (2011). Classroom walls that talk: Using online course activity data of successful students to raise self-awareness of underperforming peers. The Internet and Higher Education, 14(2), 89–97.

Furnham, A., and Alibhai, N. (1985). The friendship networks of foreign students: A replication and extension of the functional model. International Journal of Psychology, 20(6), 709.

Gasevic, D., Zouaq, A., & Janzen, R. (2013). "Choose your classmates, your GPA is at stake!": the association of cross-class social ties and academic performance. American Behavioural Scientist, 57, 1460–1479.

Gilmore, D. (2014). Goffman's Front Stage And Backstage Behaviours In Online Education. Journal Of Learning Analytics, 1(3), 187–190.

Golding, P., and Donaldson, O. (2006, October). Predicting academic performance. In Proceedings. Frontiers in Education. 36th Annual Conference (pp. 21-26). IEEE.

Graunke, S.S., And S.A. Woosley. 2005. An Exploration Of The Factors That Affect The Academic Success Of College Sophomores. College Student Journal 39: 367–76.

Greller, W. and Drachsler, H. 2012. Translating Learning Into Numbers: A Generic Framework For Learning Analytics. *Journal Of Educational Technology and Society*, 15**,** 42.

Grönlund, Å., and Andersson, A. (2006). E-Gov Research Quality Improvements Since 2003: More Rigor, But Research (Perhaps) Redefined. Proceedings Of 5th International Conference, Egov 2005. Krakow, Poland. Heidelberg, Germany: Springer.

Guney, Y., 2009, Exogenous and endogenous factors influencing students' performance in undergraduate accounting modules, Accounting Education: An International Journal 18, 51–73.

Guo, W. W. (2010). Incorporating statistical and neural network approaches for student course satisfaction analysis and prediction. Expert Systems with Applications, 37(4), 3358–3365.

Häkkinen I. "Do university entrance exams predict academic achievement?" Communications of Department of Economics Working paper Uppsala University October 2004 (unpublished)

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. Computers & education, 80, 152-161.

Harrison, N., & Peacock, N. (2009). Cultural distance, mindfulness and passive xenophobia: using integrated threat theory to explore home higher education students' perspectives on 'internationalisation at home'. British Educational Research Journal, 36(6), 2009.

Hau, K. T., and Salili, F. (1996). Prediction Of Academic Performance Among Chinese Students: Effort Can Compensate For Lack Of Ability. Organizational Behaviour And Human Decision Processes, 65, 83-94.

He, Xiaohong, Mohammad Elahee, Robert Engle, Chadwick Nehrt, and Farid Sadrieh, (2007), "Globalization and International Business", (NorthCoast Publishers; Garfield Heights, OH).

Hill, K., Storch, N., and Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. In R. Tulloh (Ed.), IELTS research reports (Vol. 2, pp. 62–73). Canberra: IELTS Australia.

Hoffman, M., J. Richmond, J. Morrow, and K. Salomone. 2002. Investigating 'sense of belonging' in first-year college students. Journal of College Student Retention 4: 227–56.

Hoffman, L., and Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. Behaviour Research Methods, 39(1), 101-117.

Hofstede, Geert, (2001). "Culture's Consequences", (2nd ed.; Sage: Thousand Oaks, CA).

Hommes, J., Rienties, B., de Grave, W., Bos, G., Schuwirth, L., & Scherpbier, A. (2012). Visualising the invisible. A network approach to reveal the informal social side of student learning. Advances in Health Sciences Education, 17(5), 743–757

Iskhakova, M. (2018). Does Cross-Cultural Competence Matter When Going Global: Cultural Intelligence And Its Impact On Performance Of International Students In Australia. Journal Of Intercultural Communication Research, 47(2), 121-140.

Jacob, E. J., and Greggo, J. W. (2001). Using Counsellor Training And Collaborative Programming Strategies In Working With International Students. Journal Of Multicultural Counseling And Development, 29, 73-88.

Juillerat, S. 2000. Assessing The Expectations And Satisfaction Of Sophomores. In Visible Solutions For Invisible Students: Helping Sophomores Succeed, Ed. L.A. Schreiner And J.

Jochems, W., Snippe, J., Smid, H. J. and Verweij, A. (1996) The Academic Progress Of Foreign Students: Study Achievement And Study Behaviour, Higher Education, 31, 325–340.

Johnson, G.M. (1996). 'Faculty Differences In University Attrition: A Comparison Of The Characteristics Of Arts, Education And Science Students Who Withdraw From Undergraduate Programs', Journal Of Higher Education Policy And Management 18(1), 75–91.

Kaisler, S., Armour, F., Espinosa, J. A. and Money, W. Big Data: Issues And Challenges Moving Forward. System Sciences (Hicss), 2013 46th Hawaii International Conference On, 2013. Ieee, 995-1004.

Katal, A., Wazid, M. and Goudar, R. Big Data: Issues, Challenges, Tools And Good Practices. Contemporary Computing (Ic3), 2013 Sixth International Conference On, 2013. Ieee, 404-409.

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R., and Hatala, M. M. (2015). Penetrating The Black Box Of Time-On-Task Estimation. Proceedings Of The Fifth International Conference On Learning Analytics And Knowledge (Pp. 184–193). Acm.

Kerstjens, M., and Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. In R. Tulloh (Ed.), IELTS research reports (Vol. 3, pp. 86–108). Canberra: IELTS Australia.

Kizilcec, R. F., Piech, C., and Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massiveopen online courses. In D. Suthers, K. Verbert, E. Duval, and X. Ochoa (Eds.), Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (pp. 170–179), New York, NY: ACM.

Koh, M. Y., and H. C. Koh, 1999, The determinants of performance in an accountancy degree course, Accounting Education: An International Journal 8, 13–29.

Lang, C., Siemens, G., Wise, A., and Gasevic, D. (Eds.). (2017). Handbook of learning analytics. SOLAR, Society for Learning Analytics and Research.

Leask, B. (2003). Venturing into the unknown—A framework and strategies to assist international and Australian students to learn from each other. In C. Bond and P. Bright (Eds.), Research and development in higher education: Learning for an unknown future (Vol. 26, pp. 380-387). Christchurch, New Zealand: Higher Education Research and Development Society of Australasia Inc.

Li, G., Chen, W., and Duanmu, J. L. (2010). Determinants of international students' academic performance: A comparison between Chinese and other international students. Journal of Studies in International Education, 14(4), 389-405.

Lu, O. H. T., Huang, J. C. H., Huang, A. Y. Q., and Yang, S. J. H. (2017). Applying learning analytics for improving students' engagement and learning outcomes in an MOOCs enabled collaborative programming course. Interactive Learning Environments, 25(2), 220–234.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., and Loumos, V. (2009a). Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers and Education, 53(3), 950–965.

Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., and Loumos, V. (2009b). Early and dynamic student achievement prediction in e-learning courses using neural networks. Journal of the American Society for Information Science and Technology, 60(2), 372–380.

Mah, D.-K. (2016). Learning Analytics And Digital Badges: Potential Impact On Student Retention In Higher Education. Technology, Knowledge And Learning, 21(3), 285–305.

Makepeace, E. and Baxter, A. (1990) Overseas Students And Examination Failure: A National Study, Journal Of International Education, 1(1), 36–48.

Mannan, M. A. (2007). Student Attrition And Academic And Social Integration: Application Of Tinto's Model At The University Of Papua New Guinea. *Higher Education*, *53*(2), 147-165.

Marshall, P. and Chilton, E. H. S. (1995) Singaporean Students In British Higher Education: The Statistics Of Success, Engineering Science And Education Journal, August, 155–160.

Matsuda, A. (Ed.). (2012). Principles and practices of teaching English as an international language (Vol. 25). Multilingual Matters.

Medved, D., Franco, A., Gao, X., & Yang, F. (2013). Challenges in teaching international students: Group separation, language barriers and culture differences. Lund, Sweden: Lund University.

Meeuwisse, M., Severiens, S. E., and Born, M. P. (2010). Reasons for withdrawal from higher vocational education. A comparison of ethnic minority and majority non-completers. Studies in Higher Education, 35(1), 93–111.

Merceron, A. (2015). Educational Data Mining/Learning Analytics: Methods, Tasks And Current Trends. Proceedings Of Delfi Workshops 2015 Co-Located With 13th E-Learning Conference Of The German Computer Society, München, Germany, September 1. Retrieved From Https://Pdfs.Semanticscholar.Org/1d3a/De2c0a5a60be82030616b99ebd8426238098.Pdf 2018-05-17

Metzger, B. S., and Bean, J. P. (1987). The estimation of a conceptual model of non-traditional undergraduate student attrition. Research in Higher Education 27(1): 15--38.

Milliron, M. D., Malcolm, L., and Kil, D. (2014). Insight and action analytics: Three case studies to consider. Research and Practice in Assessment, 9, 70–89.

Misra, R., Crist, M., and Burant, C. J. (2003). Relationships Among Life Stress, Social Support, Academic Stressors, And Reactions To Stressors Of International Students In The United States. International Journal Of Stress Management, 10, 137-157.

Mittelmeier, J., Rienties, B., Tempelaar, D., & Whitelock, D. (2018). Overcoming cross-cultural group work tensions: mixed student perspectives on the role of social relationships. Higher Education, 75(1), 149-166.

Moon, Hyoung Koo, Byoung Kwon Choi, and Jae Shik Jung, (2012), "Previous international experience, cross-cultural training, and expatriates' cross-cultural adjustment: Effects of cultural intelligence and goal orientation", Human Resource Development Quarterly 23 (3), 285–330.

Moore, P., & Hampton, G. (2015). 'It's a bit of a generalisation, but...': participant perspectives on intercultural group assessment in higher education. Assessment & Evaluation in Higher Education, 40(3), 390–406.

Morrison, J., Merrick, B., Higgs, S., and Le Métais, J. (2005). Researching The Performance Of International Students In The Uk. Studies In Higher Education, 30(3), 327-337.

Munro, B.H (1981). 'Dropouts From Higher Education: Path Analysis Of A National Sample', American Educational Research Journal 18(2), 133–141.

Neumann, H., Padden, N., and McDonough, K. (2019). Beyond English language proficiency scores: understanding the academic performance of international undergraduate students during the first year of study. Higher Education Research and Development, 38(2), 324-338.

New Zealand Ministry Of Education. (2004). The Experience Of International Students In New Zealand (Prepared By C. Ward And A. Masgoret, Victoria University Of Wellington). Wellington, New Zealand: Author.

Nistor, N., Baltes, B., Dascălu, M., Mihăilă, D., Smeaton, G., and Trăuşan-Matu, Ş. (2014). Participation In Virtual Academic Communities Of Practice Under The Influence Of Technology Acceptance And Community Factors. A Learning Analytics Application. Computers In Human Behaviour, 34, 339–344.

Öckert, B. (2001): Does pre-university background matter? In Effects of higher education and the role of admission selection, Dissertation series no. 52, Swedish Institute for Social Research, Stockholm University.

Osland, Joyce, and Asbjorn Osland, (2005), "Expatriate paradoxes and cultural involvement", International Studies of Management and Organization 35 (4), 91–114.

Özütürk, G., & Hürsen, Ç. (2013). Determination of English language learning anxiety in efl classrooms. Procedia-Social and Behavioral Sciences, 84, 1899-1907.

Pascarella, E.T., Smart, J.C. And Ethington, C.A. (1986). 'Long-Term Persistence Of Two Year College Students', Research In Higher Education 24(1), 47–71.

Pascarella, E. T., and Terenzini, P.T. (2005). How College Affects Students Revisited: A Third Decade of Research. San Francisco, CA: Jossey-Bass.

Pauley, G. F. (1988) Overseas-Fee-Paying Students At Curtin University Of Technology: A Report On Their Progress After One Year (Perth, Curtin University Of Technology).

Prinsloo, P., and Slade, S. (2017). An Elephant In The Learning Analytics Room: The Obligation To Act. Proceedings Of The Seventh International Learning Analytics and Knowledge Conference (Pp. 46–55). Acm.

Pun, A.S. (1990) 'Managing The Cultural Differences In Learning', Journal Of Management Development 9 (5): 35–40.

Rantanen, P. (2001): Valintakoe vai ei? Ammatillisen koulutuksen ja ammattikorkeakoulujen opiskelijavalinnan tarkastelua. Koulutus- ja tiedepolitiikan osaston julkaisusarja 83. Ministry of Education.

Rienties, B., Beausaert, S., Grohnert, T., Niemantsverdriet, S., and Kommers, P. (2012). Understanding academic performance of international students: the role of ethnicity, academic and social integration. Higher education, 63(6), 685-700.

Robertson, M., Line, M., Jones, S., and Thomas, S. (2000). International Students, Learning Environments And Perceptions: A Case Study Using The Delphi Technique. Higher Education Research And Development, 19, 89-102.

Rockstuhl, T., and Ng, K. Y. (2008). The Effects Of Cultural Intelligence On Interpersonal Trust In Multicultural Teams. In S. Ang and L. V. Dyne (Eds.), Handbook Of Cultural Intelligence: Theory, Measurement, And Applications (Pp. 206–220). Ny: M.E. Sharpe.

Romero-Zaldivar, V-A., Pardo, A., Burgos, D., and Kloos, C. D. (2012). Monitoring student progress using virtual appliances: a case study. Computers and Education, 58(4), 1058–1067.

Ronen, Simcha, and Oded Shenkar, (2013), "Mapping world cultures: Cluster formation, sources and implications",Journal of International Business Studies 44, 867–897.

Rose, G. (2005). Group Differences In Graduate Students' Concepts Of The Ideal Mentor. Research In Higher Education, 46, 53-80.

Rubel, A., and Jones, K. (2016). Student Privacy In Learning Analytics: An Information Ethics Perspective. The Information Society, 32(2), 143–159.

Russell, J., Rosenthal, D., and Thomson, G. (2010). The international student experience: Three styles of adaptation. Higher Education, 60(2), 235–249.

Schneider, L. J., and Spinler, D. G. (1986). Help-Giver Preference Patterns in American and International Asian Students.

Schumacher, C., and Ifenthaler, D. (2018). Features Students Really Expect From Learning Analytics. Computers In Human Behaviour, 7, 397–407.

Scot, C., Burns, A. And Cooney, G. (1996). 'Reasons For Discontinuing Study: The Case Of Mature Age Female Students With Children', Higher Education 31, 233–253.

Searle, W., and Ward, C. (1990). The Prediction Of Psychological And Sociocultural Adjustment During Cross-Cultural Transitions. International Journal Of Intercultural Relations, 14, 449-464.

Sela, R. J., and Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. Machine learning, 86(2), 169-207.

Severiens, S., and Wolff, R. (2008). A comparison of ethnic minority and majority students: Social and academic integration, and quality of learning. Studies in Higher Education, 33, 253–266.

Siemens, G., and Baker, R. (2012). Learning Analytics And Educational Data Mining: Towards Communication And Collaboration. Proceedings Of The Second International Conference On Learning Analytics and Knowledge (Pp. 252–254). Acm.

Siemens, G., and Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE review, 46(5), 30.

Slade, S., and Prinsloo, P. (2013). Learning Analytics: Ethical Issues And Dilemmas. American Behavioural Scientist, 57(10), 1510–1529.

Smith, R. A., & Khawaja, N. G. (2011). A review of the acculturation experiences of international students. International Journal of intercultural relations, 35(6), 699-713.

Smith, P. J., and Smith, S. N. (1999). Differences Between Chinese And Australian Students: Some Implications For Distance Educators. Distance Education, 20, 64-80.

Spady, W.G. (1970). 'Dropouts From Higher Education: An Interdisciplinary Review And Synthesis', Interchange 1, 64–85.

Tessema, M., Ready, K., and Malone, C. (2012). Effect of Gender on College Students' satisfaction and Achievement. International Journal of Business and Social Science, 3(10), 1-11

Tinto, V. (1993). Leaving College: Rethinking The Causes And Cures Of Student Attrition. 2nd Edition, Chicago: University Of Chicago Press.

Train 'em up. Kick 'em out. (2016, Jan 30). The Econimist.

Trompenaars, Fons, and Charles Hampden-Turner, (1998), "Riding the Waves of Culture", (McGraw-Hill: New York,NY).

Tsai, Y. S., Moreno-Marcos, P. M., Tammets, K., Kollom, K., and Gašević, D. (2018). Sheila Policy Framework: Informing Institutional Strategies And Policy Processes Of Learning Analytics. Proceedings Of The 8th International Conference On Learning Analytics And Knowledge (Pp. 320–329). Acm.

Turner, K. and George, M. (2011), "Hybrid teaching and learning in contemporary English teacher education", paper presented at the Society for Information Technology & Teacher Education International Conference 2011, Nashville, TN.

Viberg, O., Hatakka, M., Bälter, O., and Mavroudi, A. (2018). The current landscape of learning analytics in higher education. Computers in Human Behaviour.

Van Nelson, C., Nelson, J. S., and Malone, B. G. (2004). Predicting success of international graduate students in an American university. College and University, 80(1), 19–27.

Wadia-Fascetti, S., & Leventman, P. G. (2000). E-Mentoring: A Longitudinal Approach to Mentoring Relationships for Women Pursuing Technical Careers. Journal of Engineering Education, 89(3), 295-300.

Wait, I. W., and Gressel, J. W. (2009). Relationship between TOEFL score and academic success for international engineering students. Journal of Engineering Education, 98(4), 389–398. doi:10.1002/j.2168-9830.2009.tb01035.x

Wardlow, G. (1999). International Students Of Agriculture In U.S. Institutions Precursors To Academic Success. Journal Of Agricultural Education, 30, 17-22.

Watkins, D. A., and Biggs, J. B. (1996). The Chinese Learner: Cultural, Psychological, And Contextual Influences. Hong Kong: Comparative Education Research Centre, University Of Hong Kong.

Webb, V., and Read, J. (2000, September). The Challenge Of Cultural Heterogeneity To Educational Development. Paper Presented At Laoiii Conference, Pretoria, South Africa.

Wilcox, P., Winn, S., and Fyvie-Gauld, M. (2005). It was nothing to do with the university, it was just the people: The role of social support in the first-year experience of higher education. Studies in Higher Education, 30(6), 707–722.

Willcoxson, L., Cotter, J., and Joy, S. (2011). Beyond The First-Year Experience: The Impact On Attrition Of Student Experiences Throughout Undergraduate Degree Studies In Six Diverse Universities. Studies In Higher Education, 36(3), 331-352.

Williams, Sir B. (Ed.) (1989) Overseas Students In Australia: Policy And Practice (Canberra, International Development Program Of Australian Universities And Colleges).

Williamson, B. (2018). The hidden architecture of higher education: building a big data infrastructure for the 'smarter university'. International Journal of Educational Technology in Higher Education, 15(1), 12.

Wright, T. and Cochrane, R. (2000) Factors Influencing Successful Submission Of Phd Theses, Studies In Higher Education, 25(2), 181–195.

Yen, D. A., and Kuzma, J. (2009). Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students. Worcester Journal of Learning and Teaching, 3, 1–7.

Yizar Jr, J. H. (2010). Enrollment Factors That Predict Persistence Of At-Risk (Low Income And First Generation) Students' Journey Towards Completion Of A Baccalaureate Degree At Idaho State University. Idaho State University.

Zajacova, A., Lynch, S. M., and Espenshade, T. J. (2005). Self-Efficacy, Stress, And Academic Success In College. Research In Higher Education, 46(6), 677-706.

Zhang, N., and Dixon, D. N. (2003). Acculturation and attitudes of Asian international students toward seeking psychological help. Journal of Multicultural Counseling and Development, 31(3), 205-222.

Zhou, G., & Zhang, Z. (2014). A study of the first year international students at a Canadian university: Challenges and experiences with social integration. Comparative and International Education/Éducation Comparée et Internationale, 43(2), 7.

Annex

Data pre processing's codes

```
setwd("C:/Users/margh/Desktop/tesi/dati finali")
atenei = read.csv("atenei.csv")
attivita_formative = read.csv("attivita-formative.csv")
carriere = read.csv("carriere.csv")
cds_tipo = read.csv("cds_tipo.csv", sep=',')
comuni = read.csv("comuni.csv")
# indice delle province
corsi_laurea = read.csv("corsi_laurea.csv", sep =',')
# indice ID dei corsi di studio
domini = read.csv("domini.csv")
domini_valori = read.csv("domini_valori.csv")
# indici di carriera
frequenze = read.csv("frequenze.csv")
indir_res = read.csv("indir_res.csv")
# provincia residenza
mobilita = read.csv("mobilita.csv")
# mobilità all'estero dello studente
nazioni = read.csv("nazioni.csv")
persone = read.csv("persone.csv")
province = read.csv("province.csv")
qd_classi = read.csv("qd-classi.csv")
tit_acc_prec = read.csv("tit_acc_prec.csv")
##
tit_med_prec = read.csv("tit_med_prec.csv")
##
tit_med_tp = read.csv("tit_med_tp.csv")
# liceo or previous school
####################################
# STUDENT EXPLAINATORY INFORMATION #
####################################
########################################
# merge di carriere e persone
carriere[which(carriere$CARR_FIN_AA==17),10]=2017
carriere[which(carriere$CARR_FIN_AA==2008),10]=NA
boxplot(carriere$CARR_FIN_AA,na.rm=T)
## DIVIDIAMO SUBITO INGEGNERIA DA DESIGN DA ARCHITETTURA
id_architettura = corsi_laurea$CDS_ID[corsi_laurea$CDS_POLI_EDU_FLD=='A']
id_design = corsi_laurea$CDS_ID[corsi_laurea$CDS_POLI_EDU_FLD=='D']
id_ingegneria = corsi_laurea$CDS_ID[corsi_laurea$CDS_POLI_EDU_FLD=='I']
ingegneria = carriere[carriere$IMM_CDS_ID %in% id_ingegneria,]
dim(ingegneria) # 91,329 carriere
architettura = carriere[carriere$IMM_CDS_ID %in% id_architettura,]
dim(architettura) # 28,292 carriere
design = carriere[carriere$IMM_CDS_ID %in% id_design,]
dim(design) # 15,503 carriere
## DIFFERENZIAMO LE TRIENNALI DALLE MAGISTRALI DAI CICLI UNICI
table(corsi_laurea$CDS_TIPO_DN)
bachelor_id = corsi_laurea$CDS_ID[which(corsi_laurea$CDS_TIPO_DN=='Laurea Di Primo
Livello')]
magistrale_id = corsi_laurea$CDS_ID[which(corsi_laurea$CDS_TIPO_DN=='Laurea
Magistrale'|corsi_laurea$CDS_TIPO_DN=='Laurea Specialistica')]
#special_id = corsi_laurea$CDS_ID[which(corsi_laurea$CDS_TIPO_DN=='Laurea
Specialistica')]
special_unico_id = corsi_laurea$CDS_ID[which(corsi_laurea$CDS_TIPO_DN=='Laurea
Specialistica A Ciclo Unico')]
length(bachelor_id)
length(magistrale_id)
#length(special_id)
length(special_unico_id)
somma1 = 0
somma2 = 0
somma3 = 0
somma4 = 0
for(i in 1:length(carriere$IMM_CDS_ID)){
somma1 = somma1 + is.element(el=carriere$IMM_CDS_ID[i], set=bachelor_id)
somma2 = somma2 + is.element(el=carriere$IMM_CDS_ID[i], set=magistrale_id)
# somma3 = somma3 + is.element(el=carriere$IMM_CDS_ID[i], set=special_id)
somma4 = somma4 + is.element(el=carriere$IMM_CDS_ID[i], set=special_unico_id)
}s
omma1 ## 78113
somma2 ## 55592
```

```
somma3 ## 138
somma4 ## 1281
## estraggo solo le carriere delle MAGISTRALI di INGEGNERIA
magistrali = carriere[carriere$IMM_CDS_ID %in% magistrale_id,]
dim(magistrali)
# quante di ing, design e architett?
mag_ing = ingegneria[ingegneria$IMM_CDS_ID %in% magistrale_id,]
dim(mag_ing) #
mag_arch = architettura[architettura$IMM_CDS_ID %in% magistrale_id,]
dim(mag_arch) #
mag_des = design[design$IMM_CDS_ID %in% magistrale_id,]
dim(mag_des) #
length(unique(magistrali$PERS_AN_ID)) #
aa = names(which(table(as.factor(paste(magistrali$PERS_AN_ID)))==1))
length(aa) #
b = names(which(table(as.factor(paste(magistrali$PERS_AN_ID)))==2))
length(b) #
magistrali_doppie = magistrali[magistrali$PERS_AN_ID %in% b,] # 2,140
#
#####DA QUI!
c = names(which(table(as.factor(paste(magistrali$PERS_AN_ID)))==3))
head(c)
magistrali_triple = magistrali[magistrali$PERS_AN_ID %in% c,] #
d = names(which(table(as.factor(paste(magistrali$PERS_AN_ID)))==4))
head(d)
magistrali_quadruple = magistrali[magistrali$PERS_AN_ID %in% d,] # 12
#########################################################################
ab = rep(0, length(aa)+ length(b))
for(i in 1:length(aa)){
ab[i]=aa[i]
}f
or(i in (length(aa)+1):(length(b)+length(aa))){
ab[i]=b[i-length(aa)]
}h
ead(ab)
sum(ab==0)
# triennali con carriere singole o al massimo doppie. Cosa facciamo con le doppie?
triennali_n = triennali[triennali$PERS_AN_ID %in% ab, ] #
sum(triennali_doppie$CARR_FLST=='L') #
sum(triennali_doppie$CARR_FLST=='D') #
sum(triennali_doppie$CARR_FLST=='A') #
# ordiniamo per codice persona
ord=triennali_doppie[order(triennali_doppie$PERS_AN_ID),]
## non so come analizzare queste doppie carriere.
# comunque, in quanti hanno cambiato corso di studi?
dim(triennali_n[triennali_n$IMM_CDS_ID!=triennali_n$UIS_CDS_ID,])
magistrali_singole=magistrali[magistrali$PERS_AN_ID %in% aa,]
dim(magistrali_singole[magistrali_singole$IMM_CDS_ID!=magistrali_singole$UIS_CDS_ID,])
table(corsi_laurea$CDS_POLI_EDU_FLD)
## prendo solo le magistrali singole di ingegneria
ingegneria = corsi_laurea$CDS_ID[corsi_laurea$CDS_POLI_EDU_FLD=='I']
length(ingegneria)
design = corsi_laurea$CDS_ID[corsi_laurea$CDS_POLI_EDU_FLD=='D']
length(design)
architettura = corsi_laurea$CDS_ID[corsi_laurea$CDS_POLI_EDU_FLD=='A']
length(architettura)
mag_sing_ing = magistrali_singole[magistrali_singole$IMM_CDS_ID %in% ingegneria,]
dim(mag_sing_ing) # quindi sono tutti di ingegneria
mag_sing_des = magistrali_singole[magistrali_singole$IMM_CDS_ID %in% design,]
dim(mag_sing_des) # quindi sono tutti di design
mag_sing_arch = magistrali_singole[magistrali_singole$IMM_CDS_ID %in% architettura,]
dim(mag_sing_arch) # , quindi sono tutti di architettura
####solo magistrali
magistrali = carriere[carriere$IMM_CDS_ID %in% magistrale_id,]
dim(magistrali) #
length(unique(magistrali$PERS_AN_ID)) #
aa = names(which(table(as.factor(paste(magistrali$PERS_AN_ID)))==1))
length(aa)
magistrali_singole=magistrali[magistrali$PERS_AN_ID %in% aa,]
magi_sing_ing = magistrali_singole[magistrali_singole$IMM_CDS_ID %in% ingegneria,]
dim(magi_sing_ing) # , quindi sono tutti di ingegneria
magi_sing_des = magistrali_singole[magistrali_singole$IMM_CDS_ID %in% design,]
dim(magi_sing_des) # quindi sono tutti di design
magi_sing_arch = magistrali_singole[magistrali_singole$IMM_CDS_ID %in% architettura,]
dim(magi_sing_arch) # , quindi sono tutti di architettura
corsi_mag = unique(magistrali_singole$IMM_CDS_ID)
corsi_laurea[corsi_laurea$CDS_ID %in% corsi_mag,]
table(mag_sing_ing$IMM_CDS_ID)
barplot(table(mag_sing_ing$IMM_CDS_ID)) # a pochissimi (se taglio a 1000 ne ho circa 16)
```

```
write.table(mag_sing_ing,file='mag_sing_ing.txt')
write.table(mag_sing_arch,file='mag_sing_arch.txt')
write.table(mag_sing_des,file='mag_sing_des.txt')

################################################################################
########## creo dataset per le analisi delle magistrali di ingegneria ##########
################################################################################
library(dplyr)
library(plyr)
dati = read.table(file='mag_sing_ing.txt', header = T)
length(unique(dati$PERS_AN_ID)) #carriere magistrali e specialistiche a ingegneria
summary(dati$CARR_INI_AA)
persone = read.csv(file='persone.csv')
indir_res = read.csv("indir_res.csv")
tit_med_prec = read.csv("tit_med_prec.csv")
length(unique(persone$PERS_AN_ID)) # ho qualche doppione
righe_da_cancellare <- NULL
righe_da_tenere <- NULL
considerate <- NULL
for (i in 1:length(persone$PERS_AN_ID)) {
if (is.element(el=persone$PERS_AN_ID[i], set=considerate) ){
righe_da_cancellare <- c(righe_da_cancellare, i)
}
if (!(is.element(el=persone$PERS_AN_ID[i], set=considerate))){
considerate <- c(considerate, persone$PERS_AN_ID[i])
righe_da_tenere <- c(righe_da_tenere, i)
}
}
length(considerate)
length(righe_da_cancellare)
length(righe_da_tenere)
persone <- data.frame(persone[righe_da_tenere,])
carr_pers = merge(dati, persone, by ='PERS_AN_ID', all = F)
#############
# aggiungo residenza
righe_da_cancellare <- NULL
righe_da_tenere <- NULL
considerate <- NULL
for (i in 1:length(indir_res$PERS_AN_ID)) {
if (is.element(el=indir_res$PERS_AN_ID[i], set=considerate) ){
righe_da_cancellare <- c(righe_da_cancellare, i)
}
if (!(is.element(el=indir_res$PERS_AN_ID[i], set=considerate))){
considerate <- c(considerate, indir_res$PERS_AN_ID[i])
righe_da_tenere <- c(righe_da_tenere, i)
}
}
length(considerate)
length(righe_da_cancellare)
length(righe_da_tenere)
indir_res <- data.frame(indir_res[righe_da_tenere,])
y = merge(carr_pers, indir_res, by="PERS_AN_ID", all.x = T, suffixes = c(".x",".res"))
head(y)
length(unique(y$PERS_AN_ID)) ###
####################
# aggiungo titolo di licenza media superiore
z = merge(y, tit_med_prec, by="CARR_AN_ID", all.x = T, suffixes = c(".y",".med"))
head(z)
#############
# aggiungo titolo accademico precedente
tit_acc_prec=read.csv("tit_acc_prec.csv")
xx = merge(z, tit_acc_prec, by="CARR_AN_ID", all.x = T, suffixes = c(".z",".acc"))
head(xx)
dati2 <- data.frame(StudentID = xx$CARR_AN_ID,
YearOfBirth= xx$PERS_NAS_YYYY,
PlaceOfBirth_state= xx$NAS_STT_ID,
PlaceOfBirth_prov = xx$NAS_PRV_CD,
Sex=factor(xx$PERS_GENERE),
ResidenceCity = xx$GEO_PRV_CD.y,
AccessToStudiesAge = xx$CARR_INI_ETA,
Nationality=xx$PERS_CITT_STT_ID,
PreviousStudies =xx$TIT_TP_CD,
PreviousStudiesCenter_state =xx$GEO_STT_ID.med,
PreviousStudiesCenter_prov=xx$GEO_PRV_CD.med,
AdmissionScore= xx$STUD_AMM_VOTO,
YearStartCareer=xx$CARR_INGR_AA,
YearEndCareer=xx$CARR_FIN_AA,
Status = as.factor(xx$CARR_FLST),
StartingDegreeID = xx$IMM_CDS_ID,
```

```
FinalDegreeID = xx$UIS_CDS_ID,
DateEndCareer = xx$CARR_FIN_DT
)
#################
# aggiungo nazione di nascita
nazioni = read.csv("nazioni.csv")
head(nazioni)
head(dati2)
dati3 = merge(dati2, nazioni, by.x = "PlaceOfBirth_state", by.y = "GEO_STT_ID", all.x =
T)
colnames(dati3)[which(colnames(dati3) == "GEO_STT_DN")] = "Naz_nascita"
#################
# aggiungo nazione di studi precedenti
head(dati3)
dati4 = merge(dati3, nazioni, by.x = "PreviousStudiesCenter_state", by.y = "GEO_STT_ID",
all.x = T)
colnames(dati4)[which(colnames(dati4) == "GEO_STT_DN")] = "Naz_superiori"
#################
# aggiungo nazionalità
head(dati4)
dati5 = merge(dati4, nazioni, by.x = "Nationality", by.y = "GEO_STT_ID", all.x = T)
colnames(dati5)[which(colnames(dati5) == "GEO_STT_DN")] = "Nazionalita"
head(dati5)
#################
dati <- data.frame(StudentID = dati5$StudentID,
YearOfBirth= dati5$YearOfBirth,
PlaceOfBirth_state= dati5$Naz_nascita,
PlaceOfBirth_prov = dati5$PlaceOfBirth_prov,
Sex=dati5$Sex,
ResidenceCity = dati5$ResidenceCity,
AccessToStudiesAge = dati5$AccessToStudiesAge,
Nationality=dati5$Nazionalita,
PreviousStudies =dati5$PreviousStudies,
PreviousStudiesCenter_state =dati5$Naz_superiori,
PreviousStudiesCenter_prov=dati5$PreviousStudiesCenter_prov,
AdmissionScore= dati5$AdmissionScore,
YearStartCareer=dati5$YearStartCareer,
YearEndCareer=dati5$YearEndCareer,
Status = dati5$Status,
StartingDegreeID = dati5$StartingDegreeID,
FinalDegreeID = dati5$FinalDegreeID,
DateEndCareer = dati5$DateEndCareer
)
dati$ChangeDegree <- (xx$IMM_CDS_ID != xx$UIS_CDS_ID)
head(dati)
dati[which(dati$StudentID == 99804317),]
tt = data.frame(table(dati$StartingDegreeID))
View(tt)
hist(tt$Freq)
summary(tt$Freq)
# ci sono 23 corsi di triennale, elimino:
magistrali = as.vector(subset(tt$Var1, tt$Freq>300))
#corsi_laurea[corsi_laurea$CDS_ID %in% triennali,]
# tengo solo carriere iniziate e finite nei ? corsi selezionati
dati = subset(dati, is.element(dati$FinalDegreeID, magistrali))
dati = subset(dati, is.element(dati$StartingDegreeID, magistrali))
head(dati)
#dati$StartingDegreeID = paste("PoliMi", dati$StartingDegreeID, sep="_")
#dati$FinalDegreeID = paste("PoliMi", dati$FinalDegreeID, sep="_")
head(dati)
dim(dati)
summary(dati$ChangeDegree)
table(dati$PreviousStudies)
dati$PreviousStudies = as.vector(dati$PreviousStudies)
for( i in 1:length(dati$PreviousStudies)) {
if(is.element(dati$PreviousStudies[i],c('T ','TA', 'TC', 'TG', 'TI', 'TN'))) {
dati$PreviousStudies[i] = "Tecnica"
}
else if ( is.element(dati$PreviousStudies[i], c('S '))) {
dati$PreviousStudies[i] = "Scientifica"
}
else if ( is.element(dati$PreviousStudies[i], c('A '))) {
dati$PreviousStudies[i] = "Artistica"
}
else if ( is.element(dati$PreviousStudies[i], c('C '))) {
dati$PreviousStudies[i] = "Classica"
}
else if ( is.element(dati$PreviousStudies[i], c('L '))) {
dati$PreviousStudies[i] = "Linguistica"
```

```
}
else if ( is.element(dati$PreviousStudies[i], c('M '))) {
dati$PreviousStudies[i] = "Magistrale"
}
else if ( is.element(dati$PreviousStudies[i], c('P '))) {
dati$PreviousStudies[i] = "Professionale"
}
else if ( is.element(dati$PreviousStudies[i], c('-E'))) {
dati$PreviousStudies[i] = "Straniera"
}
else if ( is.element(dati$PreviousStudies[i], c('- '))) {
dati$PreviousStudies[i] = NA
}
}t
able(dati$PreviousStudies)
exp_info = dati[,c(1:12)]
exp_info_ord = arrange(exp_info, StudentID)
head(exp_info_ord)
write.table(exp_info_ord, file="PoliMi_student_information.txt")
View(exp_info_ord)
########################
# DEGREE INFORMATION #
########################
corsi_laurea <- read.csv('corsi_laurea.csv', sep = ",")
corsi_laurea = subset(corsi_laurea, is.element(corsi_laurea$CDS_ID, magistrali))
head(corsi_laurea)
deg_info <- data.frame( DegreeID = corsi_laurea$CDS_ID,
Institution = 'PoliMi',
DegreeNature = corsi_laurea$CDS_DN,
NumberECTS = 120,
NumberYears = 2,
Languages = 'Italian',
NumberAttemptsToEnroleSubject = 5,
NumberAttemptsToBeEvaluatedOneYear = 5,
ScoreImprovement = 'No'
)h
ead(deg_info)
write.table(deg_info, file="PoliMi_degree_information.txt")
#############################
# STUDENT CAREER INFORMATION #
#############################
head(dati)
perf1 = data.frame(
StudentID = dati$StudentID,
Status = dati$Status,
YearStartCareer = dati$YearStartCareer,
YearEndCareer = dati$YearEndCareer,
YearsToFinishDegree = dati$YearEndCareer +1 - dati$YearStartCareer,
StartingDegreeID = dati$StartingDegreeID,
FinalDegreeID = dati$FinalDegreeID,
DateEndCareer = dati$DateEndCareer,
ChangeDegree = dati$ChangeDegree
)h
ead(perf1)
# aggiungo mobilità
mobilita <- read.csv('mobilita.csv')
# mantengo solo anno inizio e fine
mobilita$SI_INI_DT <- format(as.Date(mobilita$SI_INI_DT, format = "%d/%m/%Y"), "%Y")
mobilita$SI_FIN_DT <- format(as.Date(mobilita$SI_FIN_DT, format = "%d/%m/%Y"),"%Y")
head(mobilita)
mobilita<-mobilita[-which(mobilita$SI_PRGRM_DN== "Doppia Laurea Extra Ue"),]
mobilita <- mobilita[-which(mobilita$SI_PRGRM_DN == "Doppia Laurea Ue"),]
# tengo prima mobilità per ogni carriera
mobilita <- arrange(mobilita, SI_INI_DT, SI_FLTP)
righe_da_cancellare <- NULL
righe_da_tenere <- NULL
considerate <- NULL
for (i in 1:length(mobilita$CARR_AN_ID)) {
if (is.element(el=mobilita$CARR_AN_ID[i], set=considerate) ){
righe_da_cancellare <- c(righe_da_cancellare, i)
}
if (!(is.element(el=mobilita$CARR_AN_ID[i], set=considerate))){
considerate <- c(considerate, mobilita$CARR_AN_ID[i])
righe_da_tenere <- c(righe_da_tenere, i)
}
}
length(considerate)
length(righe_da_cancellare)
length(righe_da_tenere)
```

```
mobilita <- data.frame(mobilita[righe_da_tenere,c(1,2,5)])
#merge di carriere e mobilità
perf2 <- merge(perf1, mobilita, by.x = "StudentID", by.y = 'CARR_AN_ID', all.x = T)
sum(is.na(perf2$SI_FLTP))
sum(is.na(perf2$SI_PRGRM_DN ))
levels(perf2$SI_FLTP)
head(perf2)
perf2$Mobility <- as.character(perf2$SI_PRGRM_DN)
for(i in 1:length(perf2$StudentID)){
if( is.na(perf2$Mobility[i] == "NA") ){
perf2$Mobility[i] = 'No'
}
else if (perf2$Mobility[i] == "Erasmus" & perf2$SI_FLTP[i]=='O'){
perf2$Mobility[i] = 'Erasmus outgoing'
}
else if (perf2$Mobility[i] == "Erasmus" & perf2$SI_FLTP[i]=='I'){
perf2$Mobility[i] = 'Erasmus incoming'
}
else{
perf2$Mobility[i] = 'Other'
}
}l
evels(as.factor(perf2$Mobility))
head(perf2)
perf_info = perf2[,-c(10,11)]
head(perf_info)
write.table(perf_info, file="PoliMi_career_information.txt")
######################
#### EXAMS #######
######################
frequenze <- read.csv('frequenze.csv')
attivita_formative <- read.csv('attivita-formative.csv')
qd_classi <- read.csv('qd-classi.csv')
head(frequenze)
# tengo solo ID carriera, ID materia, anno, semestre, voto, lode, n°tentativi, classe
d'esame
frequenze <- frequenze[,c(1,2,3,4,5,6,10,11)]
#tengo solo le righe di frequenze presenti in carriere
frequenze <- frequenze[which(is.element(frequenze$CARR_AN_ID, perf_info$StudentID)),]
#tengo solo frequenze con voto registrato oppure numero tentativi maggiore di zero
frequenze_confermate = subset(frequenze, frequenze$ESA_VERB_NUM > 0 |
is.na(frequenze$STUD_ACQSZ_CFU_VOTO) == F )
table(frequenze_confermate$ESA_VERB_NUM, frequenze_confermate$STUD_ACQSZ_CFU_VOTO)
#problema: ci sono ancora esami CON voto ma con nessun tentativo! Metto a questi esami 1
tentativo invece che zero?
si
for(i in 1:length(frequenze_confermate$CARR_AN_ID)){
if (frequenze_confermate$ESA_VERB_NUM[i] == 0) {
frequenze_confermate$ESA_VERB_NUM[i] = 1
}
}
head(frequenze_confermate)
freq_confermate = merge(frequenze_confermate, attivita_formative, all.x = T)
freq2 = merge(freq_confermate, qd_classi, by.x = "STUD_ATTFRM_CLAS_ID", by.y =
'QD_CLAS_ID', all.x = T)
freq_confermate = freq2
subset(freq_confermate, freq_confermate$CARR_AN_ID == 99804317)
SPEET_freq <- data.frame(
StudentID = freq_confermate$CARR_AN_ID,
SubjectID = freq_confermate$ATTFRM_CD.x,
NumberECTS = freq_confermate$ATTFRM_CFU,
KnowledgeArea = freq_confermate$ATTFRM_SSD,
Year = freq_confermate$STUD_ATTFRM_FRQ_AA,
Semester = freq_confermate$STUD_ATTFRM_FRQ_SEM,
Score = freq_confermate$STUD_ACQSZ_CFU_VOTO,
Lode = freq_confermate$STUD_ACQSZ_CFU_LODE,
NumberAttempts = freq_confermate$ESA_VERB_NUM,
AverageScoreYear = freq_confermate$QD_CLAS_ESA_VOTO_AVG,
FailureRateYear = freq_confermate$QD_CLAS_ESA_FAIL_RATE
)
subset(SPEET_freq, SPEET_freq$StudentID == 99804317)
write.table(SPEET_freq, file="PoliMi_exams.txt")

# This script builds
# 1) the file "PoliMi_career_information2.csv" that includes aggregate career info
# such as average evaluations, credits, ...
# 2) the file "model_dataframe.csv" that includes
# all careers that are selected to train & test the models (18,612 careers)
rm(list=ls())
```

```
# Set working directory
setwd("C:/Users/chiaramasci/Dropbox/Learning Analytics mio/dati2019totali/ingegneria")
require(plyr)
# Load dataset (it must be in the working directory)
degrees <- read.table('PoliMi_degree_information.txt')
names(degrees)
head(degrees)
students <- read.table('PoliMi_student_information.txt')
names(students)
head(students)
careers <- read.table('PoliMi_career_information.txt')
names(careers)
head(careers)
# if careers has 10 columns
#careers = careers[,-9]
exams <- read.table('PoliMi_exams.txt')
names(exams)
head(exams)
######################
# (1) PREPROCESSING #
######################
#####################################
# (1.1) COMPUTE AGGREGATE EXAMS INFO #
# select only passed exams
passed.exams = subset(exams, !is.na(exams$Score))
#### FULL CAREER INFORMATION ####
# compute weighted average for each student
avg.evals = ddply(passed.exams, .(StudentID),
function(x) data.frame(WeiAvgEval = weighted.mean(x$Score[x$Score != 0],
x$NumberECTS[x$Score !=
0])))
careers.new = merge(careers, avg.evals, by="StudentID", all.x = T)
careers.new$WeiAvgEval[which(is.na(careers.new$WeiAvgEval))] <- 0
careers <- careers.new
# compute numbers of passed exams for each student
ex.pass = ddply(passed.exams, .(StudentID),
function(x) data.frame(ExamsPassed = count(x$StudentID)))
colnames(ex.pass) = c("StudentID","ExamsPassed.x","ExamsPassed")
careers.new = merge(careers, ex.pass[,c("StudentID","ExamsPassed")], by="StudentID",
all.x = T)
careers.new$ExamsPassed[which(is.na(careers.new$ExamsPassed))] <- 0
careers <- careers.new
# compute average exams attempts for each student
avg.att = ddply(exams, .(StudentID),
function(x) data.frame(AvgAttempts = mean(x$NumberAttempts)))
careers.new = merge(careers, avg.att, by="StudentID", all.x = T)
careers.new$AvgAttempts[which(is.na(careers.new$AvgAttempts))] <- 0
careers <- careers.new
#### 1st SEMESTER INFORMATION ####
attach(careers)
enrolls <- data.frame(StudentID, YearStartCareer)
detach(careers)
# compute weighted average, first semester, first year, for each student
avg.evals2 = ddply(passed.exams, .(StudentID, Year, Semester),
function(x) data.frame(WeiAvgEval1.1 = weighted.mean(x$Score[x$Score != 0],
x$NumberECTS[x$Score
!= 0])))
unione = merge(avg.evals2, enrolls, by='StudentID')
avg.evals.11 = subset(unione[,c("StudentID","WeiAvgEval1.1")], (unione$Year ==
unione$YearStartCareer) &
unione$Semester == 1)
careers.new = merge(careers, avg.evals.11, by="StudentID", all.x = T)
careers.new$WeiAvgEval1.1[which(is.na(careers.new$WeiAvgEval1.1))] <- 0
careers <- careers.new
# compute average exams attempts, first semester, first year, for each student
avg.att2 = ddply(exams, .(StudentID, Year, Semester),
function(x) data.frame(AvgAtt1.1 = mean(x$NumberAttempts)))
unione = merge(avg.att2, enrolls, by='StudentID')
avg.att.11 = subset(unione[,c("StudentID","AvgAtt1.1")], (unione$Year ==
unione$YearStartCareer) & unione$Semester ==
1)
careers.new = merge(careers, avg.att.11, by="StudentID", all.x = T)
careers.new$AvgAtt1.1[which(is.na(careers.new$AvgAtt1.1))] <- 0
careers <- careers.new
# compute total credits obtained, first semester, first year, for each student
tot.cred2 = ddply(passed.exams, .(StudentID, Year, Semester),
function(x) data.frame(TotalCredits1.1 = sum(x$NumberECTS)))
unione = merge(tot.cred2, enrolls, by='StudentID')
```

```
tot.cred.11 = subset(unione[,c("StudentID","TotalCredits1.1")], (unione$Year ==
unione$YearStartCareer) &
unione$Semester == 1)
careers.new = merge(careers, tot.cred.11, by="StudentID", all.x = T)
careers.new$TotalCredits1.1[which(is.na(careers.new$TotalCredits1.1))] <- 0
careers <- careers.new
head(careers)
#### 1st YEAR INFORMATION ####
attach(careers)
enrolls <- data.frame(StudentID, YearStartCareer)
detach(careers)
# compute weighted average, first year, for each student
avg.evals2 = ddply(passed.exams, .(StudentID, Year),
function(x) data.frame(WeiAvgEval1tot = weighted.mean(x$Score[x$Score != 0],
x$NumberECTS[x$Score
!= 0])))
unione = merge(avg.evals2, enrolls, by='StudentID')
avg.evals.1tot = subset(unione[,c("StudentID","WeiAvgEval1tot")], (unione$Year ==
unione$YearStartCareer) )
careers.new = merge(careers, avg.evals.1tot, by="StudentID", all.x = T)
careers.new$WeiAvgEval1tot[which(is.na(careers.new$WeiAvgEval1tot))] <- 0
careers <- careers.new
# compute average exams attempts, first year, for each student
avg.att2 = ddply(exams, .(StudentID, Year),
function(x) data.frame(AvgAtt1tot = mean(x$NumberAttempts)))
unione = merge(avg.att2, enrolls, by='StudentID')
avg.att.1tot = subset(unione[,c("StudentID","AvgAtt1tot")], (unione$Year ==
unione$YearStartCareer) )
careers.new = merge(careers, avg.att.1tot, by="StudentID", all.x = T)
careers.new$AvgAtt1tot[which(is.na(careers.new$AvgAtt1tot))] <- 0
careers <- careers.new
# compute total credits obtained, first year, for each student
tot.cred2 = ddply(passed.exams, .(StudentID, Year),
function(x) data.frame(TotalCredits1tot = sum(x$NumberECTS)))
unione = merge(tot.cred2, enrolls, by='StudentID')
tot.cred.1tot = subset(unione[,c("StudentID","TotalCredits1tot")], (unione$Year ==
unione$YearStartCareer) )
careers.new = merge(careers, tot.cred.1tot, by="StudentID", all.x = T)
careers.new$TotalCredits1tot[which(is.na(careers.new$TotalCredits1tot))] <- 0
careers <- careers.new
head(careers)
write.table(careers, file="PoliMi_career_information2.txt")
################################################
# (1.2) SELECT CAREERS to include in the model #
careers$Status = as.vector(careers$Status)
table(careers$Status, careers$YearStartCareer)
# Status: A (active), S (suspended), D (dropout), L(laurea)
# remove suspended careers
#careers.ended = subset(careers, is.element(careers$Status, c('D','L')))
#table(careers.ended$YearStartCareer, careers.ended$Status)
# da qui commento perchè prendo tutte le annate...selezionerò dopo quelle che voglio
# remove careers from 2016 to 2018
#last.year.considered = 2015
#careers.ended = subset(careers.ended, careers.ended$YearStartCareer <=
last.year.considered)
#################################
# (1.3) CREATE UNIQUE DATA TABLE #
#merge of students and careers
#x = merge(careers.ended, students, by ="StudentID")
x<-merge(careers,students,by="StudentID")
# merge with degree name
mydataframe = merge(x,degrees[,c("DegreeID","DegreeNature")], by.x = "StartingDegreeID",
by.y = "DegreeID", all.x=T)
colnames(mydataframe)[names(mydataframe) == "DegreeNature"] <- "DegreeProgramme.in"
head(mydataframe)
#############################################################################################
#####
###############
# (2) MODEL #
###############
# rm(list=ls())
#
# # load .csv
# mydataframe <-
read.csv(file="C:/Users/lucaf/Desktop/TESI_CONSEGNA/NewData/model_dataframe.csv")
# LIST OF VARIABLES OF INTEREST:
# Sex
# Nationality
# PreviousStudies
```

```
# AdmissionScore
# AccessToStudiesAge
# WeiAvgEval1.1
# AvgAtt1.1
# TotalCredits1.1
# Status
# DegreeProgramme.in
################################
# (2.1) VARIABLES REDEFINITION #
# "PreviousStudies" redefinition: turn it into a 3-level factor
mydataframe$PreviousStudies = as.vector(mydataframe$PreviousStudies)
table(mydataframe$PreviousStudies)
for( i in 1:length(mydataframe$PreviousStudies)) {
if( !is.element(mydataframe$PreviousStudies[i], c("Scientifica",
"Tecnica","Classica"))){
mydataframe$PreviousStudies[i] = "Other"
}
}t
able(mydataframe$PreviousStudies)
# "Nationality" redefinition: turn it into a 2-level factor
df.nationality<- data.frame(table(mydataframe$Nationality))
#################################################################
# (2.2) SAVE MODEL DATA TABLE (to save preprocessing result) #
# save .csv
head(mydataframe)
write.table(mydataframe, file="model_dataframe_ing.txt")
```

GMLER implementation algorithm

```
setwd("C:/Users/margh/Desktop/tesi/dati finali")
df.prova<-read.table("model_dataframe_ing.txt")
head(df.prova)
### NOTA BENE: nel dataset ci sono anche le carriere attive !!!
table(df.prova$Status)
#to do: selezionare mobility, merge con ATN di provenienza e se c'è il suo stato
atenei<-read.csv("atenei.csv")
nazioni<-read.csv("nazioni.csv")
nazioni$GEO_CNT_DN<-NULL
nazioni$GEO_STT_UE_FL<-NULL
nazioni$GEO_STT_ISTAT_CD<-NULL
atn.stt<-merge(nazioni,atenei,by.x="GEO_STT_ID",by.y="ATN_GEO_STT_ID")
colnames(atn.stt)[which(colnames(atn.stt) == "GEO_STT_DN")] = "NazTitolo"
colnames(atn.stt)[which(colnames(atn.stt) == "ATN_DN")] = "UniversOrigin"
colnames(atn.stt)[which(colnames(atn.stt) == "GEO_STT_ID")] = "GEO_STT_ID_ATN"
#carriere<-read.csv("carriere.csv")
#carriere<-carriere[c(1,2)]
#df1<-merge(carriere,df.prova,by.x="CARR_AN_ID",by.y="StudentID")
#colnames(df1)[which(colnames(df1)=="PERS_AN_ID")]="StudentID"
titoloprec<-read.csv("tit_acc_prec.csv")
#colnames(titoloprec)
#titoloprec<-titoloprec[c(1,3)]
#titolopre<-NULL
df1<-merge(df.prova,titoloprec,by.x = "StudentID",by.y="CARR_AN_ID")
df2<-merge(df1,atn.stt,by.x="TIT_ATN_ID",by.y="ATN_ID")
summary(df2$Mobility)
## tolgo solo gli studenti che hanno fatto mobilità incoming, che quindi sono del
PoliMi. --> 8 studenti
df3<-subset(df2,df2$Mobility=="Erasmus
outgoing"|df2$Mobility=="No"|df2$Mobility=="Other")
summary(df3$Mobility) #### outgoing=2380 no=26670 other=776
###creo variabile con polimi, nonpolimi, internazionali
df3$StudPrevStudies<-ifelse(df3$UniversOrigin=="Politecnico Di Milano", "PoliMi",
ifelse(df3$NazTitolo=="Italia" & df3$UniversOrigin!="Politecnico Di Milano",
"ItanonPoliMi", "International"))
##creo variabile su GPA < 23
df3$GPAlower23<-ifelse(df3$WeiAvgEval<23,1,0)
df3$GPAlower23<-as.character(df3$GPAlower23)
#creo variabile dropout
df3$dropout<-ifelse(df3$Status=="D","1","0")
#tolgo active careers
df3<-df3[-which(df3$Status=='A'),]
#creo la dummy del timetodegree>3 anni solo per i laureati, i dropout hanno NA
YearsToFinishDegree = df3$YearEndCareer +1 - df3$YearStartCareer
df3$HighTimetoDegree<-ifelse(df3$Status=='L',ifelse(df3$YearsToFinishDegree>3,1,0), NA)
df3$HighTimetoDegree<-as.character(df3$HighTimetoDegree)
```

```
df3$NazTitolo<-as.character(df3$NazTitolo)
##cambio nomi paesi-->inglese
table(df3$NazTitolo)
df3$NazTitolo[df3$NazTitolo=="Turchia"]<- "Turkey"
df3$NazTitolo[df3$NazTitolo=="Italia"]<- "Italy"
df3$NazTitolo[df3$NazTitolo=="Repubblica Popolare Di Cina"]<- "China"
df3$NazTitolo[df3$NazTitolo=="Egitto"]<- "Egypt"
df3$NazTitolo[df3$NazTitolo!="Italy" & df3$NazTitolo!="India" & df3$NazTitolo!="Iran" &
df3$NazTitolo!="Pakistan" &
df3$NazTitolo!="Turkey" & df3$NazTitolo!="Iran" & df3$NazTitolo!="China" &
df3$NazTitolo!="Colombia" &
df3$NazTitolo!="Egypt"] <- "others"
naz<-data.frame(table(df3$NazTitolo,df3$YearStartCareer))
table(df3$YearStartCareer[df3$NazTitolo!="Italy"])
sum(unique(df3$NazTitolo[df3$NazTitolo!="Italy"]))
write.csv(naz,"naz.csv")
###qualche descrittiva su polimi non polimi e internazionali
X<-
write.csv(table(df3$YearStartCareer,df3$StudPrevStudies,df3$Status),"status.yy.prevstud.
csv")
###ci sono solo carriere CONCLUSE
sum(is.na(df3$Status))
sum(df3$YearStartCareer==2015&df3$Status=="L") ### i laureati della coorte del 2015 sono
3038
sum(df3$YearStartCareer==2014&df3$Status=="L") ### i laureati della coorte del 2015 sono
2699
sum(df3$YearStartCareer==2013&df3$Status=="L") ### i laureati della coorte del 2015 sono
2516
sum(df3$YearStartCareer==2015&df3$Status=="D") ### i droppati della coorte del 2015 sono
144
sum(df3$YearStartCareer==2014&df3$Status=="D") ### i droppati della coorte del 2015 sono
130
sum(df3$YearStartCareer==2013&df3$Status=="D") ### i droppati della coorte del 2015 sono
130
### seleziono le variabili
library(dplyr)
table.gpa.t2d<-
select(df3,StudentID,HighTimetoDegree,Status,GPAlower23,StudPrevStudies,YearStartCareer)
write.csv(table.gpa.t2d,"table_gpa_t2d.csv")
table(YearStartCareer,Status=='L',StudPrevStudies)
table(YearStartCareer,Status=='D',StudPrevStudies)
table(YearStartCareer,StudPrevStudies)
##rifare qualche descrittiva considerando anche le carriere attive
##considero solo coloro che hanno iniziato e finito nello stesso cds? no, tutti però
consideriamo come riferimento il
CDS in cui ha iniziato
table(df3$ChangeDegree)
#df3<-subset(df3,ChangeDegree==FALSE)
df3$MostPopulUniv<-NULL
###selezioniamo solo studenti internazionali
df.internaz<-subset(df3,StudPrevStudies=="International")
### per l'analisi seleziono solo gli enrolments dal 2013 al 2015 e le carriere concluse
df.internaz.analisi<-
subset(df.internaz,YearStartCareer==2013|YearStartCareer==2014|YearStartCareer==2015)
df.internaz.analisi<-subset(df.internaz.analisi,Status!='A'|Status!='S')
### creo variabile università di proveninenza con più carriere concluse (più di 10) (!)
##ricorda: per il ciclo for tipo questo devi trasformare in factor
df.internaz.analisi$UniversOrigin<-as.character(df.internaz.analisi$UniversOrigin)
for(i in 1:length(df.internaz.analisi$UniversOrigin)){
if(!is.element(df.internaz.analisi$UniversOrigin[i],c("Istanbul Teknik
Universitesi","Anna University","Jawaharlal
Nehru Technological University, Hyderabad","Universita' Azad Islamica","Tongji
University","Visvesvaraya
Technological University","Universidad De Los Andes","Universita' Tecnologica
Amirkabir","Universita' Di
Belgrado","University Of Teheran","Universidad Nacional De Colombia","Jawaharlal Nehru
Technological University,
Kakinada","Sharif University Of Technology"))){
df.internaz.analisi$UniversOrigin[i]="other_universities"
}
}
table(df.internaz.analisi$UniversOrigin)
## seleziono CDS con più di 20 studenti internaz. e metto a posto i nomi dei CDS
tt = data.frame(table(df.internaz.analisi$DegreeProgramme.in))
View(tt)
selez = as.vector(subset(tt$Var1, tt$Freq>20))
df.internaz.analisi = subset(df.internaz.analisi,
is.element(df.internaz.analisi$DegreeProgramme.in, selez)) ## 1245
obs per analisi
```

```
dropout.naz<-select(df.internaz.analisi,Status,YearStartCareer,StudentID,NazTitolo)
write.csv(dropout.naz,"dropout.naz.csv")
sapply(df.internaz.analisi, class)
View(table(df.internaz.analisi$DegreeProgramme.in))
table(df.internaz.analisi$DegreeProgramme.in)
df.internaz.analisi$DegreeProgramme.in<-
as.character(df.internaz.analisi$DegreeProgramme.in)
df.internaz.analisi$DegreeProgramme.in[df.internaz.analisi$DegreeProgramme.in=="Ingegner
ia Gestionale"]<-"Management
Engineering - Ingegneria Gestionale"
df.internaz.analisi$DegreeProgramme.in[df.internaz.analisi$DegreeProgramme.in=="Ingegner
ia Dell'Automazione"]<-
"Automation And Control Engineering - Ingegneria Dell'Automazione"
df.internaz.analisi$DegreeProgramme.in[df.internaz.analisi$DegreeProgramme.in=="Ingegner
ia Elettrica"]<-"Electrical
Engineering - Ingegneria Elettrica"
df.internaz.analisi$DegreeProgramme.in[df.internaz.analisi$DegreeProgramme.in=="Material
s Engineering And
Nanotechnology - Ingegneria Dei Materiali E Delle Nanotecnologie"]<-"Materials
Engineering And Nanotechnology"
table(df.internaz.analisi$DegreeProgramme.in)
df.internaz.analisi$DegreeProgramme.in<-
as.factor(df.internaz.analisi$DegreeProgramme.in)
## creo test e train
train<-subset(df.internaz.analisi,YearStartCareer==2013 | YearStartCareer==2014 )
test<-subset(df.internaz.analisi,YearStartCareer==2015)
library(lme4)
library(ggplot2)
library(lme4)
library(lattice)
library(dplyr)
train$DegreeProgramme.in<-as.factor(train$DegreeProgramme.in)
test$DegreeProgramme.in<-as.factor(test$DegreeProgramme.in)
train$NazTitolo[train$NazTitolo=="others"]<-"a.others"
test$NazTitolo[test$NazTitolo=="others"]<-"a.others"
write.table(train,"train.txt")
write.table(test,"test.txt")
#### ANALISI
train<-read.table("train.txt")
test<-read.table("test.txt")
table(train$AvgAttNew1)
table(test$AvgAttNew1)
#logit dropout e CDS
names(train)
train$dropout<-as.factor(train$dropout)
test$dropout<-as.factor(test$dropout)
m1<- glmer(dropout ~ (1|DegreeProgramme.in)+
NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex +
AccessToStudiesAge, data = train, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
m1<- glmer(dropout ~ (1|DegreeProgramme.in)+WeiAvgEval1tot+TotalCredits1tot +
AccessToStudiesAge, data = train,
family = binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
x11()
m.red = m1
rand_intercept = ranef(m.red, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
psiA = as.numeric(summary(m.red)$varcor)
psiA/(psiA +pi^2/3)
test_pred <- predict(m1,test,re.form=NULL,type="response", allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$dropout==1))
t10 = length(which(pred2 ==1 & test$dropout==0))
t01 = length(which(pred2 ==0 & test$dropout==1))
t00 = length(which(pred2 ==0 & test$dropout==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
```

```
i0 = 210
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
# il p ottimale è:
p0opt
test_pred_class <- ifelse(test_pred> p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('dropout','graduated'))
observed = factor(test$dropout, levels = c(1,0),labels = c('dropout','graduated'))
misc.table = table(predicted, observed)
misc.table
dim(predicted)
dim(observed)
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
#logit con dropout e nationalities
train$dropout<-as.factor(train$dropout)
test$dropout<-as.factor(test$dropout)
m1<- glmer(dropout ~ (1|NazTitolo) +WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex +
AccessToStudiesAge, data =
train, family = binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
m1<- glmer(dropout ~ (1|NazTitolo) +WeiAvgEval1tot+TotalCredits1tot+ AccessToStudiesAge,
data = train, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
x11()
m.red = m1
rand_intercept = ranef(m.red, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
psiA = as.numeric(summary(m.red)$varcor)
psiA/(psiA +pi^2/3)
test_pred <- predict(m1,test,re.form=NULL,type="response", allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$dropout==1))
t10 = length(which(pred2 ==1 & test$dropout==0))
t01 = length(which(pred2 ==0 & test$dropout==1))
t00 = length(which(pred2 ==0 & test$dropout==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 200
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
# il p ottimale è:
p0opt
test_pred_class <- ifelse(test_pred> p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('dropout','graduated'))
observed = factor(test$dropout, levels = c(1,0),labels = c('dropout','graduated'))
misc.table = table(predicted, observed)
misc.table
dim(predicted)
dim(observed)
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
###logit con time2degree e CDS --> tolgo i dropout
train.degree<-subset(train,Status=='L')
```

```
test.degree<-subset(test,Status=='L')
table(train.degree$DegreeProgramme.in)
table(test.degree$AvgAttNew1) ##prendendo solo i laureati rimane solo uno studente con
zero tentativi dal 2013 al
2015
table(train.degree$AvgAttNew1)
summary(train.degree$AvgAtt1.1)
sum(train.degree$AvgAtt1.1==0)
table(train.degree$Mobility)
train.degree$HighTimetoDegree<-as.factor(train.degree$HighTimetoDegree)
test.degree$HighTimetoDegree<-as.factor(test.degree$HighTimetoDegree)
m1<- glmer(HighTimetoDegree ~ (1|DegreeProgramme.in)+
NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex +
AccessToStudiesAge, data = train.degree, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
m1<- glmer(HighTimetoDegree ~ (1|DegreeProgramme.in)+
NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot, data =
train.degree, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
# Graphically:
x11()
m.red = m1
rand_intercept = ranef(m.red, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
psiA = as.numeric(summary(m.red)$varcor)
psiA/(psiA +pi^2/3)
test_pred <- predict(m1,test.degree,re.form=NULL,type="response", allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==1))
t10 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==0))
t01 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==1))
t00 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 130
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
# il p ottimale è:
p0opt
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('more 3 years','less 3
years'))
observed = factor(test.degree$HighTimetoDegree, levels = c(1,0),labels = c('more 3
years','less 3 years'))
misc.table = table(predicted, observed)
misc.table
dim(predicted)
dim(observed)
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
#logit t2d e nationalities
train.degree<-subset(train,Status=='L')
test.degree<-subset(test,Status=='L')
table(train.degree$DegreeProgramme.in)
table(test.degree$AvgAttNew1) ##prendendo solo i laureati rimane solo uno studente con
zero tentativi dal 2013 al
2015
table(train.degree$AvgAttNew1)
summary(train.degree$AvgAtt1.1)
sum(train.degree$AvgAtt1.1==0)
```

```
table(train.degree$Mobility)
train.degree$HighTimetoDegree<-as.factor(train.degree$HighTimetoDegree)
test.degree$HighTimetoDegree<-as.factor(test.degree$HighTimetoDegree)
m1<- glmer(HighTimetoDegree ~ (1|NazTitolo) +WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+
Sex + AccessToStudiesAge,
data = train.degree, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
m1<- glmer(HighTimetoDegree ~ (1|NazTitolo) +WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot,
data = train.degree, family
= binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m1)
# Graphically:
x11()
m.red = m1
rand_intercept = ranef(m.red, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
psiA = as.numeric(summary(m.red)$varcor)
psiA/(psiA +pi^2/3)
test_pred <- predict(m1,test.degree,re.form=NULL,type="response", allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==1))
t10 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==0))
t01 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==1))
t00 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 115
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
# il p ottimale è:
p0opt
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('more 3 years','less 3
years'))
observed = factor(test.degree$HighTimetoDegree, levels = c(1,0),labels = c('more 3
years','less 3 years'))
misc.table = table(predicted, observed)
misc.table
dim(predicted)
dim(observed)
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
###logit GPA E CDS, utilizzo lo stesso dataset del T2D (solo i laureati !!!)
train$GPAlower23<-as.factor(train$GPAlower23)
test$GPAlower23<-as.factor(test$GPAlower23)
m2 <- glmer(GPAlower23 ~ (1|DegreeProgramme.in)+
NazTitolo+TotalCredits1tot+AvgAtt1tot+YearsToFinishDegree+
AccessToStudiesAge, data = train.degree, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m2)
m2 <- glmer(GPAlower23 ~ (1|DegreeProgramme.in)+ NazTitolo+TotalCredits1tot+AvgAtt1tot +
AccessToStudiesAge, data =
train.degree, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m2)
# Graphically:
x11()
m.red = m2
rand_intercept = ranef(m.red, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
psiA = as.numeric(summary(m.red)$varcor)
```

```
psiA/(psiA +pi^2/3)
test_pred <- predict(m2,test,re.form=NULL,type="response", allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$GPAlower23==1))
t10 = length(which(pred2 ==1 & test$GPAlower23==0))
t01 = length(which(pred2 ==0 & test$GPAlower23==1))
t00 = length(which(pred2 ==0 & test$GPAlower23==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 275
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
# il p ottimale è:
p0opt
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
observed = factor(test$GPAlower23, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
misc.table = table(predicted, observed)
misc.table
dim(predicted)
dim(observed)
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
###logit GPA E nationality, utilizzo lo stesso dataset del T2D (solo i laureati !!!)
train$GPAlower23<-as.factor(train$GPAlower23)
test$GPAlower23<-as.factor(test$GPAlower23)
m2 <- glmer(GPAlower23 ~ (1|NazTitolo)+TotalCredits1tot+AvgAtt1tot+ YearsToFinishDegree
+ AccessToStudiesAge, data =
train.degree, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m2)
m2 <- glmer(GPAlower23 ~ (1|NazTitolo)+TotalCredits1tot+AvgAtt1tot+ AccessToStudiesAge,
data = train.degree, family =
binomial,na.action=na.omit,control=glmerControl(optimizer="bobyqa"))
summary(m2)
# Graphically:
x11()
m.red = m2
rand_intercept = ranef(m.red, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
psiA = as.numeric(summary(m.red)$varcor)
psiA/(psiA +pi^2/3)
test_pred <- predict(m2,test,re.form=NULL,type="response", allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$GPAlower23==1))
t10 = length(which(pred2 ==1 & test$GPAlower23==0))
t01 = length(which(pred2 ==0 & test$GPAlower23==1))
t00 = length(which(pred2 ==0 & test$GPAlower23==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 335
points(specificity_comp[i0], sensitivity[i0], pch = 1)
```

```
p0opt = p0[i0] # threshold value
# il p ottimale è:
p0opt
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
observed = factor(test$GPAlower23, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
misc.table = table(predicted, observed)
misc.table
dim(predicted)
dim(observed)
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
```

GMET tree algorithm

```
# funzione per fittare "mymixedtrees"
mymixedtree <- function(formula, dataset, random, subset = NULL,
family = binomial, tree.control = rpart.control(),
cv=TRUE, cpmin = 0.01, verbose = FALSE){
# Parse formula
Predictors <- paste(attr(terms(formula),"term.labels"),collapse="+")
TargetName <- formula[[2]]
# Subset the data if necessary
if(identical(subset, NULL)){
train <- dataset
} else {
train <- subset(dataset, dati[,subset])
}
################
# 1) TRAIN MODEL
################
# Fit GLM (only fixed effect)
GLM <- glm(formula = formula, data = train, family = family)
if(verbose) {
print(GLM)}
hat.p = GLM$fitted.values
# Fit regression tree on predicted probabilities
# and prune tree (eventually)
if(cv){
CART <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = train, method = "anova", control = rpart.control(cp=cpmin))
cventry <- which.min(CART$cptable[, "xerror"])
xerrorcv <- CART$cptable[cventry, "xerror"]
sexerrorcv <- xerrorcv + CART$cptable[cventry, "xstd"]
cpcvse <- CART$cptable[which.max(CART$cptable[, "xerror"] <= sexerrorcv), "CP"]
CART.pr <- prune(CART, cp=cpcvse)
if(verbose){
print(CART.pr)
rpart.plot(CART.pr)
}
} else {
CART.pr <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = train, method = "anova", control = tree.control)
if(verbose){
print(CART.pr)
rpart.plot(CART.pr)
}
}
# get indicator of leaves
leaves <- as.numeric(rownames(CART.pr$frame)[CART.pr$where])
train[,"nodeInd"] <- leaves
n.leaves = length(unique(leaves))
# Fit GLMM
fixed = paste(c(TargetName, "as.factor(nodeInd)"), collapse = "~")
GLMM <- glmer(formula = formula(paste(c(fixed, random[[2]]), collapse = "+")),
data=train, family = family, control=glmerControl(optimizer="bobyqa"))
if(verbose) {
summary(GLMM)}
#get fixed effect from the model and put it into tree
adjtarg <- unique(cbind(leaves, predict(GLMM, re = NA, type = "response")))
CART.pr$frame[as.character(adjtarg[,"leaves"]),]$yval <- adjtarg[,2]
```

```
# fitted values
fitted.GLMM <- predict(GLMM, type = "response")
result <- list(
Tree=CART.pr,
EffectModel=GLMM,
RandomEffects=ranef(GLMM),
BetweenMatrix = VarCorr(GLMM)
)
class(result) <- "mymixedtree"
return(result)
}
### Function: plot.mymixedtree
# This function plots the tree part of a mymixedtree tree
# Inputs:
# x - a RE-EM tree object
plot.mymixedtree <- function(x, ...){
require(rpart.plot)
rpart.plot(x$Tree)
}
### Function: ranef.REEMtree
# This function extracts the vector of estimated random effects
ranef.mymixedtree <- function(object,...){
return(object$RandomEffects)
}
predict.mymixedtree <- function(object, newdata, type = "response", ...){
# get underlying tree structure
tree.aux = object$Tree
# substitute the row index as predicted value in each leaf
tree.aux$frame$yval = as.numeric(rownames(tree.aux$frame))
# get the leaf indicator for each newdata
newdata[,"nodeInd"] <- predict(tree.aux, newdata= newdata)
#predict using GLMM
predict(object$EffectModel, newdata = newdata, type = type,...)
}
```

GMET implementation algorithm

```
rm(list=ls())
require(nlme)
require(rpart)
require(rpart.plot)
require(lme4)
require(dplyr)
require(lattice)
require(ggplot2)
source("C:/Users/margh/Desktop/tesi/dati finali/mymixedtree.R")
setwd("C:/Users/margh/Desktop/tesi/dati finali")
train<-read.table("train.txt")
test<-read.table("test.txt")
# Prova funzione
sapply(train, class)
formula = dropout ~ NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex +
AccessToStudiesAge
dataset = train
random = ~ (1|DegreeProgramme.in)
family = binomial
tree.control = rpart.control(cp = 0.006, maxdepth=10, minbucket = 15)
# (1) no cv
Predictors <- paste(attr(terms(formula),"term.labels"),collapse="+")
Predictors
TargetName <- formula[[2]]
TargetName
# step 2: Fit GLM (only fixed effect)
GLM <- glm(formula = formula, data = dataset, family = family)
hat.p = GLM$fitted.values
# step 3: Fit regression tree on predicted probabilities
CART.pr <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = dataset, method = "anova", control = tree.control)
# rpart.plot(CART.pr)
# get indicator of leaves
leaves <- as.numeric(rownames(CART.pr$frame)[CART.pr$where])
dataset[,"nodeInd"] <- leaves
n.leaves = length(unique(leaves))
# step 4: Fit GLMM
library(lme4)
```

```
fixed = paste(c(TargetName, "as.factor(nodeInd)"), collapse = "~")
GLMM <- glmer(formula = formula(paste(c(fixed, random[[2]]), collapse = "+")),
data=dataset, family = family, control=glmerControl(optimizer="bobyqa"))
# step 5: get fixed effect from the model and put it into tree leaves
adjtarg <- unique(cbind(leaves, predict(GLMM, re = NA, type = "response")))
CART.pr$frame[as.character(adjtarg[,"leaves"]),]$yval <- adjtarg[,2]
GMET <- list(
Tree=CART.pr,
EffectModel=GLMM,
RandomEffects=ranef(GLMM),
BetweenMatrix = VarCorr(GLMM)
)c
lass(GMET) <- "GMET"
print(GMET$Tree)
library(rpart)
plot(GMET$Tree)
text(GMET$Tree)
rpart.plot(GMET$Tree)
# plot random effect
x11()
rand_intercept = ranef(GMET$EffectModel, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
rand_intercept
# VPC
psiA = as.numeric(summary(GMET$EffectModel)$varcor)
psiA/(psiA +pi^2/3) # 3.6%
predict.GMET <- function(object, newdata, type = "response", ...){
# get underlying tree structure
tree.aux = object$Tree
# substitute the row index as predicted value in each leaf
tree.aux$frame$yval = as.numeric(rownames(tree.aux$frame))
# get the leaf indicator for each newdata
newdata[,"nodeInd"] <- predict(tree.aux, newdata= newdata)
#predict using GLMM
predict(object$EffectModel, newdata = newdata, type = type,...)
}
# ROC analysis
test_pred <- predict(GMET,test,re.form=NULL,type="response",allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$dropout==1))
t10 = length(which(pred2 ==1 & test$dropout==0))
t01 = length(which(pred2 ==0 & test$dropout==1))
t00 = length(which(pred2 ==0 & test$dropout==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 145
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
p0opt
# ONE SHOT indexes
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('Dropout','Graduate'))
observed = factor(test$dropout, levels = c(1,0),labels = c('Dropout','Graduate'))
misc.table = table(predicted, observed)
misc.table
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
rm(list=ls())
require(nlme)
require(rpart)
require(rpart.plot)
require(lme4)
```

```
require(dplyr)
require(lattice)
require(ggplot2)
source("C:/Users/margh/Desktop/tesi/dati finali/mymixedtree.R")
setwd("C:/Users/margh/Desktop/tesi/dati finali")
train<-read.table("train.txt")
test<-read.table("test.txt")
# Prova funzione
sapply(train, class)
formula = dropout ~ WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge
dataset = train
random = ~ (1|NazTitolo)
family = binomial
tree.control = rpart.control(cp = 0.006, maxdepth=10, minbucket = 15)
# (1) no cv
Predictors <- paste(attr(terms(formula),"term.labels"),collapse="+")
Predictors
TargetName <- formula[[2]]
TargetName
# step 2: Fit GLM (only fixed effect)
GLM <- glm(formula = formula, data = dataset, family = family)
hat.p = GLM$fitted.values
# step 3: Fit regression tree on predicted probabilities
CART.pr <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = dataset, method = "anova", control = tree.control)
# rpart.plot(CART.pr)
# get indicator of leaves
leaves <- as.numeric(rownames(CART.pr$frame)[CART.pr$where])
dataset[,"nodeInd"] <- leaves
n.leaves = length(unique(leaves))
# step 4: Fit GLMM
library(lme4)
fixed = paste(c(TargetName, "as.factor(nodeInd)"), collapse = "~")
GLMM <- glmer(formula = formula(paste(c(fixed, random[[2]]), collapse = "+")),
data=dataset, family = family, control=glmerControl(optimizer="bobyqa"))
# step 5: get fixed effect from the model and put it into tree leaves
adjtarg <- unique(cbind(leaves, predict(GLMM, re = NA, type = "response")))
CART.pr$frame[as.character(adjtarg[,"leaves"]),]$yval <- adjtarg[,2]
GMET <- list(
Tree=CART.pr,
EffectModel=GLMM,
RandomEffects=ranef(GLMM),
BetweenMatrix = VarCorr(GLMM)
)c
lass(GMET) <- "GMET"
print(GMET$Tree)
library(rpart)
plot(GMET$Tree)
text(GMET$Tree)
rpart.plot(GMET$Tree)
# plot random effect
x11()
rand_intercept = ranef(GMET$EffectModel, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
rand_intercept
# VPC
psiA = as.numeric(summary(GMET$EffectModel)$varcor)
psiA/(psiA +pi^2/3) # 3.6%
predict.GMET <- function(object, newdata, type = "response", ...){
# get underlying tree structure
tree.aux = object$Tree
# substitute the row index as predicted value in each leaf
tree.aux$frame$yval = as.numeric(rownames(tree.aux$frame))
# get the leaf indicator for each newdata
newdata[,"nodeInd"] <- predict(tree.aux, newdata= newdata)
#predict using GLMM
predict(object$EffectModel, newdata = newdata, type = type,...)
}
test_pred <- predict(GMET,test,re.form=NULL,type="response",allow.new.levels=T)
# ROC analysis
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$dropout==1))
t10 = length(which(pred2 ==1 & test$dropout==0))
t01 = length(which(pred2 ==0 & test$dropout==1))
```

```
t00 = length(which(pred2 ==0 & test$dropout==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 145
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
p0opt
# ONE SHOT indexes
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('Dropout','Graduate'))
observed = factor(test$dropout, levels = c(1,0),labels = c('Dropout','Graduate'))
misc.table = table(predicted, observed)
misc.table
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
rm(list=ls())
require(nlme)
require(rpart)
require(rpart.plot)
require(lme4)
require(dplyr)
require(lattice)
require(ggplot2)
source("C:/Users/margh/Desktop/tesi/dati finali/mymixedtree.R")
setwd("C:/Users/margh/Desktop/tesi/dati finali")
train<-read.table("train.txt")
test<-read.table("test.txt")
sapply(train, class)
train.degree<-subset(train,Status=='L')
test.degree<-subset(test,Status=='L')
table(train.degree$DegreeProgramme.in)
table(test.degree$AvgAttNew1) ##prendendo solo i laureati rimane solo uno studente con
zero tentativi dal 2013 al
2015
table(train.degree$AvgAttNew1)
summary(train.degree$AvgAtt1.1)
sum(train.degree$AvgAtt1.1==0)
table(train.degree$Mobility)
train.degree$HighTimetoDegree<-as.factor(train.degree$HighTimetoDegree)
test.degree$HighTimetoDegree<-as.factor(test.degree$HighTimetoDegree)
formula = HighTimetoDegree ~ NazTitolo+WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex +
AccessToStudiesAge
dataset = train.degree
random = ~(1|DegreeProgramme.in)
family = binomial
tree.control = rpart.control(cp = 0.006, maxdepth=10, minbucket = 15)
# (1) no cv
Predictors <- paste(attr(terms(formula),"term.labels"),collapse="+")
Predictors
TargetName <- formula[[2]]
TargetName
# step 2: Fit GLM (only fixed effect)
GLM <- glm(formula = formula, data = dataset, family = family)
hat.p = GLM$fitted.values
# step 3: Fit regression tree on predicted probabilities
CART.pr <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = dataset, method = "anova", control = tree.control)
# rpart.plot(CART.pr)
# get indicator of leaves
leaves <- as.numeric(rownames(CART.pr$frame)[CART.pr$where])
dataset[,"nodeInd"] <- leaves
n.leaves = length(unique(leaves))
# step 4: Fit GLMM
library(lme4)
fixed = paste(c(TargetName, "as.factor(nodeInd)"), collapse = "~")
GLMM <- glmer(formula = formula(paste(c(fixed, random[[2]]), collapse = "+")),
data=dataset, family = family, control=glmerControl(optimizer="bobyqa"))
```

```
# step 5: get fixed effect from the model and put it into tree leaves
adjtarg <- unique(cbind(leaves, predict(GLMM, re = NA, type = "response")))
CART.pr$frame[as.character(adjtarg[,"leaves"]),]$yval <- adjtarg[,2]
GMET <- list(
Tree=CART.pr,
EffectModel=GLMM,
RandomEffects=ranef(GLMM),
BetweenMatrix = VarCorr(GLMM)
)c
lass(GMET) <- "GMET"
print(GMET$Tree)
library(rpart)
plot(GMET$Tree)
text(GMET$Tree)
rpart.plot(GMET$Tree)
# plot random effect
x11()
rand_intercept = ranef(GMET$EffectModel, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
rand_intercept
# VPC
psiA = as.numeric(summary(GMET$EffectModel)$varcor)
psiA/(psiA +pi^2/3) # 3.6%
predict.GMET <- function(object, newdata, type = "response", ...){
# get underlying tree structure
tree.aux = object$Tree
# substitute the row index as predicted value in each leaf
tree.aux$frame$yval = as.numeric(rownames(tree.aux$frame))
# get the leaf indicator for each newdata
newdata[,"nodeInd"] <- predict(tree.aux, newdata= newdata)
#predict using GLMM
predict(object$EffectModel, newdata = newdata, type = type,...)
}
# ROC analysis
test_pred <- predict(GMET,test.degree,re.form=NULL,type="response",allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==1))
t10 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==0))
t01 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==1))
t00 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 115
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
p0opt
# ONE SHOT indexes
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('more 3 years','less 3
years'))
observed = factor(test.degree$HighTimetoDegree, levels = c(1,0),labels = c('more 3
years','less 3 years'))
misc.table = table(predicted, observed)
misc.table
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
rm(list=ls())
require(nlme)
require(rpart)
require(rpart.plot)
require(lme4)
require(dplyr)
```

```
require(lattice)
require(ggplot2)
source("C:/Users/margh/Desktop/tesi/dati finali/mymixedtree.R")
setwd("C:/Users/margh/Desktop/tesi/dati finali")
train<-read.table("train.txt")
test<-read.table("test.txt")
sapply(train, class)
train.degree<-subset(train,Status=='L')
test.degree<-subset(test,Status=='L')
table(train.degree$DegreeProgramme.in)
table(test.degree$AvgAttNew1) ##prendendo solo i laureati rimane solo uno studente con
zero tentativi dal 2013 al
2015
table(train.degree$AvgAttNew1)
summary(train.degree$AvgAtt1.1)
sum(train.degree$AvgAtt1.1==0)
table(train.degree$Mobility)
train.degree$HighTimetoDegree<-as.factor(train.degree$HighTimetoDegree)
test.degree$HighTimetoDegree<-as.factor(test.degree$HighTimetoDegree)
formula = HighTimetoDegree ~ WeiAvgEval1tot+TotalCredits1tot+AvgAtt1tot+ Sex +
AccessToStudiesAge
dataset = train.degree
random = ~(1|NazTitolo)
family = binomial
tree.control = rpart.control(cp = 0.006, maxdepth=10, minbucket = 15)
# (1) no cv
Predictors <- paste(attr(terms(formula),"term.labels"),collapse="+")
Predictors
TargetName <- formula[[2]]
TargetName
# step 2: Fit GLM (only fixed effect)
GLM <- glm(formula = formula, data = dataset, family = family)
hat.p = GLM$fitted.values
# step 3: Fit regression tree on predicted probabilities
CART.pr <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = dataset, method = "anova", control = tree.control)
# rpart.plot(CART.pr)
# get indicator of leaves
leaves <- as.numeric(rownames(CART.pr$frame)[CART.pr$where])
dataset[,"nodeInd"] <- leaves
n.leaves = length(unique(leaves))
# step 4: Fit GLMM
library(lme4)
fixed = paste(c(TargetName, "as.factor(nodeInd)"), collapse = "~")
GLMM <- glmer(formula = formula(paste(c(fixed, random[[2]]), collapse = "+")),
data=dataset, family = family, control=glmerControl(optimizer="bobyqa"))
# step 5: get fixed effect from the model and put it into tree leaves
adjtarg <- unique(cbind(leaves, predict(GLMM, re = NA, type = "response")))
CART.pr$frame[as.character(adjtarg[,"leaves"]),]$yval <- adjtarg[,2]
GMET <- list(
Tree=CART.pr,
EffectModel=GLMM,
RandomEffects=ranef(GLMM),
BetweenMatrix = VarCorr(GLMM)
)c
lass(GMET) <- "GMET"
print(GMET$Tree)
library(rpart)
plot(GMET$Tree)
text(GMET$Tree)
rpart.plot(GMET$Tree)
# plot random effect
x11()
rand_intercept = ranef(GMET$EffectModel, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
rand_intercept
# VPC
psiA = as.numeric(summary(GMET$EffectModel)$varcor)
psiA/(psiA +pi^2/3) # 3.6%
predict.GMET <- function(object, newdata, type = "response", ...){
# get underlying tree structure
tree.aux = object$Tree
# substitute the row index as predicted value in each leaf
tree.aux$frame$yval = as.numeric(rownames(tree.aux$frame))
# get the leaf indicator for each newdata
newdata[,"nodeInd"] <- predict(tree.aux, newdata= newdata)
#predict using GLMM
predict(object$EffectModel, newdata = newdata, type = type,...)
}
```

```
# ROC analysis
test_pred <- predict(GMET,test.degree,re.form=NULL,type="response",allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==1))
t10 = length(which(pred2 ==1 & test.degree$HighTimetoDegree==0))
t01 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==1))
t00 = length(which(pred2 ==0 & test.degree$HighTimetoDegree==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 130
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
p0opt
# ONE SHOT indexes
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('more 3 years','less 3
years'))
observed = factor(test.degree$HighTimetoDegree, levels = c(1,0),labels = c('more 3
years','less 3 years'))
misc.table = table(predicted, observed)
misc.table
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
rm(list=ls())
require(nlme)
require(rpart)
require(rpart.plot)
require(lme4)
require(dplyr)
require(lattice)
require(ggplot2)
source("C:/Users/margh/Desktop/tesi/dati finali/mymixedtree.R")
setwd("C:/Users/margh/Desktop/tesi/dati finali")
train<-read.table("train.txt")
test<-read.table("test.txt")
sapply(train, class)
train.degree<-subset(train,Status=='L')
test.degree<-subset(test,Status=='L')
table(train.degree$DegreeProgramme.in)
table(test.degree$AvgAttNew1) ##prendendo solo i laureati rimane solo uno studente con
zero tentativi dal 2013 al
2015
table(train.degree$AvgAttNew1)
summary(train.degree$AvgAtt1.1)
sum(train.degree$AvgAtt1.1==0)
table(train.degree$Mobility)
train.degree$HighTimetoDegree<-as.factor(train.degree$GPAlower23)
test.degree$HighTimetoDegree<-as.factor(test.degree$GPAlower23)
formula = GPAlower23 ~ NazTitolo+TotalCredits1tot+AvgAtt1tot+ Sex
+YearsToFinishDegree+AccessToStudiesAge
dataset = train.degree
random = ~(1|DegreeProgramme.in)
family = binomial
tree.control = rpart.control(cp = 0.006, maxdepth=10, minbucket = 15)
# (1) no cv
Predictors <- paste(attr(terms(formula),"term.labels"),collapse="+")
Predictors
TargetName <- formula[[2]]
TargetName
# step 2: Fit GLM (only fixed effect)
GLM <- glm(formula = formula, data = dataset, family = family)
hat.p = GLM$fitted.values
```

```
# step 3: Fit regression tree on predicted probabilities
CART.pr <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = dataset, method = "anova", control = tree.control)
# rpart.plot(CART.pr)
# get indicator of leaves
leaves <- as.numeric(rownames(CART.pr$frame)[CART.pr$where])
dataset[,"nodeInd"] <- leaves
n.leaves = length(unique(leaves))
# step 4: Fit GLMM
library(lme4)
fixed = paste(c(TargetName, "as.factor(nodeInd)"), collapse = "~")
GLMM <- glmer(formula = formula(paste(c(fixed, random[[2]]), collapse = "+")),
data=dataset, family = family, control=glmerControl(optimizer="bobyqa"))
# step 5: get fixed effect from the model and put it into tree leaves
adjtarg <- unique(cbind(leaves, predict(GLMM, re = NA, type = "response")))
CART.pr$frame[as.character(adjtarg[,"leaves"]),]$yval <- adjtarg[,2]
GMET <- list(
Tree=CART.pr,
EffectModel=GLMM,
RandomEffects=ranef(GLMM),
BetweenMatrix = VarCorr(GLMM)
)c
lass(GMET) <- "GMET"
print(GMET$Tree)
library(rpart)
plot(GMET$Tree)
text(GMET$Tree)
x11()
rpart.plot(GMET$Tree)
# plot random effect
x11()
rand_intercept = ranef(GMET$EffectModel, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
rand_intercept
# VPC
psiA = as.numeric(summary(GMET$EffectModel)$varcor)
psiA/(psiA +pi^2/3) # 3.6%
predict.GMET <- function(object, newdata, type = "response", ...){
# get underlying tree structure
tree.aux = object$Tree
# substitute the row index as predicted value in each leaf
tree.aux$frame$yval = as.numeric(rownames(tree.aux$frame))
# get the leaf indicator for each newdata
newdata[,"nodeInd"] <- predict(tree.aux, newdata= newdata)
#predict using GLMM
predict(object$EffectModel, newdata = newdata, type = type,...)
}
# ROC analysis
test_pred <- predict(GMET,test,re.form=NULL,type="response",allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$GPAlower23==1))
t10 = length(which(pred2 ==1 & test$GPAlower23==0))
t01 = length(which(pred2 ==0 & test$GPAlower23==1))
t00 = length(which(pred2 ==0 & test$GPAlower23==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 325
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
p0opt
# ONE SHOT indexes
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
observed = factor(test$GPAlower23, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
misc.table = table(predicted, observed)
misc.table
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
```

```
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4)))
rm(list=ls())
require(nlme)
require(rpart)
require(rpart.plot)
require(lme4)
require(dplyr)
require(lattice)
require(ggplot2)
source("C:/Users/margh/Desktop/tesi/dati finali/mymixedtree.R")
setwd("C:/Users/margh/Desktop/tesi/dati finali")
train<-read.table("train.txt")
test<-read.table("test.txt")
sapply(train, class)
train.degree<-subset(train,Status=='L')
test.degree<-subset(test,Status=='L')
table(train.degree$DegreeProgramme.in)
table(test.degree$AvgAttNew1) ##prendendo solo i laureati rimane solo uno studente con
zero tentativi dal 2013 al
2015
table(train.degree$AvgAttNew1)
summary(train.degree$AvgAtt1.1)
sum(train.degree$AvgAtt1.1==0)
table(train.degree$Mobility)
train.degree$HighTimetoDegree<-as.factor(train.degree$GPAlower23)
test.degree$HighTimetoDegree<-as.factor(test.degree$GPAlower23)
formula = GPAlower23 ~ TotalCredits1tot+AvgAtt1tot+ Sex + AccessToStudiesAge+
YearsToFinishDegree
dataset = train.degree
random = ~(1|NazTitolo)
family = binomial
tree.control = rpart.control(cp = 0.006, maxdepth=10, minbucket = 15)
# (1) no cv
Predictors <- paste(attr(terms(formula),"term.labels"),collapse="+")
Predictors
TargetName <- formula[[2]]
TargetName
# step 2: Fit GLM (only fixed effect)
GLM <- glm(formula = formula, data = dataset, family = family)
hat.p = GLM$fitted.values
# step 3: Fit regression tree on predicted probabilities
CART.pr <- rpart(formula = formula(paste(c("hat.p", Predictors), collapse = "~")),
data = dataset, method = "anova", control = tree.control)
# rpart.plot(CART.pr)
# get indicator of leaves
leaves <- as.numeric(rownames(CART.pr$frame)[CART.pr$where])
dataset[,"nodeInd"] <- leaves
n.leaves = length(unique(leaves))
# step 4: Fit GLMM
library(lme4)
fixed = paste(c(TargetName, "as.factor(nodeInd)"), collapse = "~")
GLMM <- glmer(formula = formula(paste(c(fixed, random[[2]]), collapse = "+")),
data=dataset, family = family, control=glmerControl(optimizer="bobyqa"))
# step 5: get fixed effect from the model and put it into tree leaves
adjtarg <- unique(cbind(leaves, predict(GLMM, re = NA, type = "response")))
CART.pr$frame[as.character(adjtarg[,"leaves"]),]$yval <- adjtarg[,2]
GMET <- list(
Tree=CART.pr,
EffectModel=GLMM,
RandomEffects=ranef(GLMM),
BetweenMatrix = VarCorr(GLMM)
)c
lass(GMET) <- "GMET"
print(GMET$Tree)
library(rpart)
plot(GMET$Tree)
text(GMET$Tree)
x11()
rpart.plot(GMET$Tree)
# plot random effect
x11()
rand_intercept = ranef(GMET$EffectModel, condVar=TRUE)
dotplot(rand_intercept,strip=T, lty= 4)
rand_intercept
```

```
# VPC
psiA = as.numeric(summary(GMET$EffectModel)$varcor)
psiA/(psiA +pi^2/3) # 3.6%
predict.GMET <- function(object, newdata, type = "response", ...){
# get underlying tree structure
tree.aux = object$Tree
# substitute the row index as predicted value in each leaf
tree.aux$frame$yval = as.numeric(rownames(tree.aux$frame))
# get the leaf indicator for each newdata
newdata[,"nodeInd"] <- predict(tree.aux, newdata= newdata)
#predict using GLMM
predict(object$EffectModel, newdata = newdata, type = type,...)
}
# ROC analysis
test_pred <- predict(GMET,test,re.form=NULL,type="response",allow.new.levels=T)
p0 = seq(0,1,0.001)
sensitivity <- specificity_comp <- NULL
i=0
for(k in p0){
i = i+1
pred2 <- ifelse(test_pred > p0[i],1,0)
t11 = length(which(pred2 ==1 & test$GPAlower23==1))
t10 = length(which(pred2 ==1 & test$GPAlower23==0))
t01 = length(which(pred2 ==0 & test$GPAlower23==1))
t00 = length(which(pred2 ==0 & test$GPAlower23==0))
##### le formule di specificity e sensitivity erano sbagliate, così sono giuste
specificity_comp[i] <- 1 - t00/(t10+t00)
sensitivity[i] <- t11/(t11+t01)
}x
11()
plot(specificity_comp, sensitivity, type = "l", xlab='1 - specificity', cex.lab = 1.5)
lines(seq(0,1,0.01), seq(0,1,0.01), lty = "dashed")
i0 = 350
points(specificity_comp[i0], sensitivity[i0], pch = 1)
p0opt = p0[i0] # threshold value
p0opt
# ONE SHOT indexes
test_pred_class <- ifelse(test_pred > p0opt,1,0)
predicted = factor(test_pred_class, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
observed = factor(test$GPAlower23, levels = c(1,0),labels = c('GPA < 23','GPA >= 23'))
misc.table = table(predicted, observed)
misc.table
TP = misc.table[1,1]
TN = misc.table[2,2]
FP = misc.table[1,2]
FN = misc.table[2,1]
tot = dim(test)[1]
misClasificError <- (FP+FN)/tot
print(paste('Accuracy',round(1-misClasificError,4)))
print(paste('Sensitivity',round(TP/(TP+FN),4)))
print(paste('Specificity',round(TN/(TN+FP),4
```

Results GMLER

```
> summary(m1)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: dropout ~ (1 | DegreeProgramme.in) + WeiAvgEval1tot + TotalCredits1tot +
    AccessToStudiesAge
   Data: train
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   201.9    223.5    -96.0    191.9      547

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.5381 -0.1558 -0.0857 -0.0482 11.5333

Random effects:
 Groups              Name        Variance Std.Dev.
 DegreeProgramme.in (Intercept) 0.3198   0.5655
Number of obs: 552, groups:  DegreeProgramme.in, 13

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.01637    1.85574   0.548   0.5839
WeiAvgEval1tot    -0.11308    0.04653  -2.430   0.0151 *
TotalCredits1tot  -0.10174    0.01465  -6.946 3.76e-12 ***
AccessToStudiesAge  0.13001    0.06368   2.042   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) WAvgE1 TtlCr1
WeAvgEvl1tt -0.439
TtlCrdts1tt -0.118 -0.243
AccssTStdsA -0.825 -0.074  0.033

> summary(m1)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: dropout ~ (1 | NazTitolo) + WeiAvgEval1tot + TotalCredits1tot +
    AccessToStudiesAge
   Data: train
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   201.8    223.3    -95.9    191.8      547

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.4266 -0.1605 -0.0857 -0.0487 13.3392

Random effects:
 Groups    Name        Variance Std.Dev.
 NazTitolo (Intercept) 0.5447   0.738
Number of obs: 552, groups:  NazTitolo, 8

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.50130    1.97668  -0.254   0.7998
WeiAvgEval1tot    -0.09810    0.04153  -2.362   0.0182 *
TotalCredits1tot  -0.09781    0.01397  -7.000 2.56e-12 ***
AccessToStudiesAge  0.15451    0.07055   2.190   0.0285 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) WAvgE1 TtlCr1
WeAvgEvl1tt -0.302
TtlCrdts1tt -0.020 -0.369
AccssTStdsA -0.886 -0.052 -0.023
```

```
> summary(m1)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: HighTimetoDegree ~ (1 | DegreeProgramme.in) + NazTitolo + WeiAvgEval1tot +
    TotalCredits1tot + AvgAtt1tot
   Data: train.degree
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   281.3    331.1   -128.7    257.3      454

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1770 -0.3354 -0.1854 -0.0793  6.4363

Random effects:
 Groups              Name        Variance Std.Dev.
 DegreeProgramme.in (Intercept) 0.08208  0.2865
Number of obs: 466, groups:  DegreeProgramme.in, 13

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.37549    1.78323   1.893   0.0584 .
NazTitoloChina    -0.18841    0.64417  -0.292   0.7699
NazTitoloColombia -1.50180    1.11242  -1.350   0.1770
NazTitoloEgypt   -16.25426  362.04030  -0.045   0.9642
NazTitoloIndia    -0.58178    0.55528  -1.048   0.2948
NazTitoloIran      1.05687    0.44553   2.372   0.0177 *
NazTitoloPakistan  1.26394    0.71997   1.756   0.0792 .
NazTitoloTurkey   -1.64926    1.01357  -1.627   0.1037
WeiAvgEval1tot    -0.20275    0.07190  -2.820   0.0048 **
TotalCredits1tot  -0.04849    0.01105  -4.389 1.14e-05 ***
AvgAtt1tot         0.85483    0.17392   4.915 8.88e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Correlation of Fixed Effects:
            (Intr) NzTtlCh NzTtlCl NzTtlE NzTtlIn NzTtlIr NzTtlP NzTtlT WAvgE1
NazTitolChn -0.012
NazTitlClmb -0.104  0.164
NazTtlEgypt  0.003  0.000  -0.001
NazTitolInd -0.182  0.338   0.201  -0.001
NazTitolIrn  0.004  0.410   0.237  -0.001  0.449
NazTtlPkstn -0.128  0.243   0.151  -0.001  0.292   0.397
NazTitlTrky -0.017  0.194   0.119  -0.001  0.228   0.285   0.161
WeAvgEvl1tt -0.925 -0.071   0.032  -0.002  0.088  -0.219   0.005 -0.088
TtlCrdts1tt -0.063  0.013   0.056   0.000  0.020   0.104   0.027  0.212 -0.185
AvgAtt1tot  -0.372 -0.119  -0.018  -0.001 -0.108   0.055   0.086 -0.157  0.266
            TtlCr1
NazTitolChn
NazTitlClmb
NazTtlEgypt
NazTitolInd
NazTitolIrn
NazTtlPkstn
NazTitlTrky
WeAvgEvl1tt
TtlCrdts1tt
AvgAtt1tot  -0.345
convergence code: 0
Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?
```

```
> summary(m1)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: HighTimetoDegree ~ (1 | NazTitolo) + WeiAvgEval1tot + TotalCredits1tot +
    AvgAtt1tot
   Data: train.degree
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   287.4    308.2   -138.7    277.4      461

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2161 -0.3495 -0.2030 -0.1135  4.9758

Random effects:
 Groups    Name        Variance Std.Dev.
 NazTitolo (Intercept) 0.6938   0.8329
Number of obs: 466, groups:  NazTitolo, 8

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.40137    1.55658   1.543  0.12290
WeiAvgEval1tot   -0.17460    0.06345  -2.752  0.00593 **
TotalCredits1tot -0.04755    0.01045  -4.553 5.30e-06 ***
AvgAtt1tot        0.84492    0.16507   5.119 3.08e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) WAvgE1 TtlCr1
WeAvgEvl1tt -0.911
TtlCrdts1tt -0.021 -0.220
AvgAtt1tot  -0.376  0.238 -0.345
```

```
> summary(m2)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: GPAlower23 ~ (1 | DegreeProgramme.in) + NazTitolo + TotalCredits1tot +
    AvgAtt1tot + AccessToStudiesAge + YearsToFinishDegree
   Data: train.degree
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   390.5    444.4   -182.3    364.5      453

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.8972 -0.4280 -0.2206  0.0437  3.9132

Random effects:
 Groups             Name        Variance Std.Dev.
 DegreeProgramme.in (Intercept) 2.417    1.555
Number of obs: 466, groups:  DegreeProgramme.in, 13

Fixed effects:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.53267    2.07536  -0.257 0.797437
NazTitoloChina     -0.13013    0.51110  -0.255 0.799021
NazTitoloColombia   0.57123    0.58292   0.980 0.327113
NazTitoloEgypt     -0.54553    0.76744  -0.711 0.477181
NazTitoloIndia      0.35391    0.41124   0.861 0.389462
NazTitoloIran      -2.32837    0.49192  -4.733 2.21e-06 ***
NazTitoloPakistan  -0.13031    0.70899  -0.184 0.854177
NazTitoloTurkey    -0.75941    0.53524  -1.419 0.155948
TotalCredits1tot   -0.02395    0.01010  -2.370 0.017798 *
AvgAtt1tot          1.24742    0.22161   5.629 1.81e-08 ***
AccessToStudiesAge -0.18129    0.06888  -2.632 0.008488 **
YearsToFinishDegree 0.86606    0.24615   3.518 0.000434 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(m2)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: GPAlower23 ~ (1 | NazTitolo) + TotalCredits1tot + AvgAtt1tot +
    AccessToStudiesAge
   Data: train.degree
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   426.4    447.1   -208.2    416.4      461

Scaled residuals:
     Min       1Q   Median       3Q      Max
-16.2698  -0.4921  -0.3182   0.0897   3.8033

Random effects:
 Groups    Name        Variance Std.Dev.
 NazTitolo (Intercept) 0.4078   0.6386
Number of obs: 466, groups:  NazTitolo, 8

Fixed effects:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.351442   1.645476   2.037  0.04167 *
TotalCredits1tot   -0.040374   0.008284  -4.874 1.09e-06 ***
AvgAtt1tot          1.097136   0.156257   7.021 2.20e-12 ***
AccessToStudiesAge -0.187218   0.064554  -2.900  0.00373 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) TtlCr1 AvgAt1
TtlCrdts1tt -0.387
AvgAtt1tot   0.060 -0.258
AccssTStdsA -0.945  0.185 -0.198
```

```
Correlation of Fixed Effects:
            (Intr) NzTtlCh NzTtlCl NzTtlE NzTtlIn NzTtlIr NzTtlP NzTtlT TtlCr1
NazTitolChn -0.239
NazTitlClmb  0.008  0.239
NazTtlEgypt -0.122  0.202   0.155
NazTitolInd -0.174  0.391   0.298   0.247
NazTitolIrn  0.023  0.266   0.231   0.209  0.389
NazTtlPkstn  0.029  0.187   0.170   0.146  0.292   0.276
NazTitlTrky -0.152  0.282   0.225   0.187  0.369   0.287   0.176
TtlCrdts1tt -0.472  0.011   0.008   0.041 -0.056   0.021  -0.071  0.045
AvgAtt1tot   0.013 -0.056   0.024   0.014 -0.070  -0.130   0.042 -0.036 -0.271
AccssTStdsA -0.815  0.174  -0.128   0.077  0.146   0.017  -0.006  0.094  0.172
YrsTFnshDgr -0.455  0.106   0.078  -0.004  0.016  -0.223  -0.186  0.043  0.421
            AvgAt1 AccTSA
NazTitolChn
NazTitlClmb
NazTtlEgypt
NazTitolInd
NazTitolIrn
NazTtlPkstn
NazTitlTrky
TtlCrdts1tt
AvgAtt1tot
AccssTStdsA -0.175
YrsTFnshDgr -0.131  0.028
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: GPAlower23 ~ (1 | NazTitolo) + TotalCredits1tot + AvgAtt1tot +
    AccessToStudiesAge + YearsToFinishDegree
   Data: train.degree
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   418.4    443.3   -203.2    406.4      460

Scaled residuals:
     Min       1Q   Median       3Q      Max
-21.2260  -0.4745  -0.2957   0.0862   3.8386

Random effects:
 Groups    Name        Variance Std.Dev.
 NazTitolo (Intercept) 0.5143   0.7171
Number of obs: 466, groups:  NazTitolo, 8

Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.976141   1.806145   0.540  0.58888
TotalCredits1tot    -0.027788   0.009144  -3.039  0.00237 **
AvgAtt1tot           0.942929   0.164350   5.737 9.62e-09 ***
AccessToStudiesAge  -0.185283   0.064393  -2.877  0.00401 **
YearsToFinishDegree  0.678463   0.218261   3.108  0.00188 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) TtlCr1 AvgAt1 AccTSA
TtlCrdts1tt -0.484
AvgAtt1tot   0.147 -0.331
AccssTStdsA -0.855  0.159 -0.185
YrsTFnshDgr -0.410  0.407 -0.245 -0.011
```