POLITECNICO DI MILANO

School of Industrial and Information Engineering

Master of Science in Management Engineering

"Predicting high achieving students by using learning analytics in higher education institutions"

Supervisor:
Prof. Tommaso Agasisti

Author:
Rossella Falchi 904924

Academic Year 2018/2019

Table of contents

Index of Figures

Index of Tables

Abstract

The advent of the era of Big Data has played a key role in the latest developments in the higher education sector. Many institutions have started to use business intelligence and machine learning tools to handle large quantity of information in order to improve the efficiency in teaching, learning and in the educational management.

The main purpose of the present work is to predict high achieving students in the higher education environment. From the literature review about Learning Analytics has emerged the gap between the performance's prediction of at-risk and high achieving students. There are more and more studies regarding the drop-out or the retention issue, while the topic related to high achieving students has not been properly yet discussed.

This paper describes a study with bachelor's engineering students showing that high achieving students can be predicted since the first semester of their career. The study and identification of these talents will allow the improvement of talent-program generating benefits for all stakeholders in the university environment (students, the school management, employment industry and policy makers).

Keywords: *Learning Analytics*; *Prediction*; *High achieving students*; *Higher Education*.

Abstract – Italiano

L'avvento dell'era dei Big Data ha avuto un ruolo fondamentale negli ultimi sviluppi nell'ambito universitario. Molte università hanno iniziato ad adottare strumenti di business intelligence e machine learning per gestire grandi quantità di informazioni allo scopo di migliorare l'efficienza nell'insegnamento, nell'apprendimento e nella gestione accademica.

Lo scopo principale di questo lavoro di ricerca è quello di prevedere gli studenti di successo all'interno del contesto universitario. Dallo studio della letteratura nell'ambito Learning Analytics relativamente alla previsione delle prestazioni degli studenti, è emersa una lacuna tra la previsione degli studenti a rischio e quelli talentuosi. Ci sono sempre più studi indirizzati a prevenire la rinuncia agli studi o a migliorare la permanenza accademica, mentre il tema degli studenti che hanno già prestazioni molto elevate non è stato ancora trattato a dovere.

Questo lavoro di ricerca descrive uno studio applicato ai corsi di laurea di primo livello di ingegneria di un'università italiana, dimostrando che è possibile prevedere gli studenti di talento sin dal primo semestre della loro carriera. Lo studio e la valutazione di questi talenti permetterà un miglioramento nella gestione dei programmi di eccellenza generando dei benefici non solo per gli studenti ma per le università stesse, per il mercato del lavoro e per i policymakers.

Keywords: *Learning Analytics*; *Previsione*; *Studenti talentuosi*; *Università*.

Executive Summary

The advent of the era of big data has played a key role in the latest developments in the higher education sector. Many institutions have started to use business intelligence and machine learning tools to handle large quantity of information in order to improve the efficiency in teaching, learning and in the educational management.

The use of digital technology has generated a vast amount of educational data coming from many emerging technologies, including learning management systems (LMS), mobile learning applications, virtual and augmented reality interventions, cloud learning services, social networking applications for learning, video learning, robotics, and so forth.

Nowadays, Machine Learning (ML) and data mining have supported a wide range of applications such as medical diagnostics, stock market analysis, DNA sequence classification, games, robotics, predictive analysis, etc. (*Rastrollo-Guerrero et al.*, 2020). The using of these techniques in a learning context can help in the discovery of knowledge and hidden patterns within large amounts of data and making predictions for students' outcomes or behaviors.

Learning Analytics (LA) is the discipline that guides the process of analyzing educational data. In the recent years, LA has received growing attention from educational researchers and practitioners. The generally adopted definition for LA refers as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (*Ferguson*, 2012).

The analysis of institutional learning will allow better decision making, since the academic field will have immediate information to identify elements affecting student performance, to strengthen positive factors and reduce the negative ones, hence supporting development of new pedagogical models.

This research work aims at filling the gap regarding prediction of students' performance that has emerged in the literature study between students at risk to fail and high achieving students. Much focus and discussion have arisen around the following terms "intervention", "underperforming" and "at-risk", yet limited research effort has been dedicated to the study of the students already succeeding especially for those above the average, here in after high achieving or talented students.

The ultimate purpose of this work is to understand if high achieving students can be detected and to find the best algorithms in order to predict them at the beginning of their career in higher educational contexts. To this end, educational data from a technical University in Italy have been used in order to solve the above-mentioned subjects. To achieve this, the following research questions have been identified:

- Is it possible to predict high achieving students since the earliest stage of their academic career?
- Which are the relevant attributes to predict talented students?
- Which are the best models to predict high achieving students?

Before getting to the core of the present work, an analysis about the applications and purposes that drive to the employment of LA in higher education has been conducted. According to the goal of the analysis, different LA techniques are employed that vary case by case:

- Prediction: includes regression and classification methods; the predictive function is used to suggest effective ways to improve student's learning and performance.
- Clustering: generally used for clustering students and find typical participation's behaviors (i.e participation in class or in forum discussion).
- 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 Judgment: these may be proper diagrams, on the appropriate data, helping people to make sense of their findings.
- 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.

Predictive methods are the most frequent in the learning analytics literature. In this regard, ML allows to implement complex models that are used for prediction purposes. These models can be of great help to users by providing relevant data to facilitate decision-making. *Rastrollo-Guerrero et al.* (2020) performed a qualitative research study of 64 articles (almost 90% were published in the last 6 years), in which they gathered the different techniques used in the predictive analytics arena into main four categories: supervised ML, unsupervised ML , Collaborative Filtering, Artificial Neural Networks and an additional group dealing with other DM techniques.

As previously mentioned, it appeared as analytics implementations seem to be primarily concerned with students poised to fail. This constant language of “intervention” perpetuates an institutional culture of students as passive subjects (the targets of a flow of information) rather than as self-reflective learners given the cognitive tools to evaluate their own learning processes.

Then, in order to assess which factors and models are the best to identify and predict talented students, it has been decided to review research articles coming both from the learning analytics field and from education and psychology research papers. Only in this way it has been possible to build an overview on what a high achieving student is, since there were not enough studies in the LA field to have a clear comprehension.

By pooling these studies together, it comes out that:

- Virtual Learning Environments (i.e MOOCs) are the most used data source in the predictive analytics field;
- There is not a shared definition for high achieving students;
- There are some recurring characteristics that can help researchers in identifying high achieving students. These attributes have been categorized in the following five classes: behavioral characteristics, social and personality characteristics, cognitive abilities, time related characteristics and academic achievements.

Due to the aforementioned lack of the existing literature in the learning analytics field, there are not evidences of successful machine learning models already applied to educational data for predicting high achieving students.

Different authors agree with the fact that the identification of talents involves complex or hierarchical constructs, then multidimensional measurement methods and classificatory approaches (to data analysis) are to be recommended over traditional one-dimensional (i.e QI cut-off scores) methods (*K. A. Heller*, 2004). So, also from the point of view of the experts in the field, the topic of high achieving student’s detection is perceived as a real concern.

To fill the gap existing in the present literature regarding the prediction of students’ performance, the first step was to give a definition for high achieving students. They have been described in terms of academic achievement, so the final graduation score has been taken into account, and in terms of speed, so the time to graduation has been considered. One of the novelties of this work lies on the definition of the response variable that is a combination of the above-mentioned measures: a high achieving student is defined as who graduates within 3 years at bachelor's level with a final score higher than 101.

Even if most of the studies for predicting performances are based on dataset coming from Virtual Learning Environments (VLEs), the data handled in this project come from the institution's academic management system and include only demographic and academic information, aiming to prove that it is enough to look to these variables in order to get insights about top performer students. Data about bachelor engineering students from six cohorts (2010 to 2016) have been used.
The cohorts from 2010 to 2015 are selected to train the models and the next cohort (A.Y 2016/2017) to test them, as opposed to most of the works reported in the literature which use cross-validation, which means that the same cohorts are used to train and test the classifier. The aim of using successive cohorts is to check how well the results generalize over time so as to use the experience of one cohort to put in place some policy to detect weak or strong students for the following cohort (*Meceron*, 2015).

In order to understand if it was possible to predict high achieving students since the beginning of their career, two main models have been considered: the first one uses as reference the academic data at the end of the 1-semester while the second one uses as reference the academic data at the end of the first year. The two models have been compared in terms of predictions' accuracy and sensitivity, also taking into account the impact generated by an earlier identification of talents.
Simultaneously the research of influencing attributes for high achieving students' prediction has been conducted. Several machine learning models have been employed during the analysis. It has been shown that the most impacting variables for predicting talented students are, as expected, the academic performances, previous high school studies, the gender and the admission score.
In particular:

- Academic performances: the higher the average of attempts per exam and the higher the number of CFU gained at the end of the first semester, the lower the probability to be a high achiever.
- Gender: being a female is positive correlated with the probability to be a top performer.
- Previous high school studies: having attended the technical high school decrease the probability to be a high achiever. While the diploma score shows positive correlation with the response variable.

- Admission score: the higher the score of the university admission test, the higher the probability to be a top performer.

Also multilevel models have been applied, to test if random effects exist between the factors that might influence students' academic behavior and students' performance. It has been shown that the courses of Environmental Engineering, Physical Engineering and Biomedical Engineering show a positive correlation with the response variable, while Mechanical Engineering, Computer Science Engineering, Building, Architectural Engineering and Civil Engineering have a negative impact on the probability to be a top performer. The models reveal that around the 10-30% of variation in the response is attributable to the nested structure of educational data.

Logistic Regression, Random Forest and Multilevel Logistic Regression resulted to be the best performing algorithms. Multilevel Logistic Regression shows the greatest performance since the begging. Instead, for the others mentioned algorithms further approaches have been adopted to improve the measures of predictions' accuracy and sensitivity:

- Balancing the dataset;
- Apply ensemble methods: Stacking;
- Define an optimal threshold p_0 in order to minimize False Negatives (FN).

The findings revealed that applying the ROSE or under-sampling balancing techniques leads to significant improvements on results when a minority class as that of high achieving students is present (they constitute only the 12,4% of the total sample). Also the application of the third approach has been successful, thanks to which the performance predictions of the algorithms have exceed the 90% of accuracy and sensitivity. As far as the application of ensemble methods resulted not worthy due to the high correlation between the Random Forest's and Logistic Regression's predictions.

Once the results have been presented, it was worth to consider the opportunities which come by implementing learning analytics to identify talented students. Bradshaw et al. (2001) highlight as, from a practical standpoint, the extent to which a campus can attract the most academically talented students speaks directly to the campus' ability to successfully navigate legislative and public demands for accountability and assessment outcome that center on student success. In short, high achieving students increase the local faculty and national prestige, which directly and indirectly leads to an increase in funding opportunities.

The advantage given by the earlier identification of high achieving students can be spread across different actors in the faculty, including policymakers and the management. The present work establishes the foundation about the theme of LA for predicting talents at higher education level, giving also some hints regarding the opportunities deriving from this study.

The first point of interest can be the reinforcement of talent-management programs. As highlighted by *János Szabó* (2019) the aspect of talent-management programs, in the European universities are underrepresented compared to American ones. The introduction of a data-driven decision in the selection of the applicants for the "Alta Scuola Politecnica" (restricted group of excellent students at the Master of Science level at Politecnico di Milano), not anymore based only on academic merit but driven by a multi-factor analysis that learned from historical data. It could be thought to adopt the same system also for bachelor's students.

The second point regards the theme of university research. The possibility to identify top performers students can help the faculty to select the most promising talent to include in their research teams.

The third point is related to the connection between university and the employment industry. Companies are always seeking of talents, and if the university already knows who the brilliant students are since the beginning of their career, preferential and tailored job opportunities or events could be provided for high achieving students. This will help students to get acquainted about the existent proposals in the job market, to understand their interests and moreover to get in touch with successful persons who can became role models for these students, generating also the creation of a network outside the university. This can have a positive impact when the time of applying for a job will come.

The last point of reflection is given by accelerated studies. It has been pointed out the practical utility of the *time-to-graduation* criterion, since it is of interest also to policy makers because it can be a benefit in terms of efficiency and reduction in costs. Educational acceleration, for example content or grade-based, has been claimed to be advantageous for high abilities students, because it helps to increase academic achievement of those students who were accelerated, and it saves time and frees up other resources (*Steenbergen-Hu et al.*, 2012). So, for those classified as high achieving, the application of accelerated courses could be beneficial for them and for the faculty.

To conclude, the innovative aspects deriving from the present work are several and interesting for various stakeholders. The results coming from this research are fruit of a study about the learning analytics topic combined with the point of view of pedagogical experts, who discuss the theme of talents from decades. The main contribution to the literature is given by the introduction of a definition for talented students in the learning analytics field. While in the previous studies the aim to predict students', performance focused mainly on students failing or in danger of failing, here the purpose is to offer something also to those students already succeeding.

The path to follow is still long because the confirmation of such results needs to be supported by further researchers. It is needed to quantify the benefits coming from the implementation of the above-mentioned initiatives in favor of talented students. Anyway, having defined such classification is something new for the literature.

The present work represents the first footsteps to address high achieving students. Their identification from the beginning of their bachelor's degree will allow to take actions which could enhance their skills and generate benefits for the whole community: institution management, university research field and employment industry. It prepares the ground for further researches that will apply similar models with enlarged datasets, for example introducing VLEs data can help to characterize the behavior of talented students considering the interaction in virtual class or in forum discussion. In the light of the most recent events the faculty has put in place virtual classes that can generate vast amount of data. Future studies could also investigate in deep the actions to be undertaken for talents, evaluating the benefits of their implementation.

Chapter 1 – Introduction

1.1 Big data in higher education institutions

The advent of the era of big data has played a key role in the latest developments in the higher education sector. Many institutions have started to use business intelligence and machine learning tools to handle large quantity of information in order to improve the efficiency in teaching, learning and in the educational management.

The introduction of massive open online courses trough which is possible to gather several information like for example participation, interaction in forum discussion, the number and the time of resources download. These inevitably generate a vast amount of learning-related data (*Aldowah et al.,* 2019).
Big Data analysis in the higher education institutions (HEIs) is used relatively less than in other sectors but, as previously said, its use is growing exponentially. It is necessary to combine Big Data and business processes to improve institutional operations and support institutions in offering innovative services to students (*Jha et al.*, 2018).

However, a detach between business intelligence and the use of data for supporting decision-making still exists. The know-how for using educational data to advance learning and improve the student experience is not sufficient to manage the magnitude of data generated and stored by universities thanks to technology (*S. De Freitas et al*., 2015). To fulfill the task of data analysis, it is necessary to work with new specific technologies, such intelligent data, data mining, and text mining, among others. The convergence of these technologies with educational systems will allow the analysis of these data and transform it into useful information for all stakeholders (*Buenaño-Fernández et al.*, 2019).

1.2 The impact of Artificial Intelligence and Machine Learning

In the recent years, the demand for a revolution in education becomes stronger. Artificial Intelligence (AI) is shaping the future of everything from medicine, to transportation and to manufacturing. Researches in AI have recently shown impressive leaps in development also in the education industry, with the result that machine learning introduced many AI-based techniques.

Machine learning is a part of AI that offers the ability to extract new knowledge and patterns from data with huge benefit potential (*Ciolacu et al.*,2017). Arthur Samuel gives the first definition of Machine Learning as the field of study that gives computers the ability to learn without being explicitly programmed.

The impact of digital transformation on industries will displace many tasks and activities which have been traditionally performed by human beings (*Ciolacu et al.*,2017).
In the Gartner "Hype Cycle for Emerging Technologies 2017" three trends appear: (AI) Everywhere, Transparent Immersive Experiences and Digital Platforms. Some results of this report are presented in Figure below.

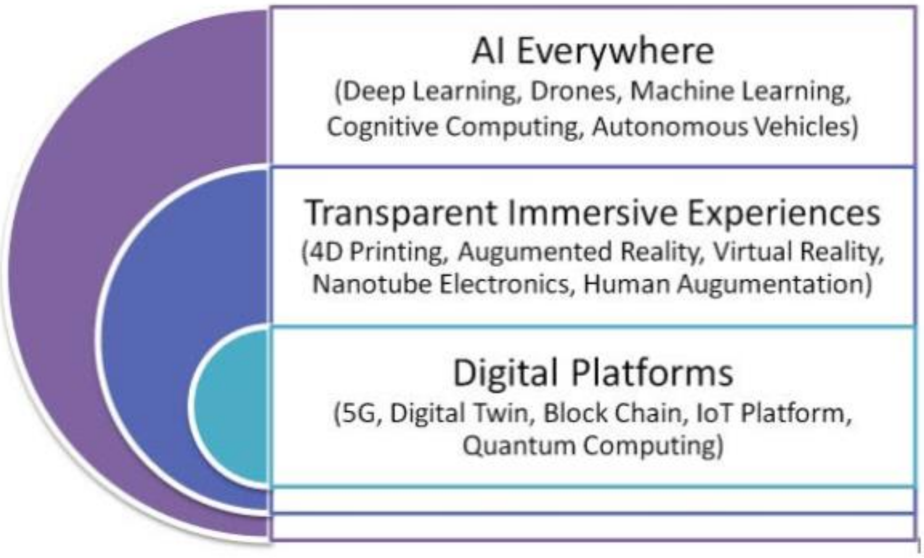

AI Everywhere
(Deep Learning, Drones, Machine Learning, Cognitive Computing, Autonomous Vehicles)
Transparent Immersive Experiences
(4D Printing, Augumented Reality, Virtual Reality, Nanotube Electronics, Human Augumentation)
Digital Platforms
(5G, Digital Twin, Block Chain, IoT Platform, Quantum Computing)

***Figure 1** "Hype Cicle for Emerging Technologies 2017". Source: Ciolacu et al., 2017*

Driving factors in real complex systems, as the education system, require handling multiple variables and multiple hypotheses with the substantial aide of computational resources for machine learning.
So, Artificial Intelligence can play a key role in Education identifying new drivers of students' performance and early disengagement cues, adopting personalizing learning, answering students' routine questions, using learning analytics and providing predictive modeling (*Ciolacu et al.*,2017).
Nowadays, machine learning and data mining have supported a wide range of applications such as medical diagnostics, stock market analysis, DNA sequence classification, games, robotics, predictive analysis, etc (*Rastrollo-Guerrero et al*., 2020). The using of data mining techniques in a learning context

can help in the discovery of knowledge and hidden patterns within large amounts of data and making predictions for outcomes or behaviors (*Aldowah et al*.,2019).

1.3 Learning Analytics drivers and purposes

Educational data mining (EDM), learning analytics (LA) and academic analytics are emerging disciplines that guide the process of analyzing educational data. This analysis is done through a variety of statistical methods, techniques, and tools, including machine learning and data mining (*Buenaño-Fernández et al.*, 2019; *Daud et al.*,2017).

In the recent years, "learning analytics" has become one of the fields in technology applied to learning that receives growing attention from educational researchers and practitioners (*Hui & Kwok*, 2019). Learning analytics involves the use of a broad range of data and techniques for analysis, which includes the development of metrics (such as predicators and indicators for various factors) to understand the current situation and measure teaching and learning effectiveness (*Lee et al.*, 2020).

Learning analytics has its roots in many fields of educational and technical research, including assessment, personal learning and social learning. It draws on theory and methodologies from disciplines such as statistics, artificial intelligence and computer science (Dawson et al., 2014).
The emergence of learning analytics as a field has been attributed to three principal drivers (Ferguson, 2012):

- **Big data**: the introduction of institutional databases and virtual learning environments (also known as learning management systems) means that educational institutions deal with increasingly large amounts of data and are looking for ways of using these to improve learning and teaching.
- **Online learning**: The rise of Big Data in education is accompanied by an increase in take-up of online and blended teaching and learning, and by growth in the number of learners worldwide learning informally using open educational resources and massive open online courses (MOOCs). There is therefore a worldwide interest in ways of optimizing learning in these settings.
- **National concerns**: Countries and international groupings are increasingly interested in measuring, demonstrating and improving performance in education and are looking for ways to

optimize learning and educational results in order to benefit society and the individuals within it.

Various political documents and laws in Europe and in the United States support the development and implementation of effective methods of LA in higher education. For example, the European Commission in 2016 noted that LA can contribute to the quality of teaching and learning and the modernization of educational systems in Europe (*Mustafina et al.,* 2018)

The enhancement in the quality of learning involves many groups of stakeholders with many more objectives (*Romero & Ventura*; 2013). Just to mention few, from the traces left by students and teachers in teaching and learning, learning analytics can enhance understanding of learning behaviors; provide useful suggestions for policymakers, instructors, and learners; and help educational practitioners to improve teaching and learning effectiveness (*Lee et al*., 2020). For recruiting fresh graduates, academic achievement is the main factor considered by the recruiting agencies (*Daud et al*.,2017).
The number of possible problems or objectives for each type of stakeholder is huge, Mustafina et al. (2018) identify different ways in which LA can be used by employees of institutions of higher education:

- LA can allow teachers to find out what resources their students use and how active they are.
- LA can allow students to implement the self- assessment on how much they have learned comparing to their fellow students.
- Real time information can give both teachers and students the opportunity to take timely, adequate action, depending on the situation.
- LA can help design the curriculum.
- LA can identify patterns of activity that lead to good learning and improve the student's performance.
- LA can identify students at risk of underachieving and enable them to improve their academic performance before they suffer any negative consequences.
- LA can offer actions and resources, which, most likely, will lead to a favorable result in studies for students.
- LA can be used to identify students with sudden changes in the learning process, which may indicate a wide range of non-academic problems. By identifying students who may face personal, emotional, medical, social or financial problems, LA can help teachers actively participate in the learning process and provide appropriate targeted support to students.

Nowadays, there is a considerable amount of research and studies that follow along the lines of learning analytics, among other related topics of interest in the educational area. Indeed, many articles have been published in journals and presented in conferences on this topic as can be seen from the following graph:

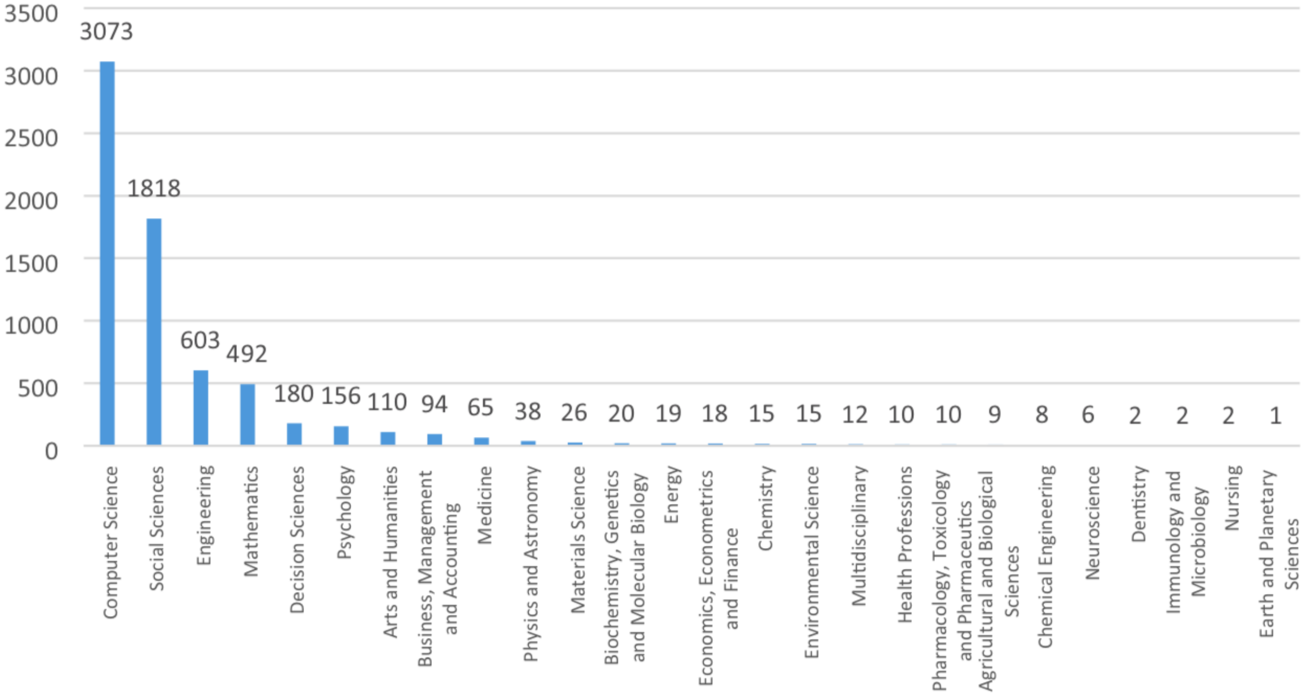


Figure 2. *Number of studies on learning analytics for individual applications domain. Source: Lee et al.,2020*

The aspect that has to be taken into account is definitely the impact of learning analytics' application on the whole university's system. Probably the major challenge that university has to be face is the arrangement of information systems and the seeking of experts in the field, otherwise the analytics' outcomes will not be successful. In order to understand the complexity hide behind analytics in the learning context, a practical example is provide comparing learning analytics with business analytics in e-commerce. The main objective of data mining in e- commerce is to increase profit. Profit is a tangible goal that can be measured in terms of sums of money, and which leads to clear secondary measures such as the number of customers and customer loyalty (Romero &Ventura; 2013). But as regards data mining in education, the performance measurements are more difficult to obtain since the main objective is the enhancement of learning. Then to evaluate the performances of data mining in this field, some proxy measures are considered like performances and dropout.

Being capable of handling such kind of data requires both expertise in the field of analytics and in the field of learning and if a solid information system is built, then the search of relevant information would be easier for the expert.

In figure 3 a diagram showing the applications of data mining in educational systems is presented. On the top are shown the activities to put in input

- To design, plan, build and maintenance: related to the issue discussed above;
- To use, interact, participate and communicate: more people-related initiatives. The application of learning analytics indeed requires the involvement of all the actors.

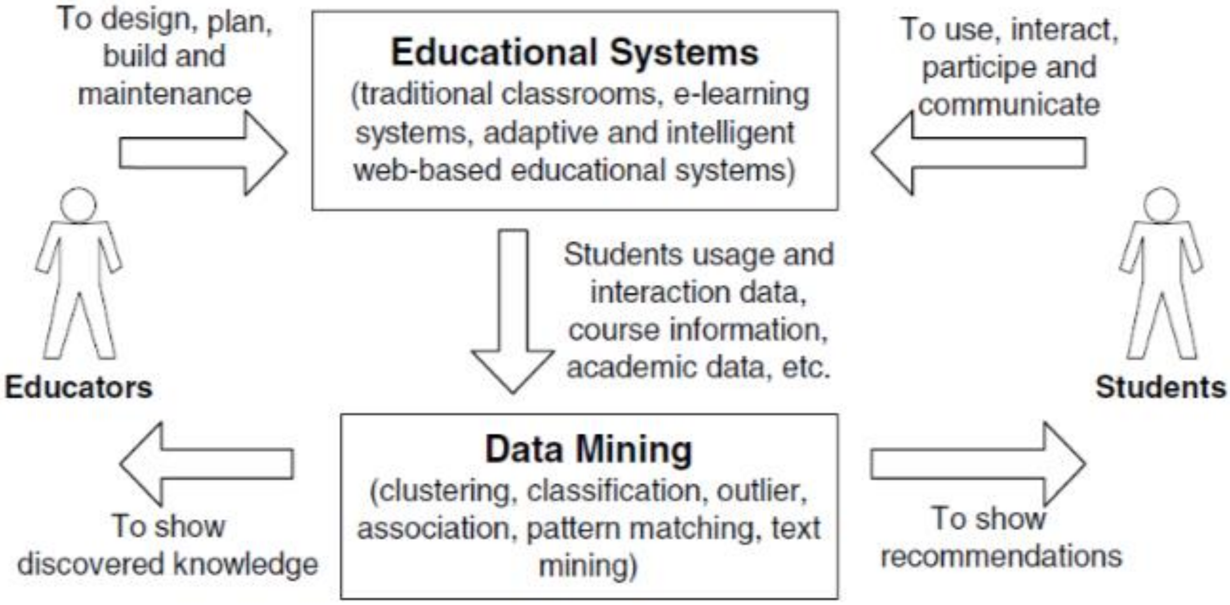

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 3. *Application of data mining at higher education level. Source: Daud et al., 2017*

On the bottom the two outgoing arrows present the outcomes generated by the implementation of data mining addressing both students and educators.
Actually, this graph is simplified, the involved stakeholders are a lot more and so also the input variables and the possible techniques, but it represents well the mechanism of analytics in a learning context.

1.4 Learning Analytics benefits and challenges

Through a careful analysis of big data, researchers can determine useful information that can benefit educational institutions, students, instructors, and researchers in various ways (*Avella et al*., 2016).

These stakeholder benefits include targeted course offerings, curriculum development, student learning outcomes and behavior, personalized learning, improved instructor performance, post-educational employment opportunities, and improved research in the field of education (*Avella et al*., 2016). Besides making it possible to diagnose problems and identify strategies for improving a course, LA also provides indicators of educational progress at local, regional and even national or international level (*Kumar & Hamid*, 2017).

A look on the ethical, legal, and risk concerns related to the application of LA should be considered too. As already mentioned, people who work with data must possess the requisite training in learning analytics to understand how to use the data productively to achieve meaningful results (*Avella et al*., 2016). Further, to protect privacy and maintain ethical standards, students should be fully informed about what data are collected, how they are used, who has access to it, and for what purposes the data will be used. Transparency is not only a legal requirement, but also the need to maintain trust between the university and students (*Mustafina et al*., 2018). In summary, mechanisms must provide appropriate transparency, data controls by students, information security, and accountability safeguards (Pardo & Siemens, 2014).

1.5 Talent management with Learning Analytics

In order to gain the potential benefits from modernizing education systems and improving learning outcomes, further work is still needed to make links between learning analytics, European priority areas for education and training, and the beliefs and values that underpin these areas (Ferguson et al., 2016). Researchers in the learning analytics field heavily focus on the performance's prediction area, e.g. reduction of drop-out rates and identification of at-risk students, while others, for example new and more learner-centered teaching methods, remain relatively untouched. Many articles claim that talent management is a very important aspect for higher education institutions. Despite of this, the studies which empirically investigate this topic are very rare.

It is true that critical attention has to be given to prevent students that are in danger of failing or dropping out of their studies, but at the same time we need to dedicate to locate new potential scientists among the university students (Cognard-Black & Spisak, 2019). Therefore, particular attention to talented students should be given, by identifying them in the mass-education environment and providing with the appropriate services.

Moreover, the market for academic top talent has become truly international, academics have become increasingly mobile, and schools have to compete for academic talent in the global arena. Without mentioning that, it is well known that European talent programs are less developed compared to the honors program in the US for example. The idea of taking advantage from data in order to work up in this direction could be beneficial for the educational system in general.

So, are the universities able to understand what drives the behavior of talents? And to understand which are the important factors to predict them? Learning analytics can help to achieve these purposes. It could be useful to understand if repeated behaviors exist among high achieving students, if it is possible to predict them and to address tailored initiatives to empower their value.

1.6 Thesis outline

The present work is organized as following.

- Chapter 2 will present the identified research questions followed by an overview on the present studies in the literature about learning analytics with a particular focus on predictions. Then, a more in-depth look will be given to studies regarding talent predictions and their enhancement, also considering the education-related scholarly literature to overcome the weakness in the learning analytics field.
- Chapter 3 will describe the data used for the analysis. An initial view on the raw data received will be presented, followed by the detailed description of the phase of data preprocessing until reaching the final picture of the dataset used for the analysis.
- Chapter 4 will describe the theoretical models employed for the analysis of data.
- Chapter 5 will describe the results obtained from the analysis of data trough graphical representation.
- Chapter 6 will deal with the discussion of results and the answer to the research questions, including also a section that considers the managerial and policy implications derived from the obtained results.

Chapter 2 – Research literature review

The present research work wishes to prove the existence of a gap in the present-day literature regarding students' performance prediction in the learning analytics field. The majority of studies, in order to increase the faculty's success, focus their attention at addressing students at-risk, but in doing so an important portion of students is not considered: the high achieving ones.

This chapter will be divided into six main sections: firstly, the research purpose and methodology are explained in detail, secondly a general overview about the Learning Analytics topic is developed and then a deepest analysis concerning predictive analytics concludes the general overview about the core theme.

Subsequently the performance of students' prediction, addressed by predictive analytics models, is analyzed in deep according to the following clusters: prediction of low performing students and prediction of high-performing students. Firstly, the existing literature and some real case of low performing students are presented.

Then the core of this project work is developed as following: few studies about talented students' prediction are investigated in the field of the learning analytics and also from the point of view of psychology and educational experts, in order to have a clearer picture of what a talented student is. Ending there will be a brief chapter of discussion about the concerns and issues arising from the implementation of LA in the university environment.

The final chapter is about the hints gathered from the literature analyzed and the closing considerations.

2.1 Research purpose and research questions

2.1.1 Research Purpose

This research work lays its foundations on the fact that educational institutions have an enormous amount of data at their disposal. It would be self-defeating do not use them to help faculties in their core activities, especially when we are in the era of data.

From the analysis of the literature in the field of learning analytics a significant gap emerged. Focusing in particular on students' performance prediction, it has been shown as there are a lot of studies focusing on detecting students at risk but very few hints are considered with regards to high achieving students.

Multiple steps have to be taken in order to meet the core objective. In fact, the first step was to gain a deep understanding about what the Learning Analytics is, which intentions and conditions led it to be a hot topic in the higher education field. Once reached a proper level of knowledge about the subject, the goal consisted in mapping the main objectives, to analyze the best methodologies and improvements obtained with the implementation of LA in the institutions' environment.

Afterwards, the analysis of predictive analytics models has been conducted, with particular attention to classification algorithms already applied in the academic area.

Having laid the foundation for the core objective, the final step was focused on catching few hints in the existing literature about the theme of high achieving students' prediction. Thereby, it has been decided to divide predictive analytics studies in two clusters: low and high-performing students' performance predictions. After a brief summary on the state-of-art regarding student at-risk, the central piece of this literature review will be focused on the high achieving students' matter.

2.1.2 Research Questions

The logical flow of the work is marked by the identified research questions emerged both at the beginning and during the progress of the research. They are displayed starting from a more general context, where they aim at addressing what learning analytics is and which are its main purposes. Then they restrict the focus on the prediction of students' performance, that is one of the main tasks covered by learning analytics.

The identified research questions at the beginning of the research are:

a) What is Learning analytics?
b) Which are the benefits derived from using learning analytics?
c) Which model are the best to predict students' performances?

At this step, the presence of a gap between prediction of students at-risk and top performer student's emerged. Subsequently, the aim of the research was focused on covering this gap. Therefore, the necessity to define new research questions raised.

The identified research questions emerged during the research are:

1. Is it possible to predict high achieving students since the earliest stage of their academic career?
2. Which are the relevant attributes to predict talented students?
3. Which are the best models to predict high achieving students?

The identified research questions constitute the backbone of the research work.

2.2 Research Methodology

In order to collect exhaustive information for the research purpose, different data and information sources have been used.

A primary source for exploring the Learning Analytics framework is represented by literature articles and papers. They represent a fundamental qualitative source for learning about previous researches and building deep understanding of a complex topic. Articles and reports were found and accessed through Scopus and Google Scholar databases.

2.2.1 Research Process Flow

Four consequently steps can be identified in the research process, which are illustrated in the figure 2.1.

A first step entails the analysis of Learning Analytics concept through a literature review, in the second the research is deepened with regard to Predictive Analytics. In the third the results of the students' performance prediction are split in two groups and in the final step the high achieving students' case studies are analyzed, leading then to conclusions.

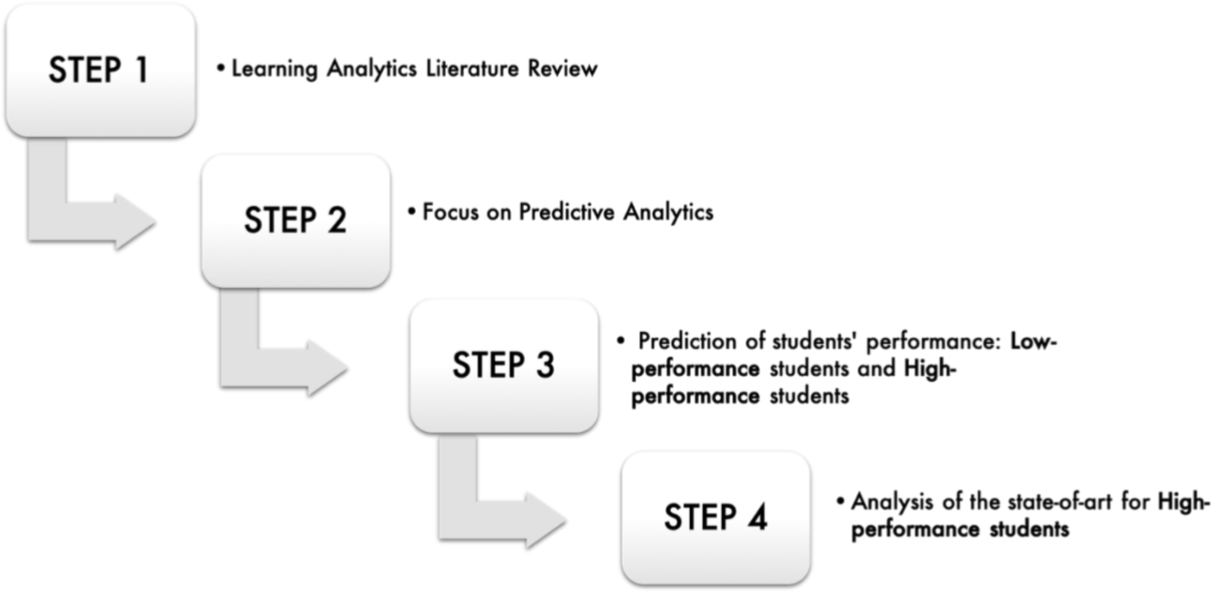

STEP 1
• Learning Analytics Literature Review
STEP 2
• Focus on Predictive Analytics
STEP 3
• Prediction of students' performance: Low-performance students and High-performance students
STEP 4
• Analysis of the state-of-art for High-performance students

Figure 4. *Literature research process flow. Source: Author's release*

The first phase started with the definition of the main keywords used, helpful to collect the appropriate material for the initial stage of the analysis.

We firstly applied the following filters:

- Search period: 2012-2020
- Keywords: learning analytics; data mining; higher education;
- Source: Google Scholar, Scopus

The search results obtained are shown in the following table, according to the different source of data:

Table 1. *Results obtained by source of data with the following keywords: Learning Analytics; data mining; higher education.*

Keywords	*Google Scholar*	*Scopus*
Learning analytics; data mining; higher education	1000+	1110

For this phase conference papers and journal publications have been used. Initially, around 30 papers have been selected after the initial filtering according to their relevance in terms of references. After having read their abstract, just eight of them have been chosen for a deep lecture and considered relevant for the research.

Then in the second phase, the keywords “prediction” and “students’ performance” have been introduced in order to find the proper articles regarding the predictive analytics tools. As it is shown in the following table, the restriction of the scope of the learning analytics application has reduce significantly the number of pertinent results obtained:

Table 2. *Results obtained by source of data with the following keywords: Learning analytics; higher education; prediction; students; performance.*

Keywords	*Google Scholar*	*Scopus*
Learning analytics; higher education; prediction; student’s performance	1000+	56

At this stage of the research analysis, some of the articles have been selected from Scopus and Google Scholar databases, others have been chosen directly looking to the references of the most interesting articles and their own cited articles, in order to find some remarkable paper to go deeper into some topics: at the end 12 articles have been selected in the STEP 2.

At the STEP 3, the analysis has been divided in two main sections: prediction of low-low performance students and prediction of top-performance students. Aiming at analyzing the studies regarding the prediction of students at-risk, the keywords "retention" has been introduced. Meanwhile for the high-achieving students different keywords have been used as "high-achieving", "high achiever", "top performance" and "talented". The following table shows the results obtained according to database sources:

Table 3. *Comparison of obtained results when searching "at-risk" and "high achieving" students*

Keywords	*Google Scholar*	*Scopus*
Learning analytics; higher education; retention;	1000+	91
Learning analytics; higher education; prediction; high achiever; students	900	1
Learning analytics; higher education; prediction; high achieving; students	1000+	4
Learning analytics; prediction; top performance; students;	-	1
Predictive; learning analytics; higher education; talented; students	-	1

From the table it is possible to highlight two major hints. The first one is that there is a relevant difference on the results obtained between the two sources (Google Scholar and Scopus). Actually, this difference is not significant since any of the results obtained on Google Scholar was relevant compared to the keywords used. Therefore, the first three rows of the table still shown the number of results obtained from Google Scholar because some articles have been chosen from the top ten results, but then to avoid misleading when reading data, it has been decided to add a dash for the following rows which indicates that there were not relevant results for the research.

The second fact that is possible to deduct, represents the big difference of results between the low-performance students (related to the focus of LA in retention) and high-performance students in the field of predictive analytics. This relevant point, that constitute the core of the project, will be discussed in deep in the following chapters.

At this step, eight articles have been selected for analyzing the studies as regards the prediction of student at-risk and the initiative put in place to improve the retention of students in the faculties.

As previously said, to understand which studies have been made about the high-achieving students' identification, several keywords have been used since the number of results obtained were not satisfying. In particular, just three specific paper has been selected as relevant since they have as a target a minority class of excellent student, as for example the prediction of top thinkers in computing. In order to compensate the absence in the LA literature regarding high achieving students detention, I explored as well the present literature in the field of education and psychology and I have selected some interesting articles published by the following journals: *Journal of the National Collegiate Honors Council*, *Canadian Center of Science and Education and European Journal of Psychology of Education*.

2.3 Learning Analytics overview

There is not a generally accepted definition of what Learning Analytics (LA) is, but the most refer to LA as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (*Ferguson*, 2012).

Learning analytics, academic analytics and educational data mining (EDM) are closely related research areas. The goal of academic analytics is to support the institutional, operational, and financial decision-making processes, while the overall purpose of LA and EDM is to understand how students learn (*O.Viberg et al*., 2018). In the following picture, Aldowah et al. show the relation between LA and EDM as the intersection between Data Mining techniques and the disposal of Big Data in the higher education field.

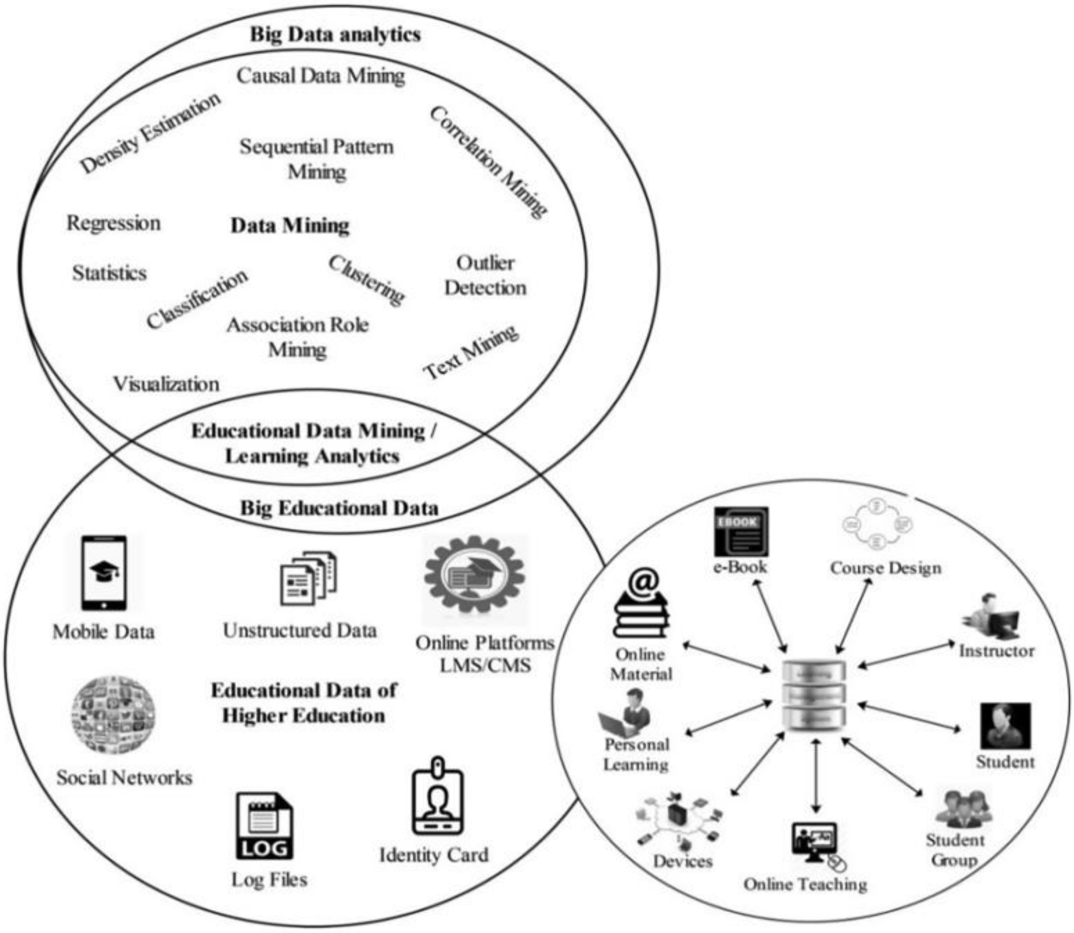


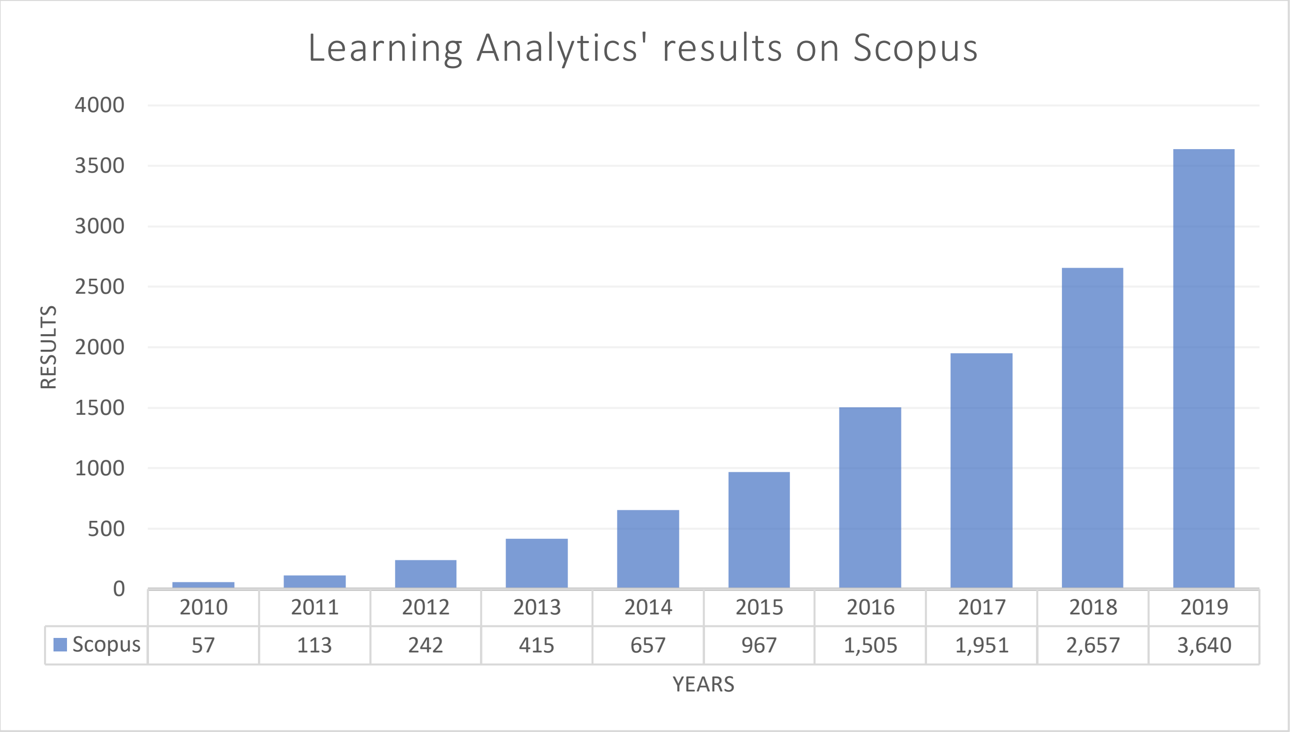
Figure 5. *Diagram highlighting the relation existing between LA and EDM. Source (Aldowah et al. ,2019)*

Both EDM and LA reflect the emergence of data-intensive approaches to education and therefore there are similarities between them, which suggests several areas of overlap, but there are also some relevant characteristics that distinguish these two approaches (O.Viberg et al., 2018):

- **Priority:** EDM focus mainly on an automated research, instead LA explores the results relaying on human judgement;
- **Application**: as a result of the priority, EDM is used as basis for automated adaption, instead LA models are implemented to inform instructors and learners;
- **Output**: the developed frameworks from EDM models are reductionist, they focus the analysis on individual components and their relationships. On the contrary, LA have stronger focus on understanding the complexity of systems as wholes.

The focus of this research work is on LA, how the LA research has been employed across different higher educational institutions and which benefits have been brought in the faculty system after the LA implementations.

The first International Conference on Learning Analytics and Knowledge (LAK conference), was in Banff, Canada, in 2011 and the second in Vancouver, Canada, in 2012. Today, the application of data mining in higher education is still in its infancy stage (H.Aldowah et al., 2018), but from the first LAK Conference the topic gained so much interest in the research community. In order to demonstrate the current increasing relevance in LA, the graph in figure 6 shows the number of results that a subscription-based tool such as *Scopus* returns when searching the exact term ‘Learning Analytics’, grouped by year from 2010 to 2019. As can be seen the numbers grow in an exponential way, showing the high interest in the topic.



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
■ Scopus	57	113	242	415	657	967	1,505	1,951	2,657	3,640

Figure 6*. Number of results per year that Scopus returns when searching the exact term 'Learning Analytics'*

Information provided by LA allows the customization of the training activities and the design of learning environments in accordance to the needs, interests and forms of interaction between teachers and students and between students themselves (*Kumar & Hamid*, 2017). Learning analytics can also provide students with timely information and recommendations as to their interests with two essential objectives: reflection and prediction. It will enable an iterative process of feedback, visual and effective, just-in-time feedback that allows the teacher or student to adopt correct teaching or learning strategies (*Kumar & Hamid*, 2017). The analysis of institutional learning will allow better decision making, since the academic field will have immediate information to identify elements affecting student performance, to strengthen positive factors and reduce negative factors, hence supporting development of new pedagogical models.

Dietz-Uhler and Hurn (2013) regard LA as a research model taking advantage of data analysis to inform on the actions and events taking place during the educational process. LA supports the education system by allowing curriculum adaptation, personalization, prediction and improvement of student success.

The implementation of learning analytics for higher education institutions brings also several benefits at an administrative level, such as improving decision-making and informing resource allocation, highlighting institution's successes and challenges, and increasing the organizational productivity (*B.Dietz-Uhler & J.E.Hurn*, 2013).

As mentioned above, there are several benefits for faculties that implement Learning Analytics models. These benefits are spread across several aspects in the university environment. Taking about the beneficiaries, they are students, professor, faculties' administrators, etc. It depends on what the purpose of the LA model is, because according to the "necessity" of universities, there are several ways in which educational data can be manipulated. Different LA techniques can indicate the number of problems and cases where each technique is more suitable (*Rastrollo-Guerrero et al*., 2020).

In this regard, *Merceron* (2015) subdivided the computational methods used to analyze educational data in the following four main categories according to the technique used:

- **Prediction**: students' performance prediction is the main assignment carried out by prediction methods. The prediction of performances can regard the drop-out rate, the fail/no-fail of the mark in a course or in a degree. The existing literature shows that it is possible to predict the drop-out ratio or the performance in a course or degree with a reasonable accuracy (mostly over 70%).
- **Clustering**: there are various clustering techniques and there are many tasks that use clustering. For instance, clusters students and find typical participation's behaviors in forums.
- **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**: these may be proper diagrams, on the appropriate data, help to grasp what happens at a glance and they form the essence of many dashboards and analytics tools.
- **Discovery with models**: this category is usually absent from conventional books about data mining or machine learning. It encompasses approaches in which the model obtained in a previous study is included in the data to discover more patterns.

Following the previous Merceron's categorization, the results of the existing review literature by *O.Viberg et al.,* show that **predictive methods** (including regression and classification) were the most frequent methods (32%), followed by **relationship mining** (24%) and **distillation of data for human judgement** (24%) including statistics and visualizations that help people make sense of their findings.

While, *Aldowah et al.* (2019) categorize data mining techniques of in four major categories according to the objective of analysis:

- **computer-supported learning techniques (CSLA)**, that uses statistical analysis in order to analyze students' information-searching and collaborative learning behaviors in a course context.
- **computer-supported predictive analytics (CSPA)**, in which the predictive function is used to suggest effective ways to improve students' learning and performance and evaluate the appropriateness of the learning materials. This is the focus for the majority of the studies.
- **computer-supported behavioural analytics (CSBA)**, which aim is discovering models of student behaviour, actions and knowledge.
- **computer-supported visualization analytics (CSVA)**, that regards method to visualize data.

They found that the application of data mining (EDM and LA) in higher education has the potential to enhance the current teaching and learning experiences by evaluating the interaction between learning materials, students' participation levels, and students' dropout and retention.

In this research we will focus in particular on predictive analytics. In the following chapters the theme will be investigated in deep with particular attention to the techniques employed and the most impacting variables for prediction purposes.

2.4 Predictive analytics

When it comes to understanding the main antecedents for promoting students' learning, Educational Data Mining and Learning Analytics can be used to predict students' performance and retention in particular course(s) based on the assessment and the evaluation of their achievement, participation, engagement, acquiring, grades, and domain knowledge in a learning activity (*Aldowah et al.*, 2019).

In this regard, Machine Learning (ML) allows to implement complex models that are used for prediction purposes. These models can be of great help to users by providing relevant data to facilitate decision-making. For example, predictive techniques allow teachers to detect students at risk and identify slow and fast learners, early intervention can thus prevent cases such as dropping out of courses or poor learning ((*Kumar & Hamid*, 2017)).

Rastrollo-Guerrero et al. (2020) performed a qualitative research study of 64 articles (almost 90% were published in the last 6 years), in which they gathered the different techniques used in the predictive analytics arena into main four categories: supervised ML, unsupervised ML , Collaborative Filtering, Artificial Neural Networks and an additional group dealing with other DM techniques, to include some works where similar objectives were tackled. The following graph in figure 7 shows the weight amount of each of these categories of techniques in the existing literature:

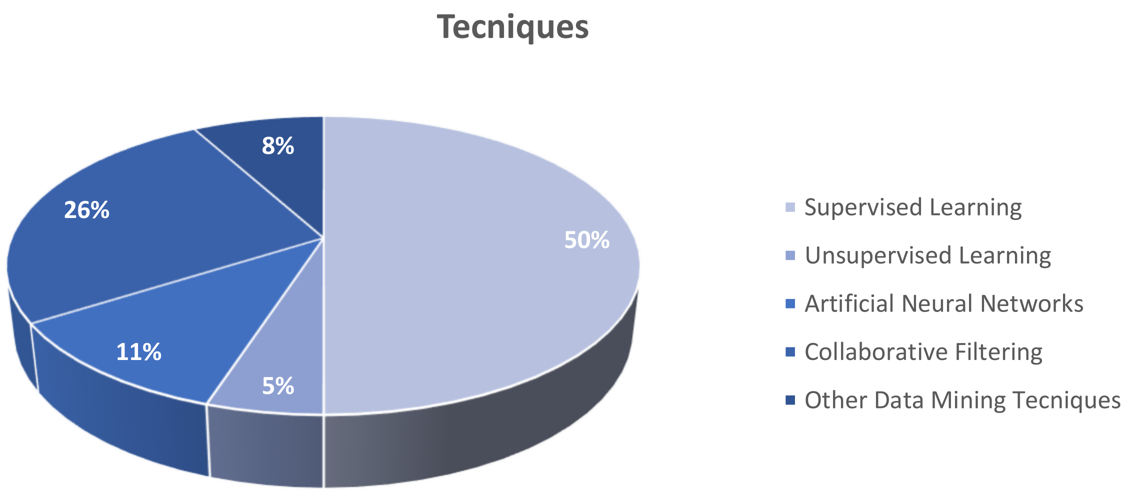


Figure 7. *Proportion of the different techniques used in predictive analytics. Source: Rastrollo-Guerrero et al., 2020*

In the past, mostly student performance is predicted by using different types of feature sets, such as, academic record, class participation and survey data.
Actually, some studies show that also variable like family income and expenditure play an important role in student performance prediction. In their research, *Daud et al.* (2017) investigates several family expenditure and personal information related feature sets based on learning analytics, in order to improve student performance prediction's results. They gathered these data from different universities of Pakistan during the period (2004 o 2011), outperforming existing methods and achieving the 86% accuracy in predicting the student that will complete the degree or dropout trough the Standar Vetor Machine algorithm.

Ciocalu et al. (2017) study the students learning behavior and its correlation with success in exams, using log data provided by the popular Learning Management System (LMS), Moodle. In this case, they put as input variables the logs of students' activities in the LMS (i.e Student ID, Time Stamp, Type of action, Object of the activity), while the response variable correspond to the exam result as a binary output variable (exam passed or failed). This study shows that the prediction accuracy is higher for Complete Virtual Course than for Blended Learning (Mix of traditional teaching in classrooms with E-Learning elements), and that non-linear kernel methods and neural networks are superior in terms of prediction accuracy.

However, the most widely used data source for predictive analytics is the student work data in a virtual learning environment or VLE (i.e LMS). Learning Management System are suites of software that provide course-delivery functions: administration, documentation, tracking, and reporting of training program, classroom and online events, e-learning programs, and training content (*Romero, Ventura*; 2013).
The trend in the use of learning systems aims to analyze the information generated by students (*Rastrollo-Guerrero et al.,* 2020). They record any student activities involved, such as reading, writing, taking tests, performing tasks in real, and commenting on events with peers (Romero, Ventura; 2013).

LMS data from course-related activities, such as discussion forums, content delivery and assessment, can be used to associate system level object with students' preferences, providing an opportunity for the instructor to gain a comprehensive view of possible learning outcomes or undesirable behavior among students (*H.Aldowah et al.*, 2018). For example, Smith, Lange, and Huston (2012) found that the frequency with which students log in to their LMS, how often they engaged in the material, their pace, and assignment grades successfully predicted their performance in the course (*Dietz-Uhler & Hurn*, 2013).

In order to show the great impact of such LMS data to predict students' performance, I catalogued 8 research articles from a total of 13 that applied techniques of predictive analytics in a real case study. The 8 articles selected use as source of information, the data coming from the Learning Management Systems of universities for predicting students' performances. In order to get a brief understanding on the use of this kind of data source and the impact in the predictive analytics environment, they are presented synthetically in the following:

Table 4. *Summary of learning analytics studies that uses LMS data for making predictions*

Document Title	*Author(s)*	*Year*	*Source*	*Model implemented*
Course signals at Purdue: using learning analytics to increase student success	Kimberly E. Arnold, Matthew D. Pistilli	2012	Proceedings of the 2nd International Conference on Learning Analytics and Knowledge	Course Signals relies on grades to predict students' performance, but also demographic characteristics, past academic history, and students' effort as measured by interaction with Blackboard Vista, **Purdue's learning management system.**
Identifying Successful Learners from Interaction Behaviour	Judi McCuaig, Julia Baldwin	2012	IEEE Transactions on Learning Technologies	Analysis of 17 blended courses with 4,989 students in a single institution using **Moodle LMS**, in which they predict student performance from LMS predictor variables and from in-between assessment grades, using both **multi-level and standard regressions**.
Foundations of dynamic learning analytics: Using	Sara de Freitas, David Gibson, Coert Du Plessis, Pat Halloran, Ed Williams, Matt	2015	British Journal of Educational Technology	Within the model's architecture, **data from the learning management system (LMS)**, financial aid system, and student system are combined to calculate

university student data to increase retention	Ambrose, Ian Dunwell, Sylvester Arnab			a likelihood of any given student failing the current course.
Multimodal Learning Analytics and Educational Data Mining: using computational technologies to measure complex learning tasks	Paulo Blikstein	2016	LAK '13: Proceedings of the Third International Conference on Learning Analytics and Knowledge	The author talks about "multimodal techniques", such techniques can enable researchers to have an unprecedented insight into the **minute-by-minute development of several activities** (multiple dimensions of interaction and social interaction).
Education 4.0 – Fostering Student's Performance with Machine Learning Methods	Monica Ciolacu, Ali Fallah Tehrani, Rick Beer, Heribert Popp	2017	23rd International Symposium for Design and Technology in Electronic Packaging	The authors study the students learning behavior and its correlation with success in exams, using **log data** provided by the popular Learning Management System (LMS), **Moodle**.
A Step towards Big Data Architecture for Higher Education Analytics	Sanjay Jha; Meena Jha; Liam O'Brien	2018	5th Asia-Pacific World Congress on Computer Science and Engineering	In this paper, an experimental study is conducted on data about 309 postgraduate students to find the **correlation between student's success rate and online activity/** student behavior.
Personalizing Computer Science Education by Leveraging Multimodal Learning Analytics	David Azcona; I-Han Hsiao; Alan F. Smeaton	2018	2018 IEEE Frontiers in Education Conference (FIE)	Predictions of students' performance were generated every week, by leveraging historical student data, prior academic history, **logged interactions between students and online resources**, and students' progress in programming laboratory work.

Rastrollo-Guerrero et al. (2020) categorized the objectives that predictive analytics aim to achieve in four main categories:

- Student dropout;
- Students' performance;
- Recommended activities and resources;
- Students' knowledge.

The focus of this research work will be initially on the first two objectives, student dropout and students' performance prediction, in order to understand the landscape of the existing literature concerning these two issues. Then, a more detailed review is made regarding the prediction of high achieving students' performance.

2.4.1 Student's Dropout

Student drop-out has been one of the primary focus of the literature in higher education. Indeed, university financing issues as well as the employment implications of university drop-out have made the understanding of withdrawing decisions a central concern for higher education policies and institutions' organization (*F.Belloc et al,* 2009). The dropout rates of university students generate a waste of resources for all actors in the education sector and even affect the evaluation processes of the institutions and among engineering students the dropout rate is even higher (*Buenaño-Fernández et al.*, 2019). Therefore, the issue of student retention is understood as a complex one involving support, social and pedagogic as well as performance factors.

Arnold & Pistilli (2012) develop an early intervention solution called "*Course Signal*" to provide real-time feedback to students at-risk. For predictions, they rely not only on grades to predict students' performance, but also demographic characteristics, past academic history, and students' effort as measured by interaction with Purdue University's learning management system.

On the same line, is the goal of the work of Buenaño-Fernández et al. (2019) which is to decrease the dropout rate, as well as provide real-time student follow-up to improve the education system. For predicting the dropout rate, they initially divide students according to similar behavior and then they apply classification trees to data, composed by students' historical grades. Finally, they develop friendly-use dashboards in order to give real-time feedbacks to students and give them the possibility to constantly monitor their career situation.

Thanks to a prompt action in an early stage of the students’ career, both these studies show that is possible to reduce the drop-out rate since the beginning, thus improving the number of graduates per cohort and so the quality of education in general.

Although there is a wide range of tools to predict students’ dropout intention, there is still no consensus about the best ways to understand the changing nature of this phenomenon regardless of the pedagogical style or activity used in the course (*Aldowah et al*., 2019).

2.5 Prediction of high achieving students' performance

2.5.1 Overview

The objective of this research work is proper to show the present gap between predictions of students at-risk and high achievers. Since there is not any research in the field of LA that specifically addresses the identification of top achiever students, additional specialized educational papers have been selected and evaluated, in order to understand which factors and implications has to be taken into considerations when seeking talented students.

The studies investigated in the following chapters relate either directly or in part to the goal of this study in order to provide either a contrast to, or corroboration of, the findings of my study. In the present research work we are trying to develop a model that can predict brilliant students in an early stage of their career. In this way, it would be possible to provide them with additional courses or learning experiences that can exploit fully their capabilities, consequently increasing the success of the institutions they belong.

2.5.2 What about high achieving students?

As previously observed, the topic regarding prediction of students' performance is the most questioned and studied in the learning analytics existing literature, but once the best techniques of prediction are identified, all the initiatives implemented for improving the quality of education and the universities' success are addressed only to one category of students: at-risk ones. But what about high achieving students?

In support of this statement, a survey of the current state of art in the learning analytics field reveals few popular words that offer a telling snapshot of the present focus of learning analytics; "intervention," "underperforming," "at-risk," and "prediction", these are some of the top repeats (*A.Kruse & R.Pongsajapan*, 2012). In particular, A.Kruse & R.Pongsajapan (2012) highlight precisely this matter: analytics implementations seem to be primarily concerned with students poised to fail, this constant language of "intervention" perpetuates an institutional culture of students as passive subjects (the targets of a flow of information) rather than as self-reflective learners given the cognitive tools to evaluate their own learning processes.

From a psychological point of view it has been adequately proved that a continual lack of challenge (due to giftedness not having been recognized), pressure to conformity (e.g. based on the fear of negative labeling effects) [..] and feelings of threat and envy could lead to behavior problems and conflicts between gifted individuals and their social environment (*K. A. Heller*, 2004).

For this reason, the purpose of this research work is to highlight the absence of studies regarding top performer students. In doing this, I have tried to understand which are the determinants for identifying high-achieving students. Firstly, from the point of view of learning analytics field and then I have picked some interesting researches from the point of view of psychology and education experts.

2.5.3 High achieving students from learning analytics perspective

As stated above, the current landscape of learning analytics is studded with studies about the prediction of students' performance with the target of student at-risk interventions.

From a practical standpoint, the extent to which a campus can attract the most academically talented students speaks directly to the campus' ability to successfully navigate legislative and public demands for accountability and assessment outcome that center on student success. In short, high achieving students increase the local faculty and national prestige, which directly and indirectly leads to an increase in funding opportunities (*Bradshaw et al.*, 2001).

I investigate such studies below that relate either directly or in part to the goal of this study (predict high-achieving students) in order to provide either a contrast to, or corroboration of, the findings of my study. Most of these researches, however, are limited either by relatively small sample sizes and/or by the lack of further researches on the topic. They have then, more specific additional limitations, which I have indicated below each presented article. While the data presented in these studies cumulatively begin to design a picture of how a high achieving student should be defined, they lack of a shared methodology and limiting the evaluation of the generalizability of the varied characteristics under consideration.

The first one is the work paper "*Can computational talent be detected? Predictive validity of the Computational Thinking Test*", in which *Gonzalès et al.* (2018) focus on the identification of particular talented students in order to boost the knowledge in the computational thinking (CT) area. They try to predict “**top thinkers of coding**” since an early stage., even before to learn coding. Identifying them, they can provide a fully and specific education to develop their high ability in the topic.

According to some studies they found in the literature, there is high correlation (50 < **r** < 85) between cognitive abilities (e.g., reasoning, processing speed, working memory) and academic performance (related to math subjects).

The sample consisted in 314 students enrolled in the subject of ‘Informatics’ who belong to different Spanish middle schools. As predicting variables, they use the following ones:

- **Academic variables**: *grade-point average* (GPA) in three different subjects (Informatics, Mathematics and Language);
- **Time related variables**: *Total minutes* (as the time spent by the student in a course) which is divided in *Video minutes* (time spent on video tutorial) and *Skill minutes* (time spent on coding task);
- **Learning variables**: expressed in two measures, *Badges earned* and *Points Earned*.

They collected these data through the information gathered from the Computational Thinking test (CTt), that is a multiple-choice test of 28 items to be completed in a maximum time of 45 min. According to Gonzalès et al. *computational top thinkers* could emerge as the students moved along coding course much faster than their regular peers did. At the end they show that the most impacting variables for predict talented coding students are:

- Participation in Free/Open Source Software (FOSS);
- The CTt has predictive validity with respect to the grade-point average (GPA) in Informatics, Mathematics and Language. A remark must bear in mind about this, findings suggested that ‘computationally talented’ might need specific tests to be detected, such as the CTt seems to be insensitive to them, because is not strictly related to compute but it is a *usual problem-solving test*;
- ‘Computationally talented’ students detected in middle school might have the ability to accelerate around 1 or 2 years in terms of curricular standards.

The most relevant remarks of this study in relation with my research are certainly the fact that the aim of *Gonzalès et al*. (2018) is to predict a niche category of top performer students since an early stage, so the variables use as predicting ones could provide an initial view on what we have to consider during the further phase of data preparation. But however, this study has some limitations if we regard to the general scope of the research. The first is that they implement the analysis on a sample of middle school students, and this can have different impacts both on the predicting variables and the results obtained. The second limitation could be given by the fact that all the students of the sample already chose the course of Informatics, so in a certain manner they were a influenced since the beginning, and for this the results of the study could be biased.

Still on the theme students able in coding, *Azcona et al.* (2018) developed a Predictive Analytics platform (PredictCS) for Computer Science courses that notifies students based on their performance using past student data and recommends most suitable resources for students to consult. The notification system is inspired by that of Purdue University. The innovation in PredictCS consists in its own design: the goal is to enhance and personalize the student's learning in programming modules. The platform sends notifications about performance, suggested programs and material on a weekly basis to students that opt-in. The input variables include a combination of student characteristics, past academic results, programming submissions and interactions with the material to generate predictions and recommendations to the students with the application of collaborative filtering techniques and other machine learning models. The recommendation engine as a module for PredictCS suggests code solutions for top performance students in the class as seen in the figure below:

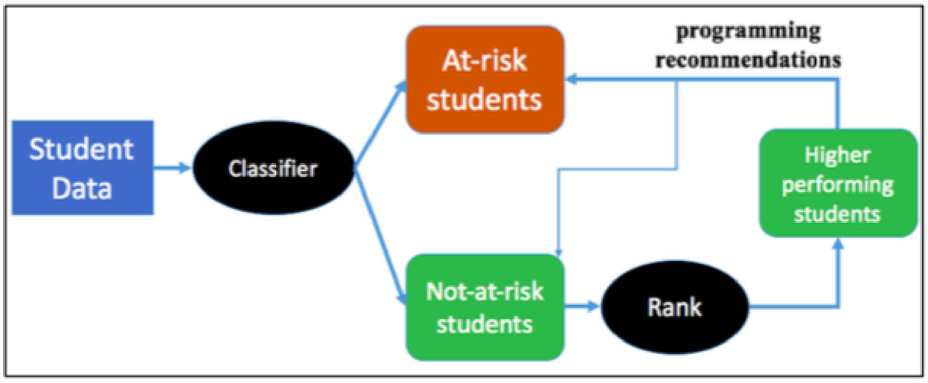


Figure 8*. Predict reccomandation engine. Source Azcona et al. (2018)*

The second investigated paper is published on the International Journal of Technology Enhanced Learning, it is entitled "*Time to focus on the temporal dimension of learning. A learning analytics study of the temporal patterns of students' interactions and self-regulation*". The authors are Mohammed Saqr, Jalal Nouri and Uno Fors (thereafter *Saqr et al.,* 2019).

In their study, they inspected the **temporal patterns** of students' interactions and self-regulation since it has been shown that the delay in performing learning tasks (known as procrastination) is a consistent negative predictor of academic achievement.

The sample included students from four different courses of dentistry at the second year of College. In order to predict the high and the low achievers' students from the sample, they apply the K-nearest neighbors (KNN) and the naïve Bayes algorithms, performing the 10-fold cross-validation as model validation. They obtain very good performance in terms of class recall: around the 88.3% for high achievers and 90.9% for low achievers.

They choose the **performance as target variable** for predictions, defining high achievers as the students with the score in the top 2/3rds of the class in each course. Their findings show that **high achievers** are more prone to have higher level of participation at the beginning of the course. Then as the time goes by, the difference in participation rates between low and high achievers is going to narrow. However, the time seems to have a great power in terms of prediction of academic performances.

The limitation of this study in regard to the purpose of my research work is mainly the temporality identified, that might be contextual and might prove to be generalizable if more papers will be authored regarding this topic in the future. Moreover, the sample of this research refers to student enrolled in a dentistry College, which differ in terms of methodology and subjects treated compare to a scientific College environment.

2.5.4 High achieving students from psychology and education-related scholarly literature

According to psychology and education experts, talented students are more opened to the experiences and more conscientious than their non talented peers (*Long & Lange*, 2002). An investigation used the Big Five (BIG5) questionnaire which proved that talented students are more conscientious, open, emotionally instable, and introverted (*Achtenberg*, 2005; *Cross,* 2018). Besides psychological features, also the GPA, learning strategy (*Cuevas, Schreiner, Kim& Bloom,* 2017), and behavioral features are used methods at investigation of this topic. The GPA is higher at honor students than non-honor mates (*Cognard-Black & Spisak*, 2019), but this does not prove to be true in every case (*Shushok*, 2006).

In particular, *Cognard-Black & Spisak* (2019) try to the define the honors student profile. In general, for referring to honors student there is not a conventional definition, but they are described as academically talented undergraduate students participating in an honors program or college. This lack of a specific definition is due to the absence of data-driven studies that focus on defining the characteristics of this talented profiles. Thus, it seems that also in the field of psychology and education a gap exists regarding this matter, but it seems more questioned in the education field from years compare on what is present in the LA literature. Precisely because of that, it has been decided to pick some of the most interesting and pertaining research papers and articles in order to have a more complete view about what a top performer student is.

In their the study *Cognard-Black & Spisak* (2019) gather information about 119 000 undergraduate students which took the Student Experience in the Research University (SERU) Survey in 2018. The SERU is a survey of undergraduate degree-seeking students, it collected information regarding nineteen university around the U.S. The 12,84% (15,280 students) of the sample reported participation in or completion of an honors program.

Later the sample will be reduced in order to have more significant data. In fact, the authors administer the survey during the summer, so for all the students at the first year the analysis would not be so meaningful because it would map a profile of a high school student, thus not representing the life of a student on campus. Instead the core of this research was that of targeting the higher education environment. So, the dataset includes just senior students.

Thanks to data gathered with the SERU Survey, the authors have information about college admission score, both high school GPA and undergraduate GPA, undergraduate major, frequency of engage in class and participation in undergraduate research.

As expected, the findings show that students in the honors group have substantial **higher GPAs** than non-honors students. Moreover, honors students are more likely to help the university in **conducting scientific researches**, to report having **study abroad** and to conduct **their own research** or project with the help of the faculty.

The limitation of this study is that the model for honors education in the United States differs significantly from that used in Europe and in Italy as well. Obviously, the concepts of these programs are different in institutes, universities, but still there are some common features. For example: the "talented" refers to the academically talented undergraduate students everywhere.

However, the aspect of talent-management programs, in the European universities are underrepresented compared to American ones. Much less universities have talent program in Europe, and the majority of these are found in the Netherlands (*János Szabó*, 2019).

János Szabó (2019) studied in particular the profile of academic/scientific talent, referring to those students interested to became academic researchers or professors. The study has been carried out in the setting of talent management programs of a Hungarian University. They recognize as talented students who take part to additional courses/opportunities besides the obligatory studies, in activities like research-seminars, lecture-series, vocational trainings, workshops or learning-groups.

The source of data was a questionnaire with 23 questions (open-ended questions; multiple choices questions; yes/no questions) addressed to faculty instructors. This questionnaire was around two main topics:

- Identifying talented students and cooperation with talented students;
- Own career of supervisor university teachers.

So, the aim of the author was to gather meaningful information to help the applying process of talent management programs (even honor programs) or even at PhD-applying process. The results of the questionnaire show that the topic is very interesting in the academic environment, since more than the 80% of the respondents declare that they care about the predictions of future scientific researchers.

If people surveyed agreed on what stated before, they were asked to provide a suggestion about the indicators that can be useful in order to predict the scientific career of an undergraduate student. The main factors emerged from the survey are listed below:

- The student puts the balance at his essays/papers on the quality rather than the length of the text [75.5% of respondents];
- In his questions, comments and papers, the student refers often to other learning materials and other science domains [60.9%];
- The student is interested **in scholarships** and **study tours abroad** [49.7%].

The results may suggest conceptions for talent-programs (honor programs) based on academic talent, for doctoral schools, and for any other institutes who works with career entrant scientist. The scientific reinforcement would be more effective if scientific programs/scholarships/PhD-programs used professional methods during selection process, instead of subjective choices, based on CV and motivation letter (*János Szabó*, 2019).

Kappe et al. (2012) have conducted a study with the aim to predict academic success in higher education. The subjects of the study were four cohorts of students enrolled in a course of human resource management (HRM) from a College in the Netherlands. Data were collected from the response of a survey that measured intelligence, BIG5 personalities, motivation and other personality traits. In addition, they consider also measures of academic achievement collected from the faculty's information system. The sample was constituted by 148 (net of 26 student dropped out before the end of the first year). So, the variables were summarized as follows:

- Predictor variables: includes all the information gathered from the survey. Then intelligence, BIG5 personalities (neuroticism, extroversion, openness, agreeableness, conscientiousness), intrinsic motivation, anxiety, environmental pressure, needs for status and motivation to study;
- **Academic achievements variables**: lectures courses, skills training, team project, internship, thesis, the overall **GPA** and the **time to graduation**.

Kappe et al. (2012) try to understand if there is correlation between the predictor variables and the academic achievement measures. According to the authors, the findings of this study have practical utility for admission boards and counsellors. Students who are intelligent and conscientious, and who therefore do not need additional pressure to study, seems to have the higher correlation with the academic achievements variables; they could participate in a special honors program in which they are challenged more during their coursework and for which they receive an additional recommendation on their diploma that distinguishes it from a normal diploma. Thus, their extraordinary potential would be reached, and their talent would not be lost.

In particular, it has been pointed out the practical utility of the *time-to-graduation* criterion, since it is of interest also to policy makers because it can be a benefit in terms of efficiency and reduction in costs. Educational acceleration for example, content or grade-based, has been claimed to be advantageous for high abilities students, because it helps to increase academic achievement of those students who were accelerated, and it saves time and frees up other resources (*Steenbergen-Hu et al.*, 2012).

2.5.5 Concerns/issues

There are a number of issues and concerns that should be highlighted in any discussion of learning analytics. In fact it is from the beginning that faculty have for the most part, relied on their intuition and hunches to know when students are struggling, or to know when to suggest relevant learning resources, or to know to encourage students to reflect on their learning (*Dietz-Huler Hurn*, 2013). But it is important to stress the fact that faculty intuitions are no going to disappear due to the implementation of LA models. Instead learning analytics promises to make these hunches and the resulting action more data-driven and easier to detect (*Dietz-Huler & Hurn*, 2013).

Campbell et al. (2007) provide a list of the issues and concerns that must be addressed before implementing any program or course of action on learning analytics. Some of these concerns include:

- Big brother: It may be threatening to some students and faculty to know that someone can "watch" and track all that they do.
- Holistic view: There is a concern that any data set, no matter how comprehensive, cannot take in to account other issues, such as interpersonal ones.
- Faculty involvement: Faculty need to be involved in order for learning analytics to have its greatest impact.
- Obligation to act: Are faculty and institutions obligated to use data to increase the probability of student success?

An issue frequently highlighted in discussions of learning analytics include profiling and how learning-analytics data will be used. Specifically, there is a danger of creating a profile of successful and unsuccessful students. More importantly, there is concern that a profile creates a set of expectations for the student and faculty (*Dietz-Huler Hurn*, 2013). Of course, students and faculty already have expectations – the issue is that learning analytics might add a set of data-driven expectations. The purpose of our study can help to overcome this issue in that sense. In fact, from our point of view the current studies are so focused on the identification of at-risk students, that the risk is became that of neglecting talented ones. The success of faculty instead, can be covered in both directions:

- identifying student at-risk and developing initiatives to help them;
- identifying high achieving students, providing additional resources in order to exploit their talent;

In this way the problem of labelling and profiled can be mitigate.

Data privacy is another very controversial theme in terms when we talk about learning analytics. There are legal and ethical issues, such as FERPA, that need to be addressed before faculty or institutions can make use of some student data (Campbell et al., 2007; Greller & Drachsler, 2012). Similarly, there is the issue of who the data belong to. Once the data have been warehoused, can anyone have access to it (Campbell et al., 2007; Greller & Drachsler, 2012)?

Finally, there is the issue of whether or not we are really measuring student learning, or are we just attempting to boost student retention and course completion (Watters, 2012). *Dietz-Huler Hurn* (2013) argues that if the types of data that are mined for learning analytics were consider, such as the number of course tools accessed in an LMS, or the number of posts “read” on the discussion forum, are these really proxies for learning? This is not to suggest that learning analytics cannot boost learning, but we need to be clear about what we are measuring and predicting.

Predicting high achieving students is a matter of academic performance and talent. At this level it is evident how the learning and the talent of the student are strictly correlate. It is not just a matter of score and success, above all the goal of this study is the identification of a tailored and talented group of students that can access to additional resources in order to take full advantage of their learning careers at university, generating benefits also for the whole educational community.

2.6 Conclusions

After having introduced the definition and the role of learning analytics today in higher education institutions, the main objectives and methodologies have been discussed, with a particular focus as regards of predictive analytics.

It has been highlighted the gap in the existing literature on predicting students' performance between students at-risk and high achieving students. Simply trough a research by keywords on Scopus database the huge difference in terms of results emerged. The proportion of research papers dealing with at-risk student (or retention) is 90 to 1, compared to the topic of high achieving students.

For this reason, in order to assess which factors and models are the best to identify and predict talented students, it has been decided to review research articles coming both from the learning analytics field and from education and psychology research papers.

These studies were very different each other, most because of the sample selected (too small/ from high school/from higher education/etc..) or because the educational system differs country by country. Even so they all agree with the absence of a common definition for talented student and the deficiency in providing them with tailored resources in order to improve their talent and the community's success.

However, by pooling these studies all together the findings show some repeated characteristics that can help researchers in identifying high achieving students among the whole. For a better understanding these attributes have been categorized in the following five classes: behavioral characteristics, social and personality characteristics, cognitive abilities, time related characteristics and academic achievement. More details are provided in the table below.

Table 5*. Summary of the main classes of typical attributes for a high achieving student*

	High achieving student typical attribute
Behavioral characteristics	Level of participation in class (traditional and virtual)
	Level of participation in discussion (traditional and virtual)
	Participation in extra activities/courses
	Prone to conducting scientific researches
	Report studying abroad experience
Social and personality characteristics	Conscientious
	Open to experience
	Emotionally instable

	Introverted
Cognitive abilities	Processing speed
	Working memory
	Intelligence (QI)
Time related characteristics	Time to graduation
	The participation in class is constant since the beginning
Academic achievement	First term GPA
	GPA at the end of the first year
	Final course grade
	Internship
	High school education
	Test scores

These five categories, all together, could provide an overall view of the significant attributes of talented students and could help universities and policymakers to take actions in order to use them for improving the success of the faculty and the quality of learning.

Due to the aforementioned lack of the existing literature in the learning analytics field, there are not successful machine learning models that have been applied to educational data for predicting specifically high achieving students.

However, different authors agree with the fact that the identification of talents involves complex or hierarchical constructs, then multidimensional measurement methods and classificatory approaches (to data analysis) are to be recommended over traditional one-dimensional (i.e QI cut-off scores) methods (*K. A. Heller*, 2004).

Furthermore *Cognard-Black & Spisak* (2019) have extracted the thought on the subject of *Achterberg* (2005), as well as several others who have surveyed the literature, she notes the lack of reliable, data-driven studies on the characteristics of honors students and call for more to be done. She concludes that honors students "*are not a homogeneous group with a set of absolute or fixed characteristics*" and that any "*firm conclusions about them should be held as suspect because empirical data about honors students are in extremely short supply*".

So, also from the point of view of the experts in the field, the topic of high achieving student's detection is perceived as a real concern.

Chapter 3 - Data available and preprocessing

3.1 Overview about the project

The aim of the present work is to predict top performer students with the application of machine learning models. Data was given by a technical university in Italy, hereinafter PoliMI (Politecnico of Milano), including Bachelor and Master Students data coming from engineering, architecture and design faculty. In our analysis we will consider only bachelor engineering students' data.

The final goal is to find out which are the relevant indicators for predicting high achieving students. Because, once these indicators are detected it will be possible to address personalized and challenging activities according to the student profile, thus increasing the satisfaction of these students and the success of the university. In fact, this research work can generate benefits for students, who could have personalized activities/courses/learning material/access to selective research programs/work experience; for the university, that could improve the level of efficiency in terms of resource management and also improve its reputation; for the government and labor industry in terms of gain by the employment of brilliant talent.

Basing on demographic and academic data, I aim to prove that it is enough to look into the first-semester careers in order to get insights about who are top performers students. Data about bachelor engineering students from six cohorts (2010 to 2016) will be used, trying to find out which are the characteristics that can define a talented student.
Binary classification models will be implemented in order to predict if the student is a top performer [Response variable = 1] or not [Response variable = 0]. High achieving students were defined as those who have a final graduation score higher than 101 and time to graduation within 3 years.

In this work the data were extracted from the institution's academic management system and stored in CSV format file. This information is periodically retrieved from the university's grades system and stored in an integrated data repository.
The data collection for this study follows the guidelines of the university and were in line with the privacy and user protection rights agreement. The personal or identifying data were completely anonymized and excluded from analysis.

The following graph could be helpful for the reader, giving him a comprehensive view to understand the logical path followed from data collection to the knowledge representation:

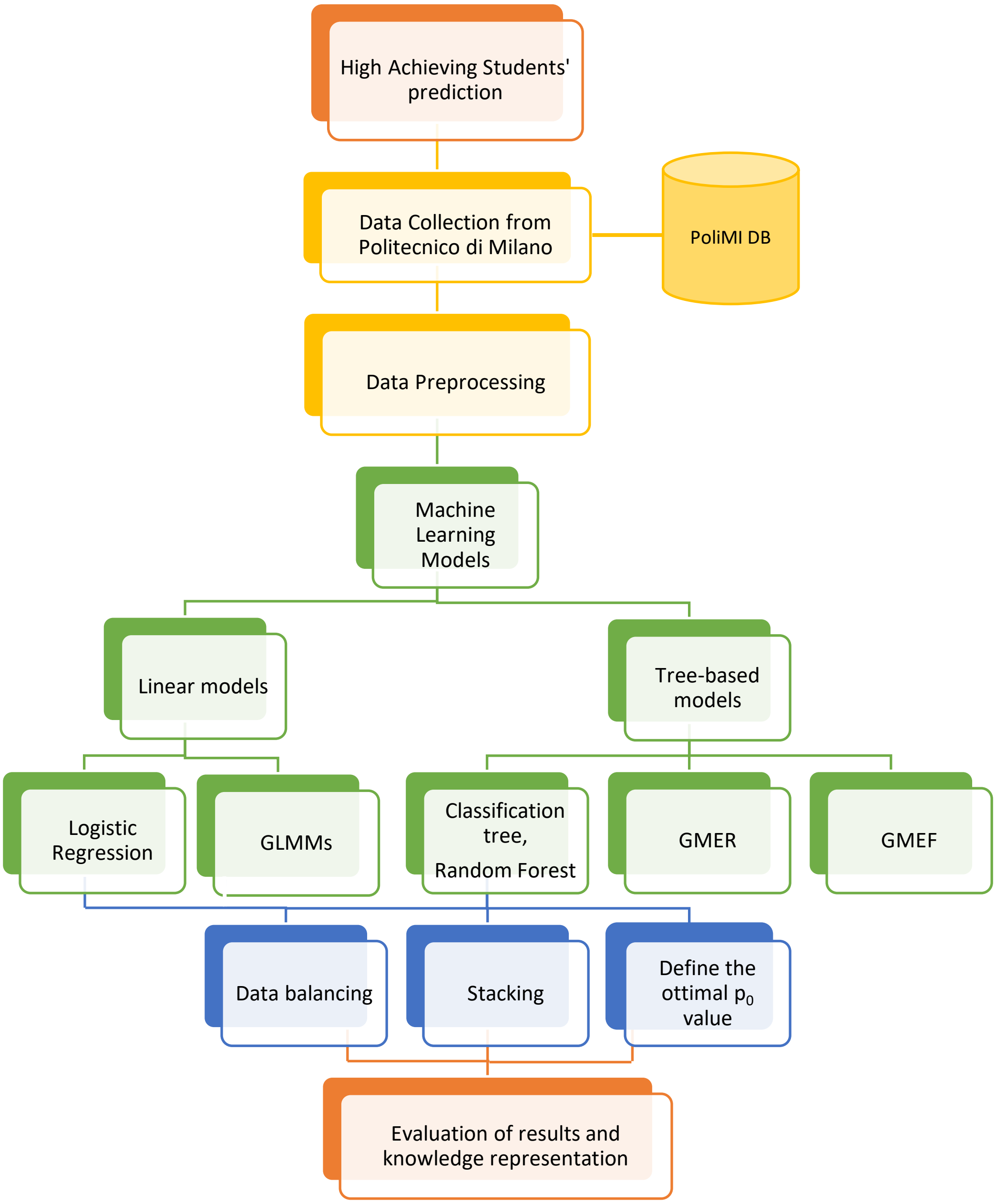

High Achieving Students' prediction
Data Collection from Politecnico di Milano
PoliMI DB
Data Preprocessing
Machine Learning Models
Linear models
Tree-based models
Logistic Regression
GLMMs
Classification tree, Random Forest
GMER
GMEF
Data balancing
Stacking
Define the ottimal p_0 value
Evaluation of results and knowledge representation

Figure 9. *Diagram showing the main steps of the project work*

To run the code, I used the R programming language. The structure and the relations among the data, on which I have worked on, will be described in the following sections. There will be an initial presentation of the datasets received from the university' information system, followed by the detail of the data preprocessing preparation phase, crucial to arrange the final dataset for the analysis.

3.2 Data preprocessing

3.2.1 Data available

The literature review analysis (chapter 2) reveal that LMSs are the most used data source in the predictive analytics field. For the present work we don't have data regarding the interaction with online educational platform but only demographic and academic data. Anyway, the introduction of data coming from LMSs could be kept in mind for a future development of the present work.

The data handled in this project comes from the institution's academic management system and as previously mentioned they include demographic and academic information of each students.

In the following picture the 10 datasets received are presented. The arrows and the different colors represent the distinctive relationship between datasets that helps to understand which keys have been used when merging the datasets.

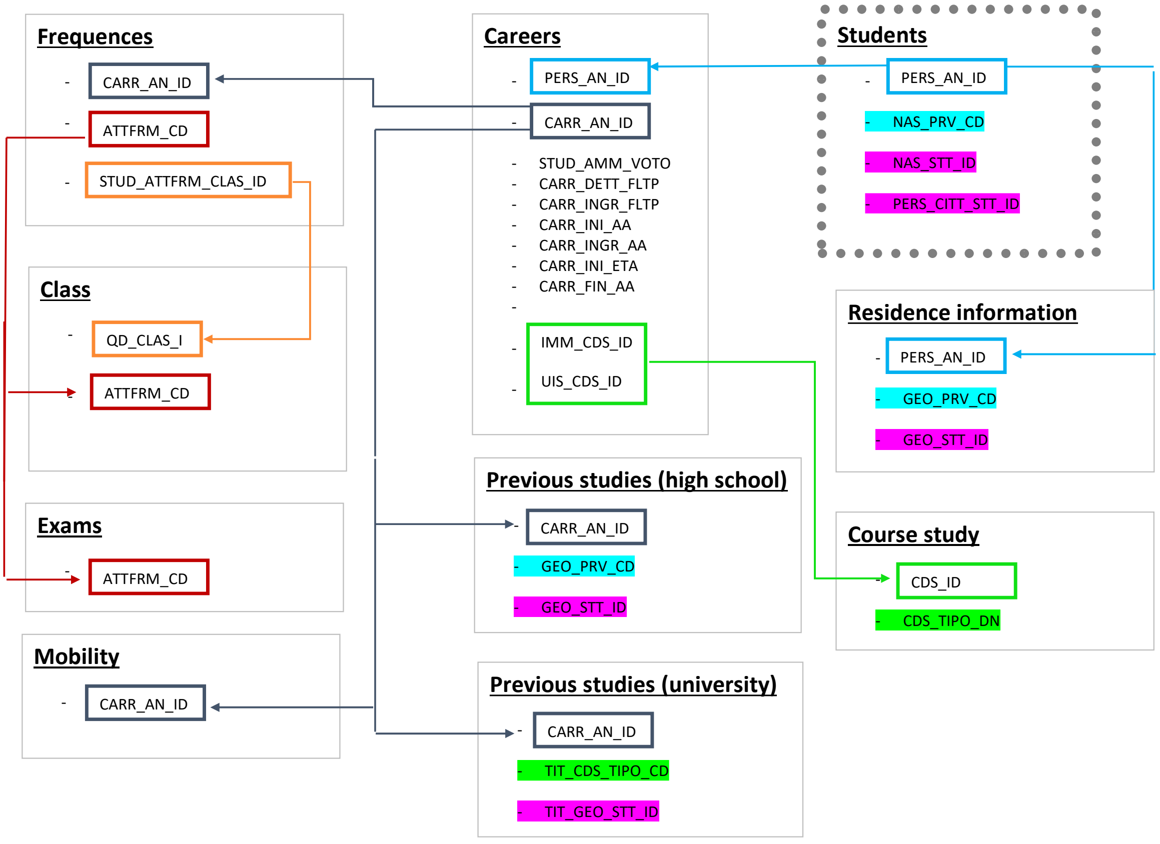


Figure 10. Raw datasets received from the university's information system

The picture has been simplified in order to be more easily understood, but actually the datasets include more variables than those present in the above representation. Then, the full datasets are presented below:

STUDENTS:

- **PERS_AN_ID**: random number which identifies each student
- PERS_NASC_YYYY: year of birth
- NAS_COM_PRV_CPL: if the student is 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 citenzship

CAREERS:

- **PERS_AN_ID**: random number which identifies each student
- CARR_AN_ID: identity number for each career
- STUD_AMM_VOTO: admission score
- CARR_DETT_FLTP: type of career (students enrolled in the university or incoming)
- CARR_INGR_FLTP: type of admission
- CARR_INI_AA: year of enrolment
- CARR_INGR_AA: year of enrolment at the university
- CARR_INI_ETA: enrolment age
- CARR_FLST: career status (A active, S suspended, D drop out, L graduated)
- CARR_FIN_AA: graduation year
- IMM_CDS_ID: ID course at the time of enrolment
- UIS_CDS_ID: ID course at the time of graduation

FREQUENCY:

- CARR_AN_ID: identity number for each career
- ATTFRM CD: identity number for identifying the subject taken by the student
- STUD_ATTFRM_FRQ_AA: year of n
- STUD_ATTFRM_FRQ_SEM: semester of the exam
- STUD_ACQSZ_CFU_VOTO: grade in the exam

- STUD_ACQSZ_CFU_LODE: if s/he gets 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
- 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
- STUD_ATTFRM_CLAS_ID: ID number for the association of the class and professor of the single course

CLASS:

- QD_CLAS_ID: identity number for each class
- ATTFRM_CD: identity 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: identity number of the subject taken by the student
- ATTFRM_DN: name of the subject
- ATTFRM_CFU: subject ETCS
- ATTFRM_SSD: area of the subject

MOBILITY:

- CARR_AN_ID: identity number for each career
- SI_FLTP: type of mobility [incoming; outgoing]
- SI_INI_DT: starting date of the mobility
- SI_FIN_DT: ending date of the mobility
- SI_PRGRM_DN: type of mobility [double degree, Erasmus...]

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

RESIDENCE INFORMATION:

- **PERS_AN_ID**: a random number which identifies each student
- GEO_PRV_CD: district of residence
- GEO_STT_ID: code of state' residence

STUDY COURSE:

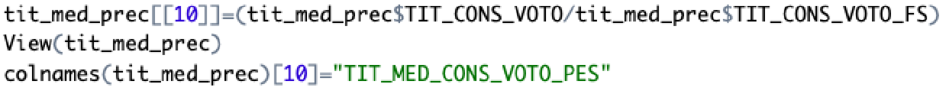
- CDS_ID: code of the course of study
- CDS_TIPO_DN: type of degree

3.2.2 Data gathering and integration

In the previous chapter the data available have been presented. Now, looking forward to the application of machine learning models for handling data and making predictions, I selected only the relevant data for the purpose of my analysis, creating a single and tailored dataset with the most impacting variables and with no redundancies.

As first step, I merge the datasets presented before in order to have all the information needed in a single dataset. Consequently, the merge with the dataset of **STUDENTS** and **CAREERS** has been done using the unique key PERS_AN_ID. Then I add the information about the address of residence (**RESIDENCE INFORMATION**), the previous studies at the high school (**PREVIOUS STUDIES HIGH SCHOOOL**), the previous studies at another university (**PREVIOUS ACADEMIC STUDIES**), the information about the mobility (**MOBILITY)** and the exams (**EXAMS**).

For what concern the two datasets including the history of previous studies (high school and university), the measures of the final grade were standardized, since not all students attended the same type of high school/university and thus the evaluation was different case by case. To overcome this problem the final grade (TIT_CONS_VOTO) has been weighted according to the total credits (TIT_CONS_VOTO_FS; i.e 100 for high school in Italy):



```
tit_med_prec[[10]]=(tit_med_prec$TIT_CONS_VOTO/tit_med_prec$TIT_CONS_VOTO_FS)
View(tit_med_prec)
colnames(tit_med_prec)[10]="TIT_MED_CONS_VOTO_PES"
```

The same has been done also for previous academic studies data.

3.2.3 Data validation and transformation

The purpose of data validation is to identify and remove anomalies and inconsistencies in the initial data, then data integration and transformation are to improve the accuracy and efficiency of learning algorithms.

Is extremely common that some records in the dataset may contain missing values corresponding to one or more attributes, so to partially correct some incomplete data I adopted techniques of elimination or substitution.

For those attributes containing just few missing records (Residence City and High School Previous Studies) I simply remove the missing records. Then for those variables with a significant number of missing rows I decided case by case:

- Diploma Lode: it is a categorical variable (Y/N). I choose to substitute the missing values with the value of the class No since it was the most frequent;
- Diploma Grade and Admission Score: supposing that past academic measures will have a great impact on our predictions, I just eliminate the missing rows avoiding the creation of arbitrary data. Thanks also to the large dimension of the dataset that allow to erase some records without having a huge loss of information.

Next, I differentiate data according to faculty following this classification: Engineering, Architecture and Design; then I split a second time the data regarding the Engineering faculty based on the type of course study: Bachelor’s and Master of Science degree.

Thanks to this initial categorization I was able to extract from the dataset Careers, all those careers concerning just undergraduate engineering students, obtaining a sample of 55 724 students enrolled at the university from the 2010 until September 2019.

Then I investigated how many students have undertaken more than one career (it means that for a single student ID there is more than one career ID) and I found out that the 12% of students have a double ID career or more, so this is means that they have change their degree course during their bachelor. Then I selected only the students with a single or a double **CAREER ID**, excluding those with triple or quadruple **CAREER ID**.

Afterward I plot the distribution per year of enrollment (see the graph below) and for my analysis I picked only those students enrolled in the first seven cohorts (2010 – 2016). In fact, since the last update of the received datasets was on October 2019, I would not have any relevant information regarding the students enrolled from the 2017 onwards, because the minimum duration of a bachelor degree course is 3 years at least, so the time interval wouldn't be sufficient to collect proper information about their graduation.

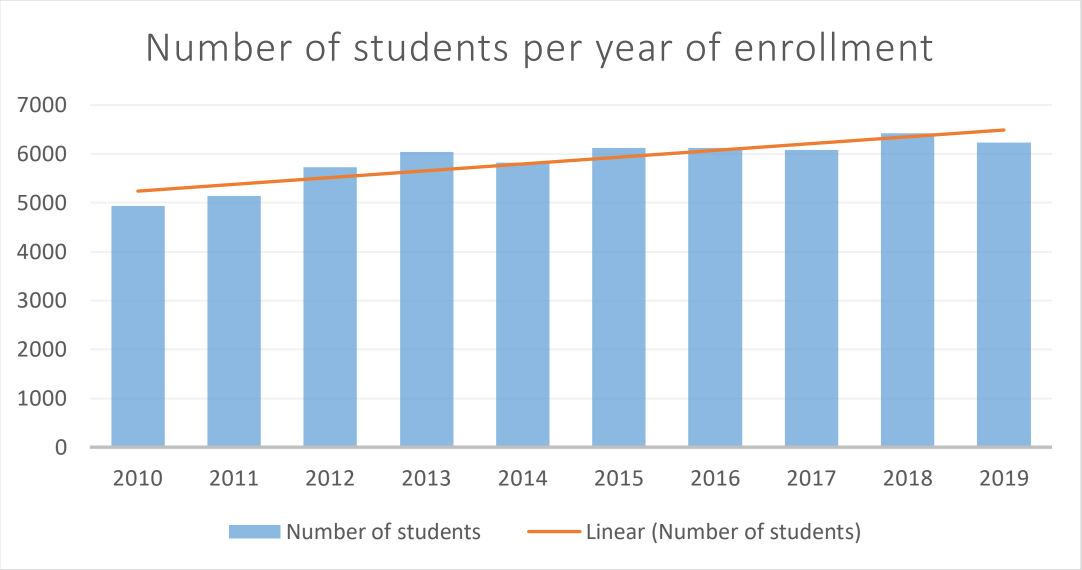
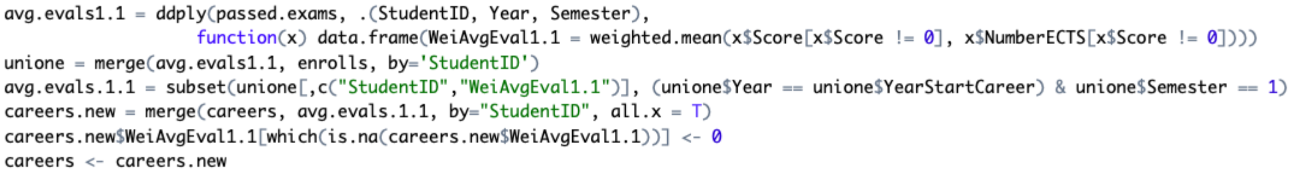


Figure 11. *Distribution of the number of students per year of enrollment*

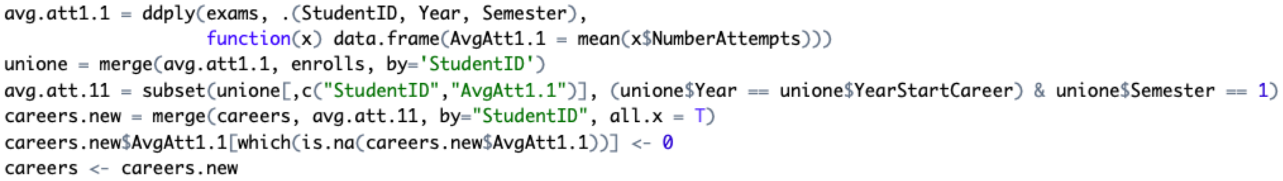
Once I obtained a more hands-on dataset, I enrich it with new descriptive variables in order to simplify the explanatory analytics. The first phase was that of creating 6 new variables related to performance:

- Compute the weighted average evaluation for the first semester for each student:



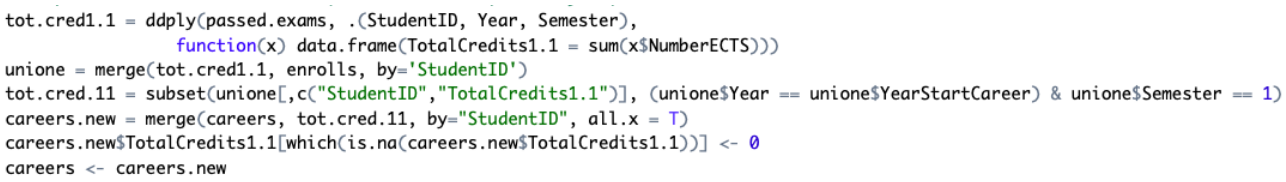
```
avg.evals1.1 = ddply(passed.exams, .(StudentID, Year, Semester),
                     function(x) data.frame(WeiAvgEval1.1 = weighted.mean(x$Score[x$Score != 0], x$NumberECTS[x$Score != 0])))
unione = merge(avg.evals1.1, enrolls, by='StudentID')
avg.evals.1.1 = subset(unione[,c("StudentID","WeiAvgEval1.1")], (unione$Year == unione$YearStartCareer) & unione$Semester == 1)
careers.new = merge(careers, avg.evals.1.1, by="StudentID", all.x = T)
careers.new$WeiAvgEval1.1[which(is.na(careers.new$WeiAvgEval1.1))] <- 0
careers <- careers.new
```

- Compute the average exams attempts for the first semester for each student:



```
avg.att1.1 = ddply(exams, .(StudentID, Year, Semester),
                   function(x) data.frame(AvgAtt1.1 = mean(x$NumberAttempts)))
unione = merge(avg.att1.1, enrolls, by='StudentID')
avg.att.11 = subset(unione[,c("StudentID","AvgAtt1.1")], (unione$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 in the first semester for each student:



```
tot.cred1.1 = ddply(passed.exams, .(StudentID, Year, Semester),
                    function(x) data.frame(TotalCredits1.1 = sum(x$NumberECTS)))
unione = merge(tot.cred1.1, enrolls, by='StudentID')
tot.cred.11 = subset(unione[,c("StudentID","TotalCredits1.1")], (unione$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
```

It has been done the same for the Weighted Average Evaluation, Average Exams Attempt and Total Credits Obtained at the end of the first year.

Finally, I create few dummy variables for those qualitative predictors present in the dataset:

Mobility.d { 0 the student did not study abroad
1 the student studied abroad

ChangeDegree.d { 0 the student did not change degree
1 the students changed degree course

The same logic has been applied for DiplomaLode.d [0- No; 1-Yes] and Sex.d [0- Male; 1-Female]. At this point, I decided to remove the information about previous academic studies [Yes/No] since the class of Yes included just few records and it wouldn't be significant for the analysis.

Some qualitative attributes in the dataset have more than two levels, so a single dummy variable could not represent all possible values, so I create additional dummy variables (n-1 levels), as shown for the following variables Residence.d and PreviousStudies.d:

LEVEL /RESIDENCE.D	2	3
MI	0	0
OFF-SITE	1	0
PENDULAR	0	1

LEVEL /PREVIOUSSTUDIES.D	2	3	4
SCIENTIFIC	0	0	0
CLASSIC	1	0	0
OTHER	0	1	0
TECHNICAL	0	0	1

3.3 Descriptive analysis

At the end of the data preparation phase we obtained an overall sample of 36 899 students enrolled from 2010 and 2016. In this chapter the final data will be described in-depth in order to highlight the relevant features of each attribute contained in the dataset, using graphical methods and calculating summary statistics, and to identify the intensity of the underlying relationships among the attributes.

Before entering in the analysis, I want to stress the aim of the study; I choose to investigate the prediction of high achieving students due to the lack of the theme in the existing literature. In order to do so, I tried to define what a high achieving student is, thanks also to the findings coming from the literature and the common concept. Then, I considered as performance attributes the final graduation score and the years to finish the degree. The following graphs show the distributions of these attributes:

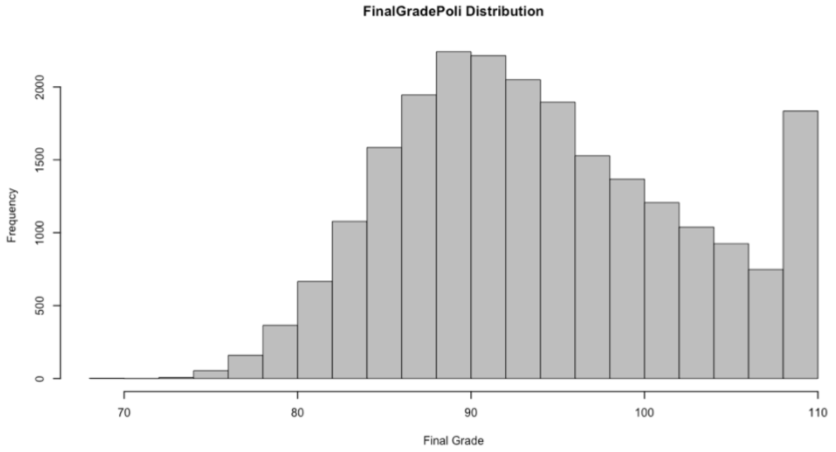


Figure 12. *Distribution of the graduation score in the sample*

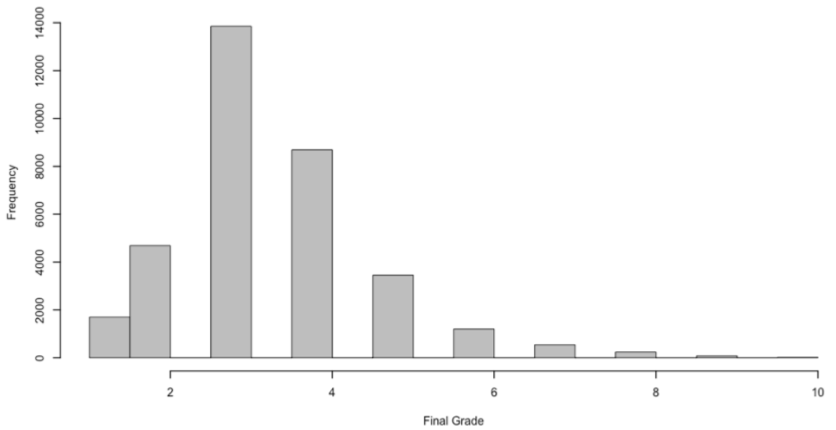


Figure 13. *Distribution of the years to finish the degree in the sample*

After having analyzed these two attributes the response variable has been defined according to the following considerations:

- I set a cut-off value of 101/110 on the final graduation score since it represents the boundary of the 3rd quartile. So, the group I select represents the 25% of the total sample.
- Then I reduced this group of students according to the time-depending variable. Imposing 3 years (and lower) as the cut-off value (12,4% of the sample).

According to these considerations the response variable **Y** is coded as a binary variable in this way:

- Y = 1 if the graduation score is higher than 101 and the time to finish the degree is equal or lower than 3 years;
- Y = 0 if FinalGradePoli < 101 and YearsToFinishDegree > 3years.

In the following graph is possible to see the trend of high achieving students per year of enrollment, it seems that the numbers increase as the time goes by.

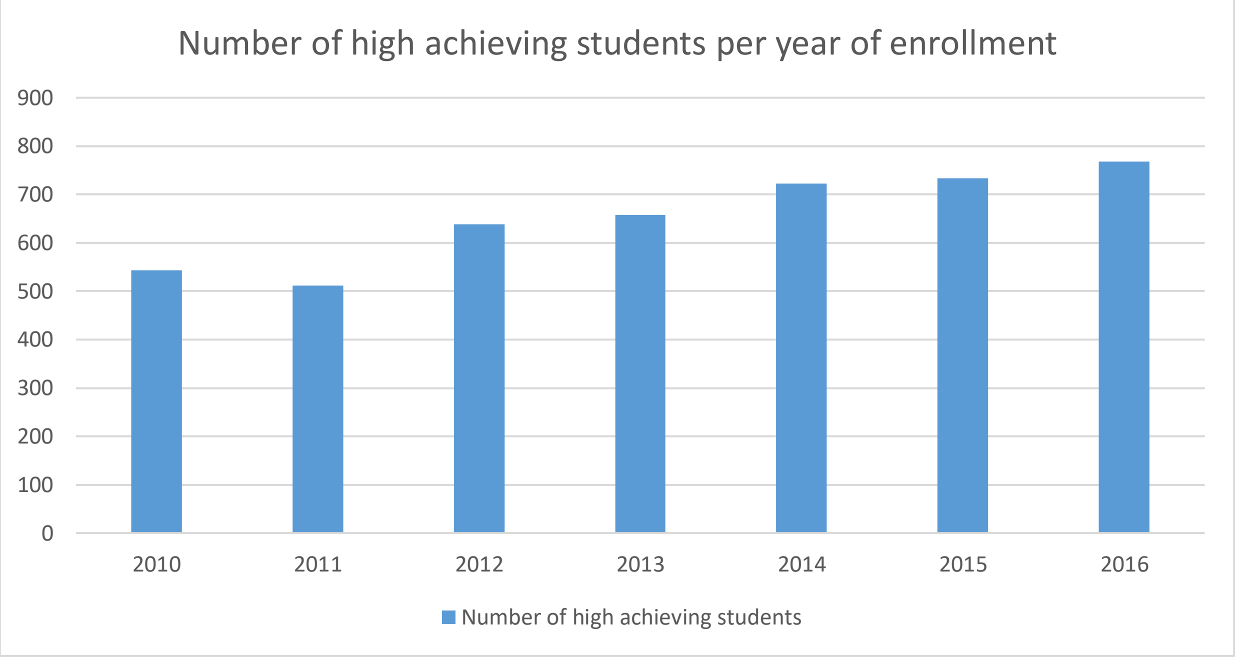


Figure 14. *Number of high achieving students per year of enrollment*

Table 6. *Summary of the final dataset used for the core analysis*

Variables	Variable Type	Variables Description	Possible Values
WeiAvgEval1.1	Numerical	Weighted average evaluation of each student at the end of the first semester	From 0 to 30
AvgAtt1.1	Numerical	Average attempts per exam of each student at the end of the first semester	From 0 to 7
TotalCredits1.1	Numerical	Total credits obtain by each student at the end of the first semester	From 0 to 80
WeiAvgEval1tot	Numerical	Weighted average evaluation of each student at the first year	From 0 to 30
AvgAtt1tot	Numerical	Average attempts per exam of each student at the end of the first year	From 0 to 5
TotalCredits1tot	Numerical	Total credits obtain by each student at the end of the first year	From 0 to 127
AccessToStudiesAge	Numerical	Age at the time of enrollment	From 17 to 60
DiplomaGrade	Numerical	High school diploma score	From 0.6 to 1
AdmissionScore	Numerical	The score performed by each student at PoliMI admission test	From -12.33 to 100
Residence.d	Categorical	It gives an indication on the residence of each student	From Milan; from countryside; off-site
Mobility.d	Categorical	If a student has reported an exchange study during his/her career	exchange; no exchange
ChangeDegree.d	Categorical	If a student has changed the degree course during his/her career	degree course changed; same degree course
Sex.d	Categorical	Gender	female; male
PreviousStudies.d	Categorical	It indicates the type of high school studies	Scientific, Classic, Technical, Other

DiplomaLode.d	Categorical	If the student graduated in high school cum laude or not	cum laude; no laude
DegreeNature	Categorical	The course of study in which the student graduated	21 different course study
Y	Categorical	It is the **response variable.**	1 if the final grade is higher than 101/110 and time to graduation <= 3 yrs; 0 otherwise

Below the most significative results of the descriptive analysis are presented, trying to understand which is the distribution of the response variables according different attributes:

The following Pie Chart shows the distribution of the response variable in our dataset:

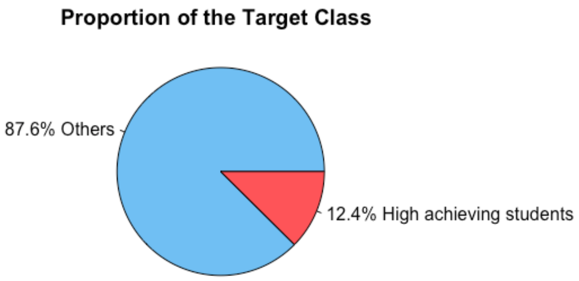


Figure 15. *Proportion of the response variable in the sample*

From the following graph the response variable's percentage are shown according to gender's attributes. It is possible to highlight that even if the share of male students is significantly higher than female ones, the proportion of the high achieving students is higher for female students.

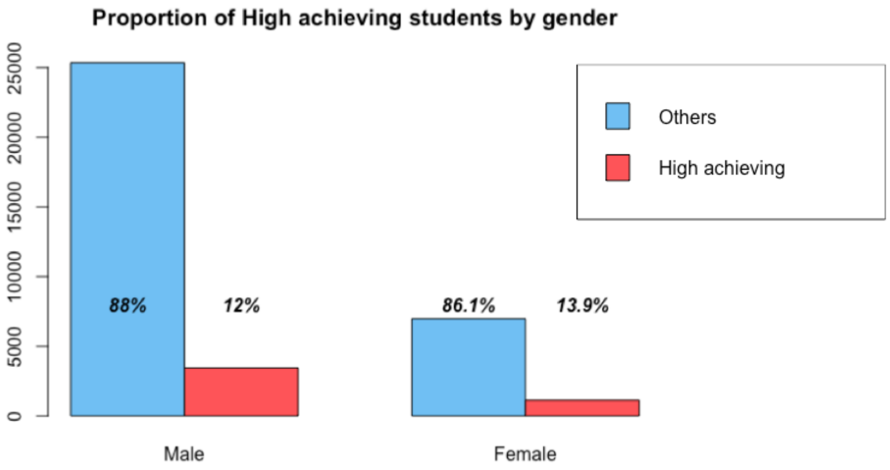


Figure 16. *Proportion of the target class according to gender*

The same has been done according to the Residence of students. The students coming from out of the Lombardy region (off-site) constitute the predominant class, while the category with the higher percentage of high achieving students is that of who come from the countryside.

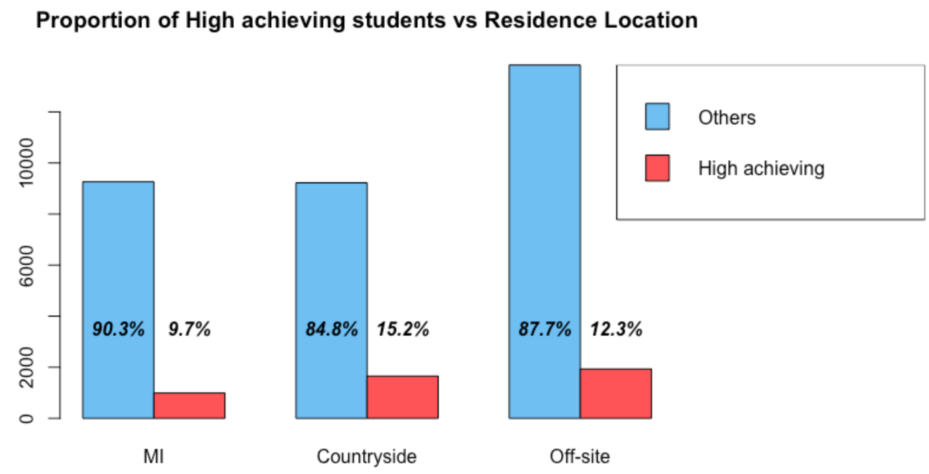


Figure 17. *Proportion of the target class by Residence Location*

In the graph below, only the class of high achieving student is reported by type of high school. The scientific high school is that with the highest percentage of high achieving students within its category, followed by classic, technical and other high school.

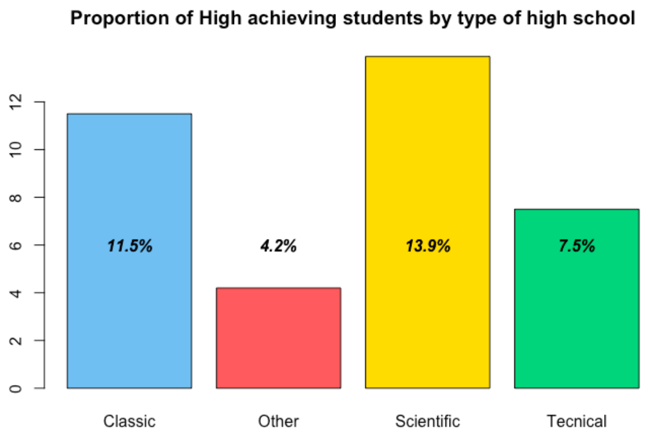


Figure 18. *Percentage of high achieving students by previous high school studies*

Then, the distribution of numerical attributes is shown split by the response class **Y**. The first explored variables have been the Total Credits at the end of the first semester and of the first year.

The below boxplots show two significant patterns according to the response variable considered. In fact, regarding the category of High achieving students we can see a tighter box that represents lower variability in terms of ECTS gained and it becomes even smaller at the end of the first year. While the other students are characterized by a greater variability in both scenarios.

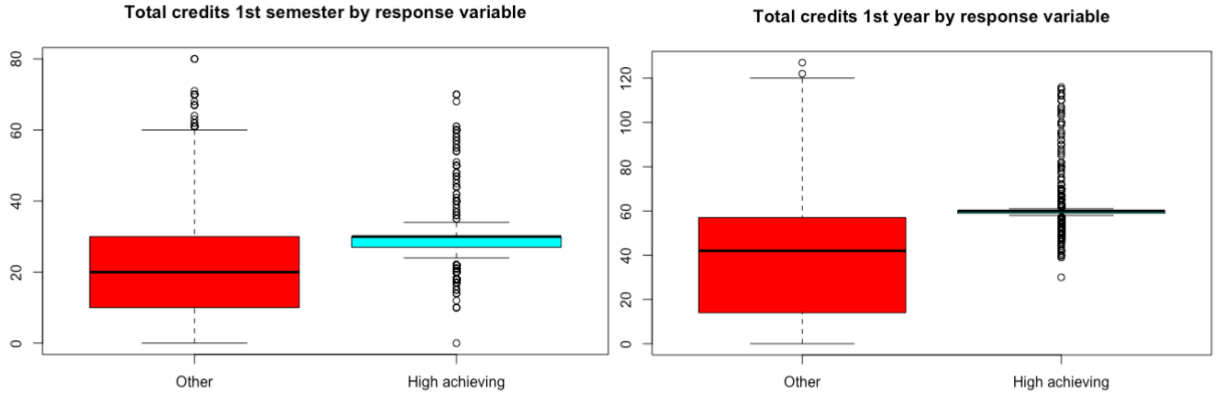


Figure 19. *Comparison of boxplots presenting the total credits gained at the end of the 1st semester (left) and at the end of the 1st year according to the response variable Y*

Also, the frequency distribution of the attribute Average Attempt per exam seems to be significant for the response variable **Y**. As can be seen from the figure below the class of high achieving students has a steady trend with a slight increase of the median at the end of the first year. For other students instead, we can see a larger variability for both the periods. It is possible to see how at the end of the first semester the class of other students have a minimum of 0 attempt per exam and a maximum of 3.5. Regarding high achieving students the range of attempts is between 1 and 2 with the median of 1.

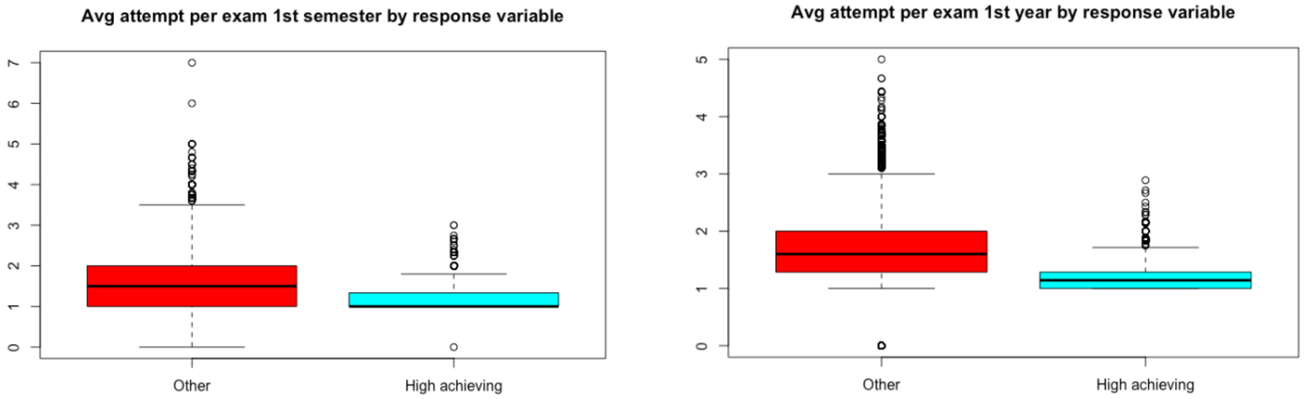


Figure 20. *Comparison of boxplots presenting the average number of attempts per exam at the end of the 1st semester (left) and at the end of the 1st year according to the response variable Y*

Also, the diploma score attributes has been split according to the binary variable Y. The result is represented in the following graph:

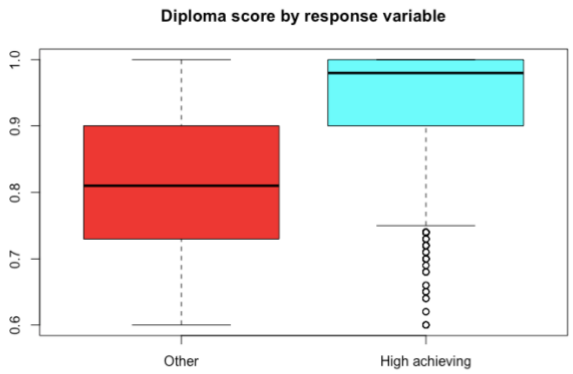


Figure 21. *Boxplots presenting the variation of the diploma score according to the response variable Y*

As expected, the class of high achieving students is characterized by higher score at high school with a median value around 0.97-0.98.

A similar pattern is given by the frequency distribution of the attributes Admission score according to the response variable Y:

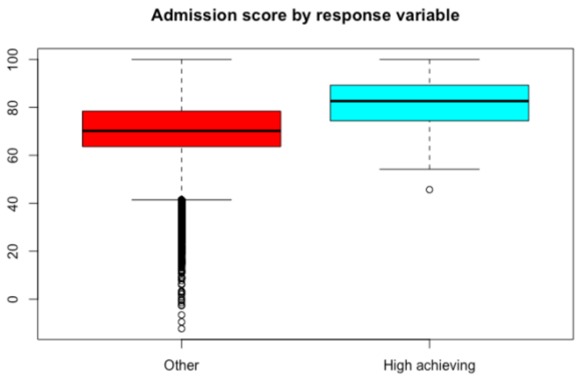


Figure 22. *Boxplots presenting the variation of the admission test score according to the response variable Y*

The difference between the two class [Other; High achieving students] is significant and indicates that the variables listed above are potentially relevant in explaining the target value.

Ultimately, I measured the collinearity between features by using correlation measure. As it can be seen from the following pictures the attributes related to the academic performance are high correlated:

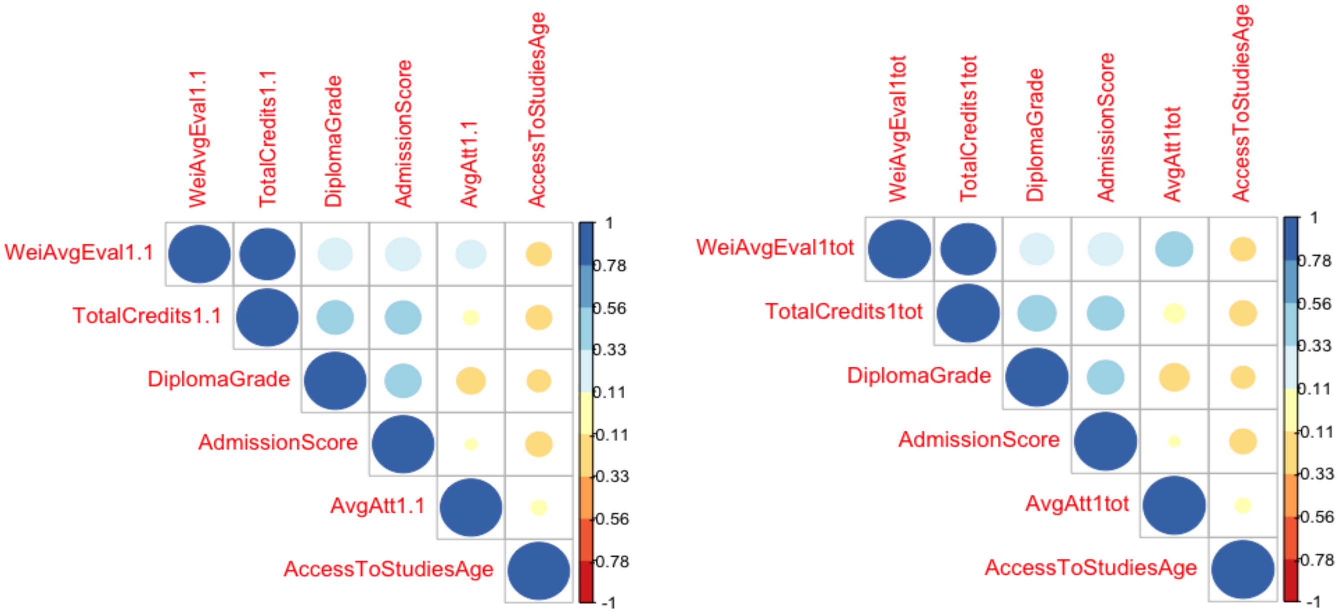


Figure 23. *Collinearity matrices for numerical variables. On the left the academic data referring to the 1st semester, on the right the academic data referring to the 1st year*

These matrices show high value of collinearity between the WeiAvgEval and TotalCredits variables. In order to avoid dependency between these two variables, I removed the WeiAvgEval1.1 for the first semester and WeiAvgEval1tot for the first year when using the linear models (GLM, GLMER).

Chapter 4 - Implemented models

4.1 Classification models

Classification models are supervised learning methods for predicting the value of a categorical target attribute, unlike regression models which deal with numerical attributes. Starting from a set of past observations whose target class is known, classification models are used to generate a set of rules that allow the target class of future examples to be predicted. A classification problem consists of defining an appropriate hypothesis space ***F*** and an algorithm ***AF*** that identifies a function $f* \in$ ***F*** that can optimally describe the relationship between the predictive attributes and the target class.

In the following chapters the classification models employed in the present work will be presented more in detail.

4.1.1 Logistic regression

Logistic regression is a technique for converting binary classification problems into linear regression ones. Given a response variable Y that takes the values [0,1] logistic regression models the probability that Y belongs to a particular category. The logistic regression model postulates that the posterior probability P(y|**x**) of the response variable conditioned on the vector **X** follows a logistic function that gives outputs between 0 and 1 for all values of **X** and is given by:

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

The standard logistic function *S(t),* also known as the *sigmoid function*, can be found in many applications of statistics and is defined as:

$$S(t) = \frac{1}{1+e^{-t}}$$

The function S(t) has the graphical shape of that shown in the below figure (In the example the variable *balance* is the response variables):

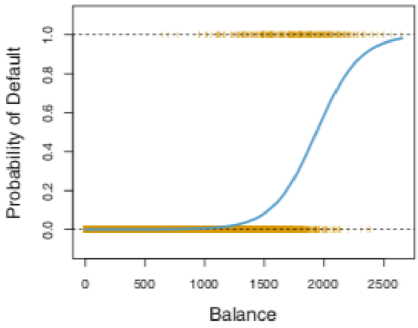


Figure 24. *Example of a sigmoid function S(t)*

After a bit of manipulation of the logit function, we find that:

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

The left-hand side is called the ***log-odds*** or ***logit***. In a logistic regression model, increasing X by one unit changes the log odds by β1, or equivalently it multiplies the odds by $e^{\beta 1}$. However, because the relationship between p(X) and X in the logit function is not a straight line, β1 does not correspond to the change in p(X) associated with a one-unit increase in X. The amount that p(X) changes due to a one-unit change in X will depend on the current value of X. But regardless of the value of X, if β1 is positive then increasing X will be associated with increasing p(X), and if β1 is negative then increasing X will be associated with decreasing p(X). The fact that there is not a straight-line relationship between p(X) and X, and the fact that the rate of change in p(X) per unit change in X depends on the current value of X, can also be seen by inspection of the sigmoid function in the graph above.

4.1.2 Classification trees

Classification trees are perhaps the best-known and most widely used learning methods in data mining applications. The reasons for their popularity lie in their conceptual simplicity, ease of usage, computational speed, robustness with respect to missing data and outliers and, most of all, the interpretability of the rules they generate.

For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.

The development of a classification tree is regulated by a recursive procedure of heuristic nature, based on a divide-and-conquer partitioning scheme referred to as top-down induction of decision trees. The following picture shows an example of a binary (if each node has at most two branches) classification tree:

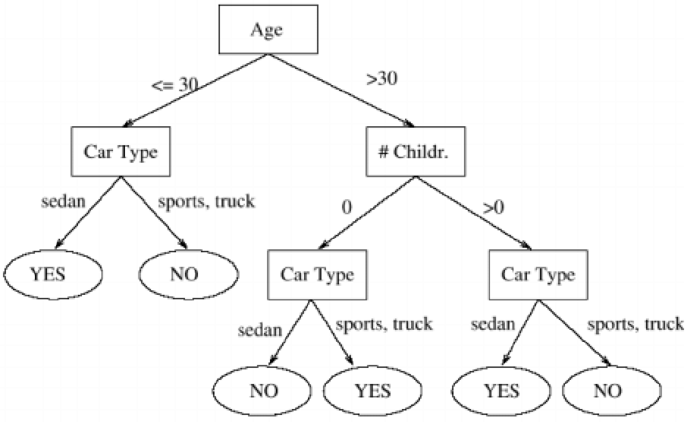


Figure 25. *Example of the structure of a classification tree*

In the splitting phase numeric variables are divided into X < a and X > a instead the levels of an unordered factor are divided into two non-empty groups. There are three main criteria used for making binary splits:

- **Classification error rate**. It is the fraction of the misclassified predicted observations over the total number of predicted observations. In the following formula $\hat{p}_{mk}$ represents the proportion of training observations in the *m*th region that are from the *k*th class. However, it turns out that classification error is not sufficiently sensitive for tree-growing, and in practice the other two measures are preferable.

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

- The **Gini index** is defined by:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

It is a measure of total variance across the K classes. It is not hard to see that the Gini index takes on a small value if all of the $\hat{p}_{mk}$'s are close to zero or one. For this reason, the Gini index is referred to as a measure of node *purity*—a small value indicates that a node contains predominantly observations from a single class.

- **Entropy** is given by:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

Since 0 ≤ $\hat{p}_{mk}$ ≤ 1, it follows that 0 ≤ – $\hat{p}_{mk}$ log$\hat{p}_{mk}$. One can show that the entropy will take on a value near zero if the $\hat{p}_{mk}$'s are all near zero or near one. Therefore, like the Gini index, the entropy will take on a small value if the *m*th node is pure. In fact, it turns out that the Gini index and the entropy are quite similar numerically.

When building a classification tree, either the Gini index or the entropy are typically used to evaluate the quality of a particular split, since these two approaches are more sensitive to node purity than is the classification error rate. Any of these three approaches might be used when pruning the tree, the classification error rate is preferable if prediction accuracy of the final pruned tree is the goal.
Splitting continues until the terminal nodes are too small or too few to be split.

4.1.3 Random Forests

Random forests, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. With random forests we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split is allowed to use only one of those *m* predictors. A fresh sample of *m* predictors is taken at each split, and typically we choose m ≈ $\sqrt{p}$ —that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors.
Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below):

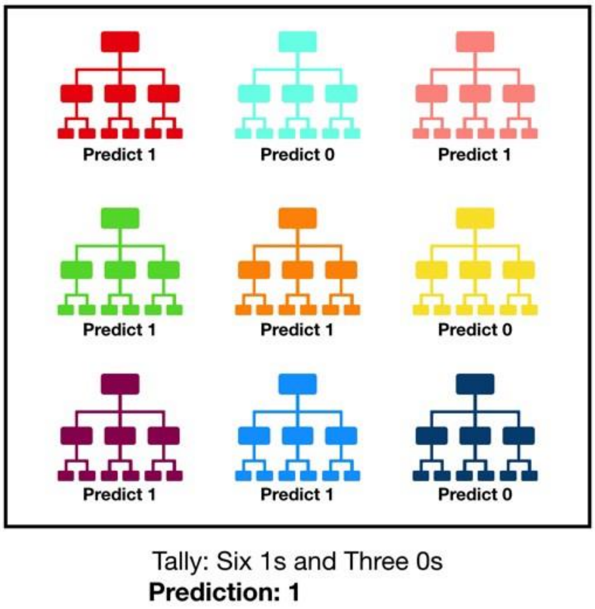


Figure 26. *The approach to make predictions adopted by the Random Forest model. Source: Tony Yiu – Understanding Random Forest*

The prerequisites for random forest to perform well are:

- There needs to be some actual signal in our features so that models built using those features do better than random guessing;
- The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

4.2 Multilevel classification models

Generally, when handling academic data, we have to consider that students are naturally nested within the degree program they are attending. While investigating the learning process, it is necessary to separate the effects given by each level of hierarchy (*L.Fontana et al.*, 2018). Indeed, if the clustered aspect of the data is not inspected, it may result in a loss of likely valuable information. Multilevel models take into account the hierarchical nature of data and are able to quantify the portion of variability in the response variable that is attributable to each level of grouping. In addition, these methods allow a clear graphical representation of the results that is easy to communicate.

4.2.1 Logistic generalized linear mixed-effects model (GLMM)

A Logistic Generalized linear mixed model (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Alternatively, you could think of GLMMs as an extension of generalized linear models (e.g., logistic regression) to include both fixed and random effects (hence mixed models).

The conventional multilevel logistic regression model incorporates cluster-specific random effects to account for the within-cluster correlation of subject outcomes:

$$\operatorname{logit}\big(\Pr(Y_{ij}=1)\big) = \alpha_0 + \alpha_{0j} + \alpha_1 x_{1ij} + \cdots + \alpha_k x_{kij} + \beta_1 z_{1j} + \cdots + \beta_m z_{mj}$$

where α_{0j}~$N(0,\tau^2)$.

In the formula above, Y_{ij} denote the binary response variable measured on the *i*-th subject within the *j*-th cluster (Y*ij* = 1 denotes success or the occurrence of the event, while Y*ij* = 0 denotes failure or lack of occurrence of the event). Furthermore, X_{1ij}, through X_{kij} denote the *k* predictor or explanatory variables measured on this subject (e.g., age of enrollment). Finally, Z_{1j}, through Z_{mj} denote the *m* predictor variables measured on the *j*-th cluster (e.g., degree course nature).

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 (degree course nature). In this case the log odd (chapter 4.1.1) becomes:

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

Where:

- M is the 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.

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 VPC (variance partition coefficient), also called Intraclass correlation (ICC), defined as:

$$\text{VPC} = \frac{\partial_m^2}{\partial_m^2 + \pi^2/3}$$

where ∂_m^2 is the estimated variance of random effects: the higher it is, the larger the variation of the average log-odds between clusters; while π2/3 represents the residual variability that can neither be explained by fixed effects, nor through the group features that are represented by the random intercept, since the variance of the standard logistic distribution is π2/3 ≃ 3.29 ($\pi^2/3$) = 3.29.

The VPC measured the homogeneity of the outcome within clusters, it represents the proportion of the between-cluster variation.

4.2.1 Generalized mixed-effects tree (GMET)

As in the multilevel logistic regression, clustered data could be analyzed through mixed models. The mixed-effects regression 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 in the present work was the one designed by Fontana et al. (2018). 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 the response variable and extract the random effects from the estimation 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.

The GLM in step 2 is fitted through the maximum likelihood, it determines values for the parameters found such that it maximizes 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.

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

4.2.2 Generalized mixed-effects forest (GMEF)

The Generalized mixed-effects forest (GMEF) as the name suggests, combines a mixed effect generalized linear model with a random forest. It is an extension of the GMET algorithm, with whom shares the same formulation, but it tries to improve it (as far as prediction accuracy is concerned), because it estimates the fixed effects by means of a random forest instead of a single tree. Again, the response Y given the random effects bi is assumed to have a distribution which is part of the exponential family with link function g.

Regression tree model produces, other than a prediction for the response, also terminal nodes, which can be used in GMET procedure as fixed effect covariates in the mixed effects model. With random forests there is no more a tree structure, since the final prediction is an average of lots of single trees, so all we can extract from the model is the predicted values.

Since neither the random effects nor the fixed effects are known, we alternate between estimating the random forest, assuming that our estimates of the random effects are correct, and estimating the random effects, assuming that the forest estimate is correct. This alternation continues until we have convergence, which means that estimates of b_i change a little from one iteration to another, id est the difference from an iteration to another one is smaller than a fixed threshold.

To predict the following formula is used:

$$\hat{\eta}_{ij} = \hat{f}(x) + z^T \hat{b}_i,$$

where:

- $\hat{f}$ is the random forest output of the algorithm;
- $\hat{b}_i$ is the i-th column of the b output of the algorithm;
- The prediction $\hat{\mu}_{ij}$ is obtained by applying to the corresponding $\hat{\eta}_{ij}$ the inverse link function g−1.

Chapter 5 – Findings

Chapter 5 is about the presentation of the results of data analysis which final aim is to answer to the initial research questions and try to "fill the literature gap" that has been left open for Learning Analytics:

- Is it possible to predict high achieving students since the earliest stage of their academic career?
- Which are the relevant attributes to predict a talented student?
- Which are the best models to predict high achieving students?

The present chapter of results will follow the layout given by the previous research questions.

To start, two main model have been considered: the one related to the 1-semester academic attributes and the one with the upgraded information at the end of the first year. Obviously, we expect that the performances of the 1-year model will be greatest since it provides more information closest to the time of graduation, but one of the objectives of this work is to is to understand if it is possible to predict high achieving students since the beginning of their career and to achieve this, we compared these two models in terms of sensitivity and accuracy.

Simultaneously to the previous goal, the research of relevant attributes to predict a high achieving student has been conducted. Several machine learning models habe been employed during the analysis, that can be divided in two main group:

- Single level-classification models: logistic regression, classification trees and random forest;
- Multilevel classification models: GLMER, GMET and GMEF.

The idea of introducing also multilevel classification models comes from the nature of educational data itself. In fact, this kind of data contains a well-defined hierarchy of levels that could compromise the independence of observations which belongs to the same cluster, represented by the same Course of study in our case.

Once the best performing algorithms have been selected, we tried to improve their performance in terms of sensitivity and accuracy of predictions, with the application of three further approaches:

- Balancing the dataset;
- Ensemble methods: Stacking;

- Optimal threshold in order to minimize False Negatives (FN).

Our response variable **Y** is a two-level factor that we coded as a binary variable:

- Y = 1 for high achieving students with graduation score > 101 and graduation within 3 years after the enrolment (YearsToFinishDegree <= 3years);
- Y = 0 for all other students that has a graduation score < 101 and graduated in more than 3 years.

The choice of this coding is given by the fact that we are interested in detecting high achieving students who are likely to graduate with high scores and without delay.

In this work the cohorts from 2010 to 2015 are selected to train the models and the next cohort (A.Y 2016/2017) to test them, as opposed to most of the works reported in the literature which use cross-validation, which means that the same cohorts are used to train and test the classifier. The aim of using successive cohorts is to check how well results generalize over time so as to use the experience of one cohort to put in place some policy to detect weak or strong students for the following cohort (*Meceron*, 2015).

5.1 Research question 1: Is it possible to predict high achieving students since the earliest stage of their academic career?

The aim of this section is to understand if it is possible to predict high achieving students since the earliest stage of their career. In order to do so we employed several machine learning algorithms considering two distinct models:

- *1st - semester data*. It includes only academic performance variables of the first semester: **WeiAvgEval1.1**, **TotalCredits1.1** and **AvgAtt1.1**;
- *1st – year data*. It includes upgraded academic performance variable at the end of the first year (2 consecutive semesters): **WeiAvgEval1tot**, **TotalCredits1tot** and **AvgAtt1tot**.

The results are evaluated in terms of Accuracy, Sensitivity and Specificity. These performance metrics are based on misclassification tables. Given the true labels $y_1,..,y_N$ of the response Y of a test set and the corresponding predictions (which we assume coded as 0 and 1), a misclassification table is a table of this type:

Table 7. *General layout for a misclassification table*

Y	**Actual Y = 1**	**Actual Y = 0**	
Predicted Y = 1	TP	FP	Predicted condition Positive
Predicted Y = 0	FN	TN	Predicted condition Negative
Total	P	N	

Where:

- True positive (TP) is the number of correct predictions for the positive condition (Y=1);
- True Negative (TN) is the number of correct predictions for the negative condition (Y=0);
- False Positive (FP) is the number of misclassified predictions for the negative condition;
- False Negative (FN) is the number of misclassified predictions for the positive condition;

Then, the above-mentioned metrics are calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity}= \frac{TN}{TN+FP}$$

Accuracy represents the percentage of correctly classified units and it is the main measure to evaluate the overall quality. However often is it more important to correctly detect True Positives than True Negatives, for example in medicine: an ill patient is a Positive and misclassifying him as healthy is far worse than considering ill a healthy patient. For this reason, the other indexes were introduced. Sensitivity represents the percentage of positive which are correctly classified, while Specificity represents the percentage of negative predicted on the whole class of negative.

Also in this project detecting the true positive is more important than detecting true negatives one, in order to be sure that the higher number of high achieving students will be addressed with the proper resources. Then predictions' sensitivity will be considered as well as the accuracy measure for evaluating the different algorithm, in order to understand the portion of true positive correctly classified.

The results are presented below considering the metrics specified above and the graphical representations.

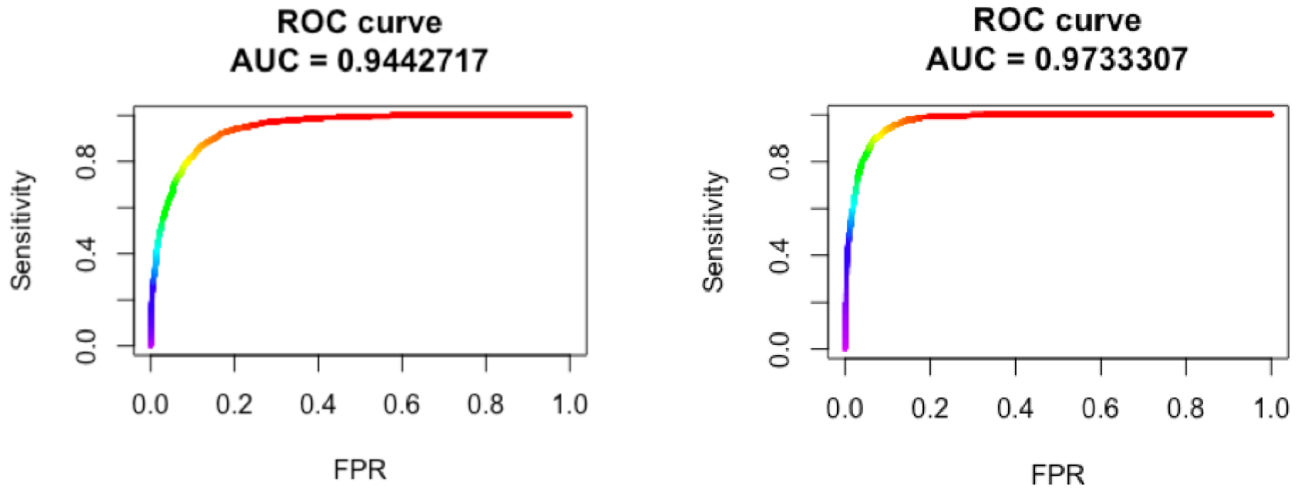


Figure 27. *Comparison of the ROC curves resulting from the estimation of the Logistic Regression model to the validation dataset. On the left the 1st semester mode (p_0 = 0.18) and on the right the 1st year model (p_0 = 0.22).*

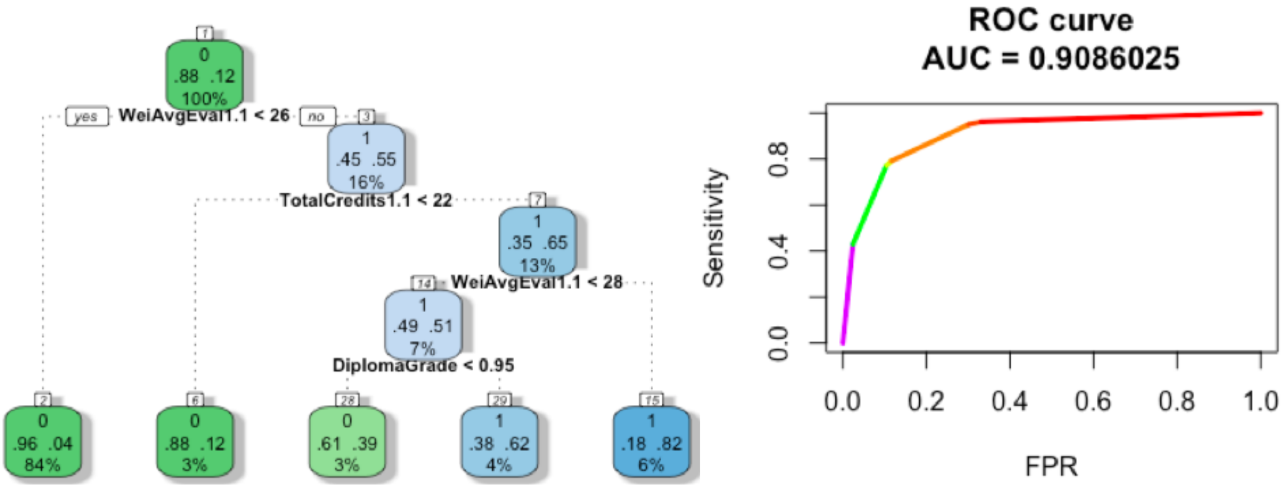


***Figure 28.** 1st SEMESTER MODEL: On the left the graphical representation of the decision classification tree applied to the training set. On the right the ROC resulting from the estimations on the test data*.

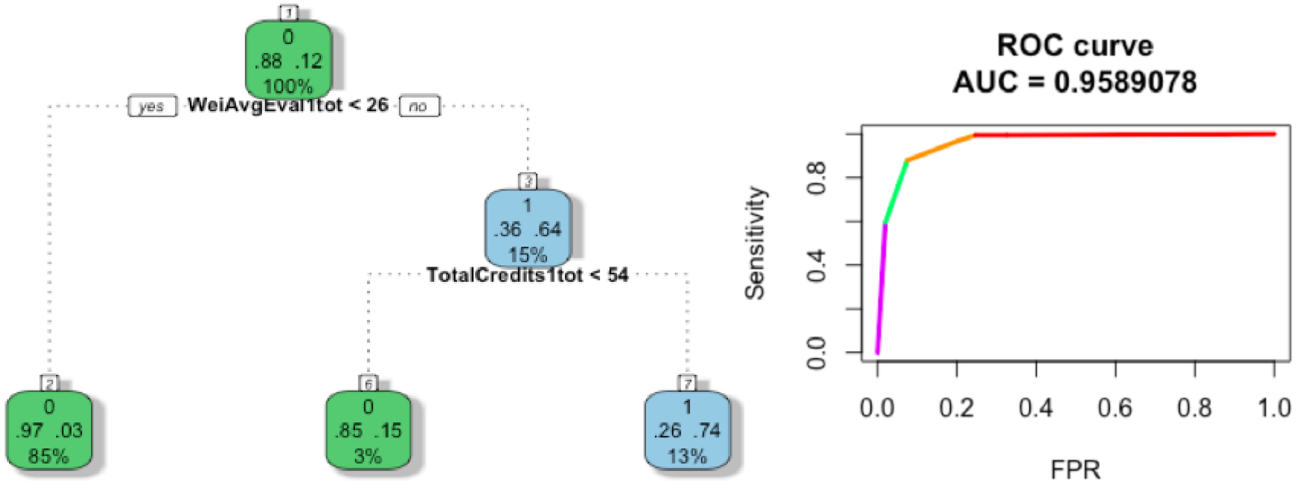


***Figure 29.** 1st YEAR MODEL: On the left the graphical representation of the decision classification tree applied to the training set. On the right the ROC resulting from the estimations on the test data.*

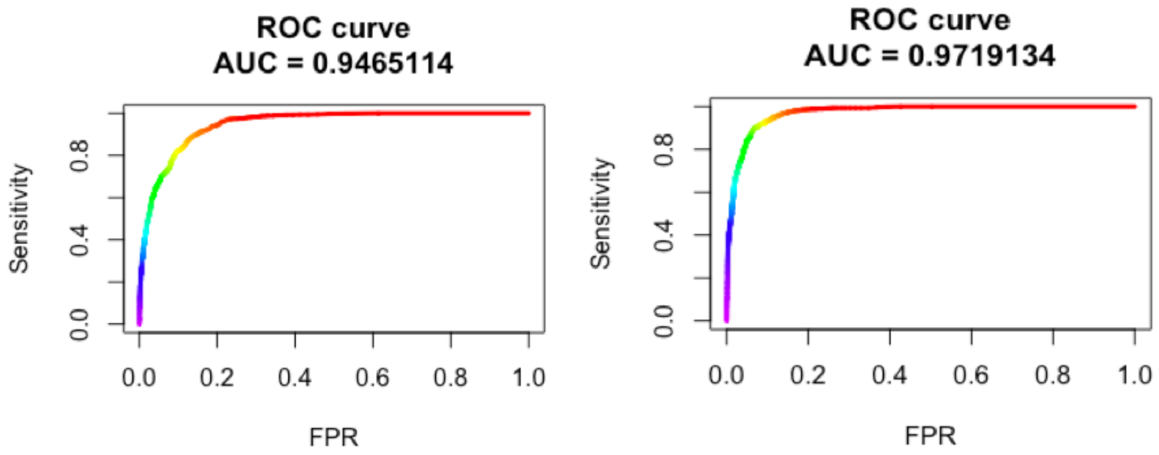


Figure 30*. Comparison of the ROC curves resulting from the estimation of the Random Forest model to the validation dataset. On the left the 1st semester model and on the right the 1st year model.*

Table 8. *Summary of the performance measures of the 1st semester model*

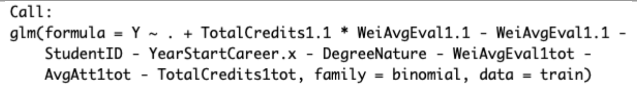
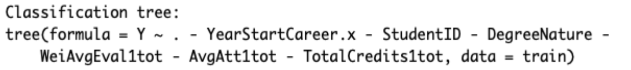
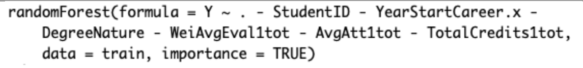
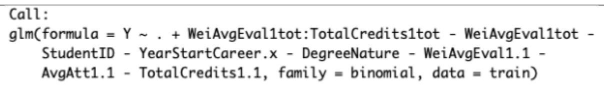
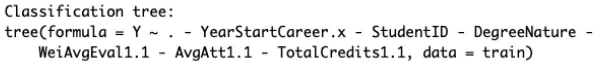
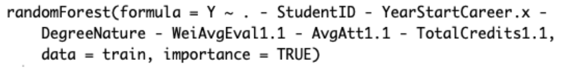
1st semester					
	Formula	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
GLM	 Call: glm(formula = Y ~ . + TotalCredits1.1 * WeiAvgEval1.1 - WeiAvgEval1.1 - StudentID - YearStartCareer.x - DegreeNature - WeiAvgEval1tot - AvgAtt1tot - TotalCredits1tot, family = binomial, data = train)	0.8883	0.8229	0.8984	0.944
Classification Trees	 Classification tree: tree(formula = Y ~ . - YearStartCareer.x - StudentID - DegreeNature - WeiAvgEval1tot - AvgAtt1tot - TotalCredits1tot, data = train)	0.9023	0.4297	0.9759	0.909
Random Forest	 randomForest(formula = Y ~ . - StudentID - YearStartCareer.x - DegreeNature - WeiAvgEval1tot - AvgAtt1tot - TotalCredits1tot, data = train, importance = TRUE)	0.9153	0.5586	0.9708	0.947

Table 9. *Summary of the performance measures of the 1st year model*

1st - year					
	Formula	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
GLM	 Call: glm(formula = Y ~ . + WeiAvgEval1tot:TotalCredits1tot - WeiAvgEval1tot - StudentID - YearStartCareer.x - DegreeNature - WeiAvgEval1.1 - AvgAtt1.1 - TotalCredits1.1, family = binomial, data = train)	0.9202	0.9076	0.9222	0.973
Classification TREE	 Classification tree: tree(formula = Y ~ . - YearStartCareer.x - StudentID - DegreeNature - WeiAvgEval1.1 - AvgAtt1.1 - TotalCredits1.1, data = train)	0.9824	0.5951	0.9803	0.959
Random Forest	 randomForest(formula = Y ~ . - StudentID - YearStartCareer.x - DegreeNature - WeiAvgEval1.1 - AvgAtt1.1 - TotalCredits1.1, data = train, importance = TRUE)	0.937	0.7357	0.9694	0.972

As expected, the 1st year model generates better results with all the employed algorithms in terms of accuracy, sensitivity and specificity. But what I want to highlight from these results, is that the difference in these performance metrics is not so huge compared to the benefit for predicting high achieving students since the first semester. In fact, if we will wait to provide relevant resources for these talented students, they will lose 6 months of opportunities. Moreover, the algorithms applied to the 1st semester model generates already very good results in terms of accuracy and sensitivity.

Especially, as regards to Accuracy and Sensitivity measures, Random Forest and Logistic Regression have better performances. Hereinafter just considering the 1st semester model, it can be seen that Random forest shows the highest value of Accuracy (0.9153) while the Logistic Regression has the highest value in terms of Sensitivity (0.8229).

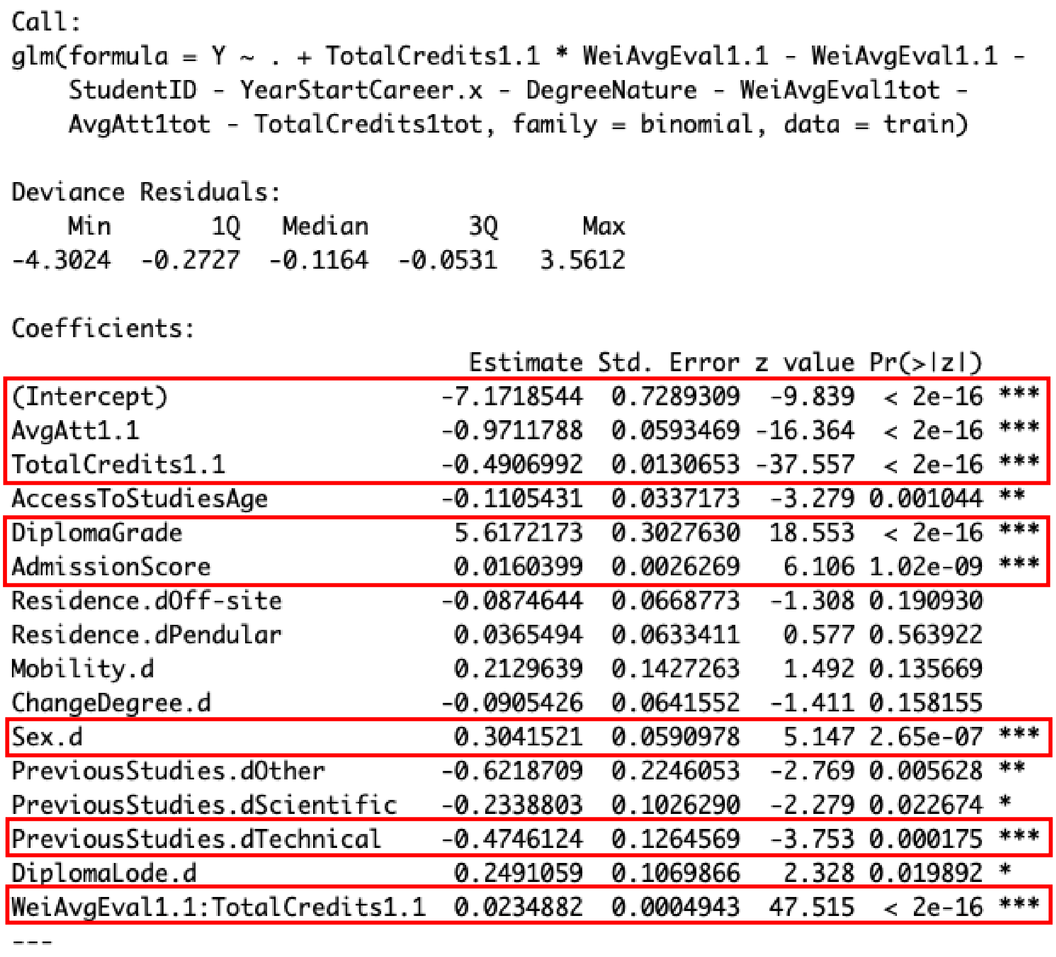
5.2 Research question 2: Which are the relevant attributes to predict a talented student?

The purpose of this paragraph is to show a more in depth understanding of which are the most impacting variable for predicting high achieving students. To do this, we restricted the focus to Logistic Regression and Random Forest algorithms applied to the 1st semester model, since they had the best performances.

Moreover, multilevel classification models have been applied in order to understand if the Degree Course, to which the student belong, can explain a significant portion of variance in our predictions.

The results which were found relevant are set out below, starting from single level classification algorithms and then analyzing the behavior of the multilevel ones.

The following picture summarizes the results obtained from the Logistic Regression model:



```
Call:
glm(formula = Y ~ . + TotalCredits1.1 * WeiAvgEval1.1 - WeiAvgEval1.1 -
    StudentID - YearStartCareer.x - DegreeNature - WeiAvgEval1tot -
    AvgAtt1tot - TotalCredits1tot, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3024  -0.2727  -0.1164  -0.0531   3.5612

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -7.1718544  0.7289309  -9.839  < 2e-16 ***
AvgAtt1.1                   -0.9711788  0.0593469 -16.364  < 2e-16 ***
TotalCredits1.1             -0.4906992  0.0130653 -37.557  < 2e-16 ***
AccessToStudiesAge          -0.1105431  0.0337173  -3.279 0.001044 **
DiplomaGrade                 5.6172173  0.3027630  18.553  < 2e-16 ***
AdmissionScore               0.0160399  0.0026269   6.106 1.02e-09 ***
Residence.dOff-site         -0.0874644  0.0668773  -1.308 0.190930
Residence.dPendular          0.0365494  0.0633411   0.577 0.563922
Mobility.d                   0.2129639  0.1427263   1.492 0.135669
ChangeDegree.d              -0.0905426  0.0641552  -1.411 0.158155
Sex.d                        0.3041521  0.0590978   5.147 2.65e-07 ***
PreviousStudies.dOther      -0.6218709  0.2246053  -2.769 0.005628 **
PreviousStudies.dScientific -0.2338803  0.1026290  -2.279 0.022674 *
PreviousStudies.dTechnical  -0.4746124  0.1264569  -3.753 0.000175 ***
DiplomaLode.d                0.2491059  0.1069866   2.328 0.019892 *
WeiAvgEval1.1:TotalCredits1.1  0.0234882  0.0004943  47.515  < 2e-16 ***
---
```

Figure 31. *Summary of the Logistic Regression model' results*

The p- values associated with the variables indicated in the red boxes are very small, indicating that each of these variables is associated with the probability to be a high achieving student. In particular, we can see that the high school score (DiplomaGrade), the admission score and being female (Sex.d) have a positive impact on the likeliness to be a top performer student. While the average attempt per exam, the number of CFU obtained and having attended a technical high school are negative correlated with the response variable.

Similar results are given by the Random Forest model. In the following graph the predictor variables are plotted according to their importance in predicting the response class (Y). The importance of variables is explained by the two following measures:

- **The mean decrease in accuracy**: is determined during the out of bag error calculation phase. The more the accuracy of the random forest decreases due to the exclusion (or permutation) of a single variable, the more important that variable is deemed, and therefore variables with a large mean decrease in accuracy are more important for classification of the data.
- **The mean decrease in Gini coefficient**: is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. Each time a particular variable is used to split a node, the Gini coefficient for the child nodes are calculated and compared to that of the original node. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). The changes in Gini are summed for each variable and normalized at the end of the calculation. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient.

As expected, the weighted average evaluation (WeiAvgEval1.1) significantly prevails compared to the other predictors. Then as follows the number of CFU obtained at the end of the 1st semester (TotalCredits1.1), the average attempt per exam (AvgAtt1.1), DiplomaGrade and AdmissionScore.

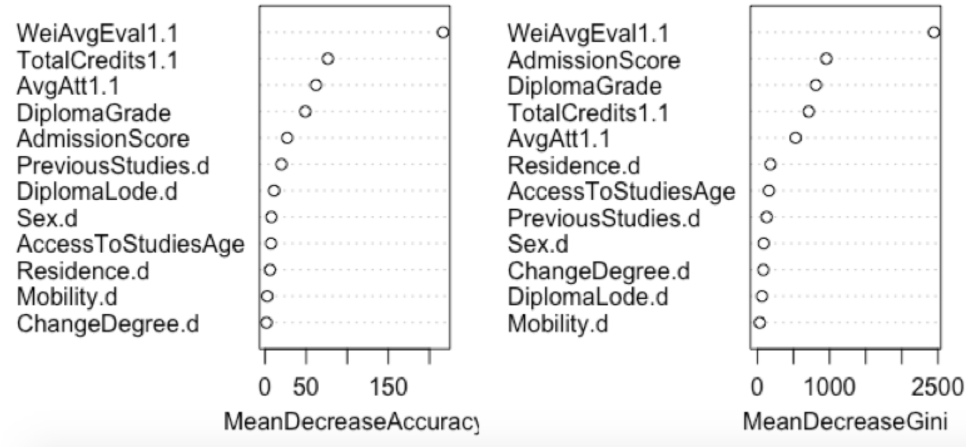
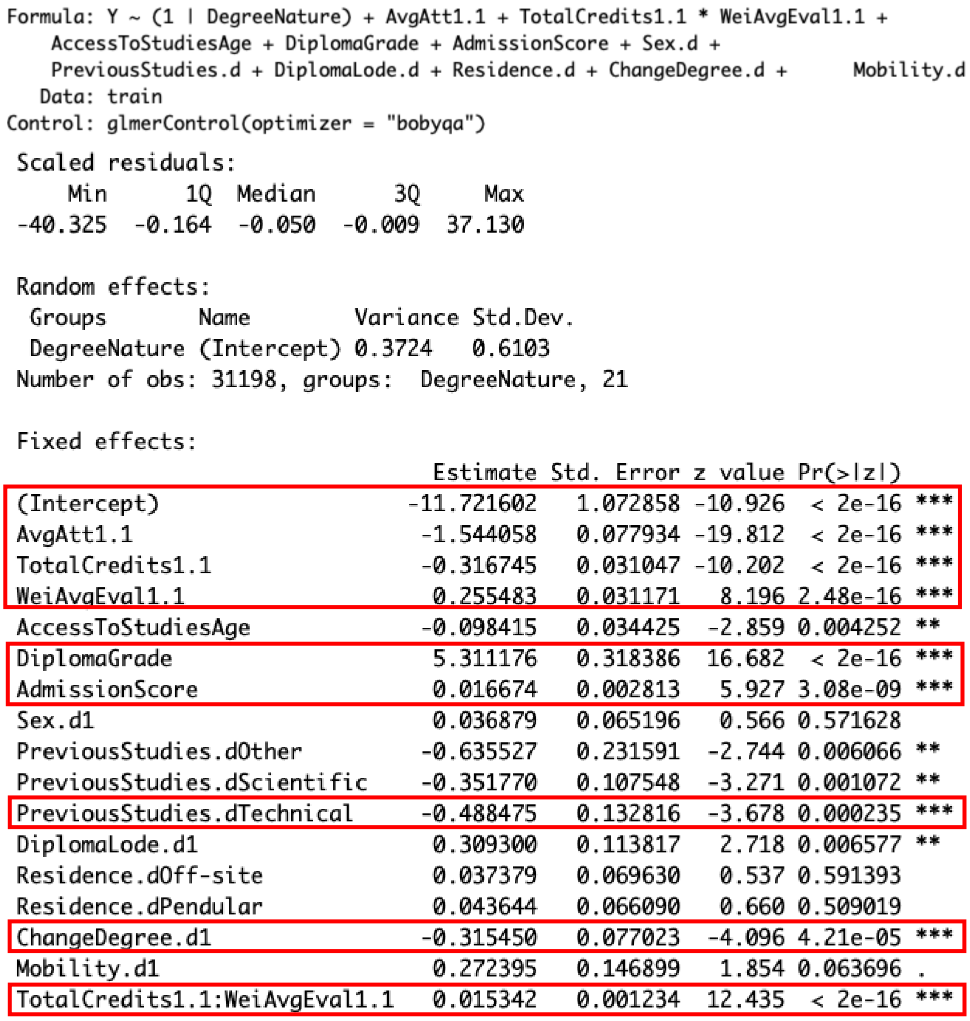


Figure 32. *Plot of the variable importance resulting from the fitting of Random Forest model by Mean Decrease Accuracy and Mean Decrease Gini*

5.2.1 Multilevel classification models

The first model employed for multilevel analysis is the *glmer* from the *lme4* package for R.



```
Formula: Y ~ (1 | DegreeNature) + AvgAtt1.1 + TotalCredits1.1 * WeiAvgEval1.1 +
    AccessToStudiesAge + DiplomaGrade + AdmissionScore + Sex.d +
    PreviousStudies.d + DiplomaLode.d + Residence.d + ChangeDegree.d +      Mobility.d
   Data: train
Control: glmerControl(optimizer = "bobyqa")

Scaled residuals:
    Min      1Q  Median      3Q     Max
-40.325  -0.164  -0.050  -0.009  37.130

Random effects:
 Groups       Name        Variance Std.Dev.
 DegreeNature (Intercept) 0.3724   0.6103
Number of obs: 31198, groups:  DegreeNature, 21

Fixed effects:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -11.721602   1.072858 -10.926  < 2e-16 ***
AvgAtt1.1                    -1.544058   0.077934 -19.812  < 2e-16 ***
TotalCredits1.1              -0.316745   0.031047 -10.202  < 2e-16 ***
WeiAvgEval1.1                 0.255483   0.031171   8.196 2.48e-16 ***
AccessToStudiesAge           -0.098415   0.034425  -2.859 0.004252 **
DiplomaGrade                  5.311176   0.318386  16.682  < 2e-16 ***
AdmissionScore                0.016674   0.002813   5.927 3.08e-09 ***
Sex.d1                        0.036879   0.065196   0.566 0.571628
PreviousStudies.dOther       -0.635527   0.231591  -2.744 0.006066 **
PreviousStudies.dScientific  -0.351770   0.107548  -3.271 0.001072 **
PreviousStudies.dTechnical   -0.488475   0.132816  -3.678 0.000235 ***
DiplomaLode.d1                0.309300   0.113817   2.718 0.006577 **
Residence.dOff-site           0.037379   0.069630   0.537 0.591393
Residence.dPendular           0.043644   0.066090   0.660 0.509019
ChangeDegree.d1              -0.315450   0.077023  -4.096 4.21e-05 ***
Mobility.d1                   0.272395   0.146899   1.854 0.063696 .
TotalCredits1.1:WeiAvgEval1.1 0.015342   0.001234  12.435  < 2e-16 ***
```

Figure 33. *Summary of the Multilevel Logistic Regression model' results*

The results are pretty similar compare to those of the linear logistic regression model. Only the attributes related to gender lose significance in the nested model. The intercept represents the average of our data for the Y condition.

Having a look to the Variance Partition Coefficient, hereinafter VPC, allows us to assess whether or not the random effect is present in the data. It is equal to the percentage of variation that is found at the higher level of a hierarchical model over the total variance. The VPC explained by the course of study is the following:

$$\text{VPC} = \frac{\partial_m^2}{\partial_m^2 + \pi^2/3} = 0.102$$

This means that nearly the 10% of variation in the response is attributable to the nested structure. This is quite significant.

The caterpillar plot, reported below, represents the random intercepts of Degree Programs estimated by the GLMER. The blue dots represent the conditional models with error bars (gray horizontal bar). The error bars, net of few cases, are very concentrated meaning that the prediction is quite accurate. Specifically, the courses of Environmental Engineering, Physical Engineering and Biomedical Engineering shows a positive correlation with the dependent class Y. While Mechanical Engineering, Computer Science Engineering, Building and Architectural Engineering and Civil Engineering are negative correlated. The other courses near to the zero value are not so significant for our predictions.

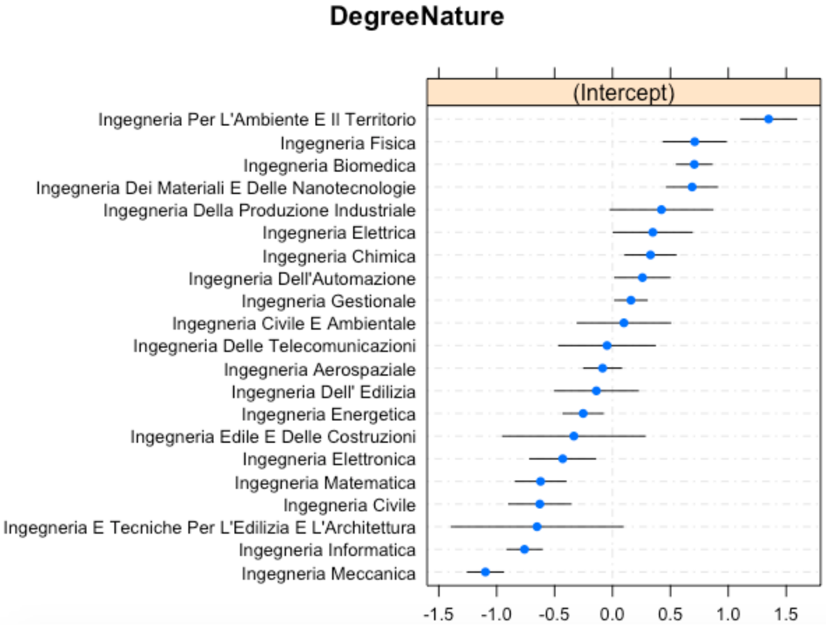


Figure 34. *Random intercepts of Degree Programs estimated by the GLMER model.*

The improvement in results can be also confirmed by the misclassification table indicated below, in which the number of FN has shrunk to only 51 observation and TP equal to 717 observations, reaching a prediction'sensitivity equal to 0.9336.

Table 10. *Misclassification table of test set predictions by GLMER*

Y	**Actual Y = 1**	**Actual Y = 0**
Predicted Y = 1	717	900
Predicted Y = 0	51	4033

Then the *GMET* algorithm has been applied and it generates similar results to the *glmer* one, but whit a slightly decrease on performances. The following picture gives a graphical understanding of the logic behind the definition of decision trees in the GMET model. It is easier to see that the first attribute considered for the classification is the weighted average evaluation, then the average attempt per exam and the diploma grade. So, if a student has a weighted average score higher than 26 and an average attempt per exam lower than 1.5, he is likely to be a top performer student. Otherwise, if the average attempt per exam is higher than 1.5 then the Diploma Grade became also a critical factor with a threshold higher than 90% of the higher score (i.e in Italy higher than 90/100).

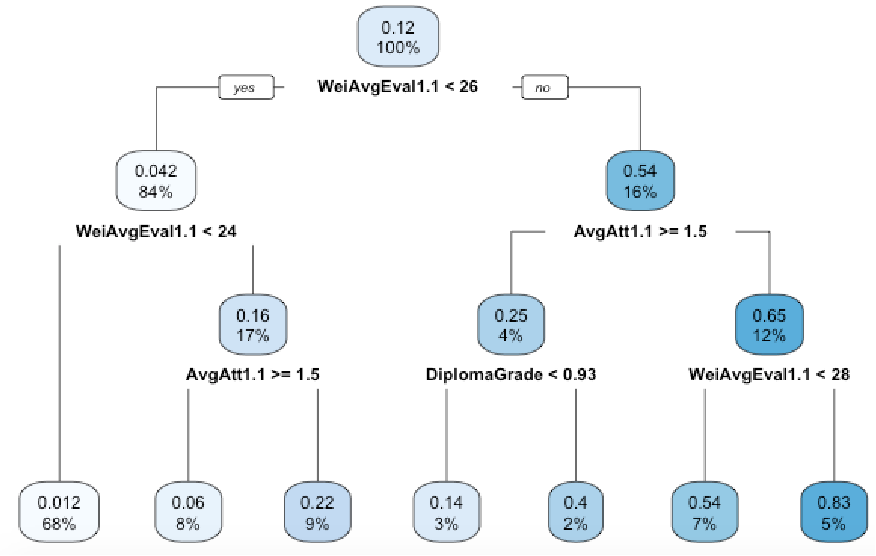


Figure 35*. Tree of the fixed effects estimated by GMET model.*

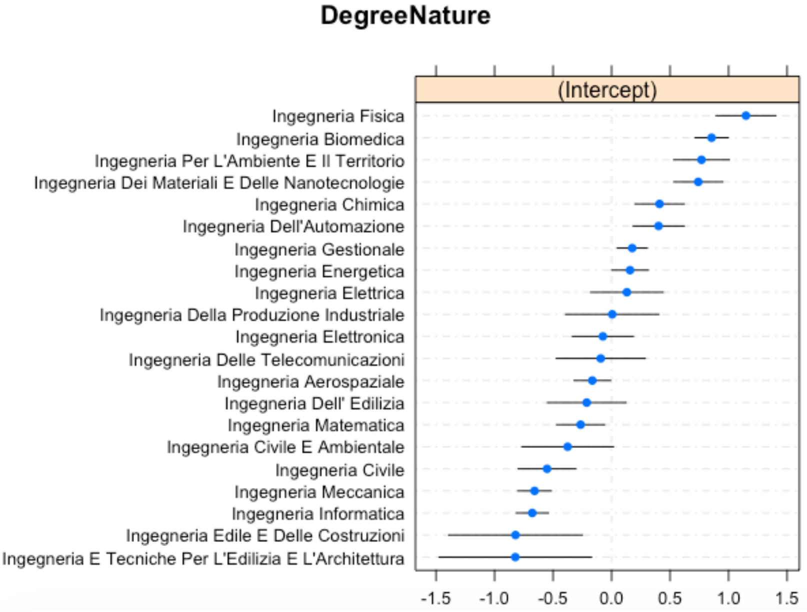


Figure 36. *Random intercepts of Degree Programs estimated by the GMET model.*

Table 11. *Misclassification table of test set predictions by GMET*

Y	Actual Y = 1	Actual Y = 0
Predicted Y = 1	655	754
Predicted Y = 0	113	4179

$$\text{VPC} = \frac{\partial_m^2}{\partial_m^2 + \pi^2/3} = 0.0920033$$

The VPC is slightly lower compared to that obtained from the *glmer* and also the misclassification table highlight the worsening of the performances.

Then the GMEF has been applied. The following results show an improvement in the classification performances of the negative class but a worst score in the prediction of high achieving students.

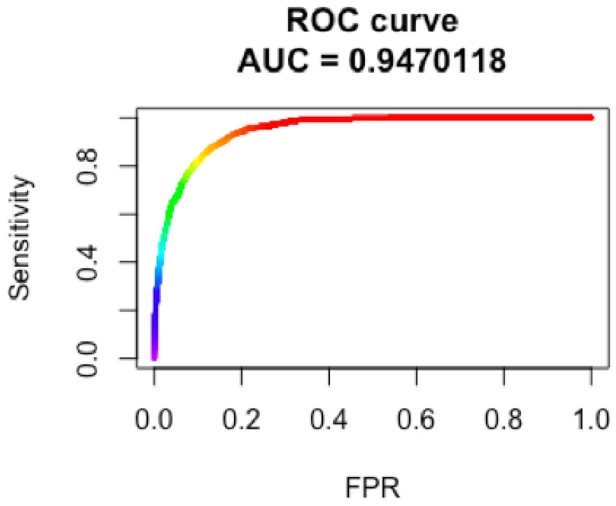


Figure 37. ROC curve resulting from the estimations on the test data by the GMEF algorithm

Similar results to the other multilevel algorithms are presented in the plot below. The only exception, compared to previous results, is given by Energetic Engineering that with the GMEF comes out as those courses of study which have negative impact on the response variable Y.

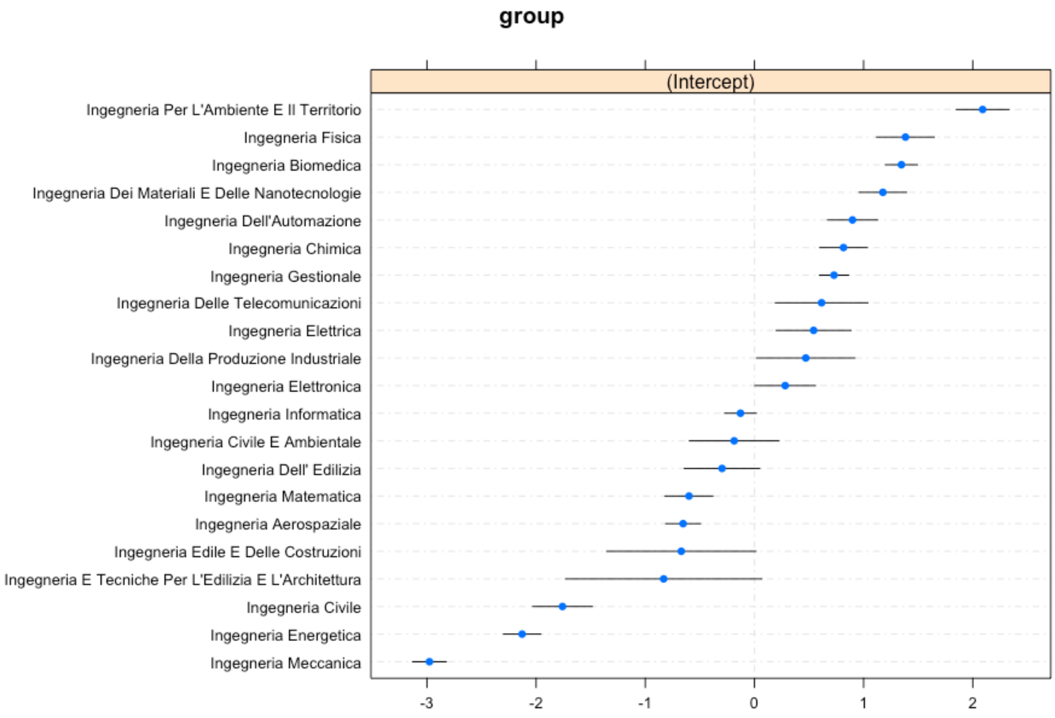


Figure 38. *Random intercepts of Degree Programs estimated by the GMEF model.*

The VPC is the highest compare to GLMER and GMET results, but then the predictions given by GMERF are very accurate for negative observation but with low levels of sensitivity, that is our core measure of performances.

$$\text{VPC} = \frac{\partial_m^2}{\partial_m^2 + \pi^2/3} = 0.315181$$

***Table 12.** Misclassification table of test set predictions by GMEF*

Y	**Actual Y = 1**	**Actual Y = 0**
Predicted Y = 1	427	139
Predicted Y = 0	341	4794

In the following table the results obtained from the different models are presented in order to get a general summary.

Table 13. Summary of performance results employing data at the end of the 1st semester

1st - semester			
	Accuracy	**Sensitivity**	**Specificity**
Logistic Regression	0.8883	0.8229	0.8984
Random Forest	0.9153	0.5586	0.9708
GLMER	0.9202	0.9076	0.9222
GMET	0.8479	0.8529	0.8472
GMEF	0.9158	0.556	0.9718

Then a graph is proposed in order to give an overview on the distribution of high achieving students among the course of studies. Since the different courses of study are 21, for simplicity I decided to indicate only six courses. I chose those that appear to be relevant for the response variable and so Environmental Engineering, Physical Engineering and Biomedical Engineering that contribute positively to the probability of being a talent (in green in the graph) and Civil Engineering, Mechanical Engineering and Computer Science Engineering that are negatively correlated in red. The percentage indicate the proportion of high achieving student per course of study.

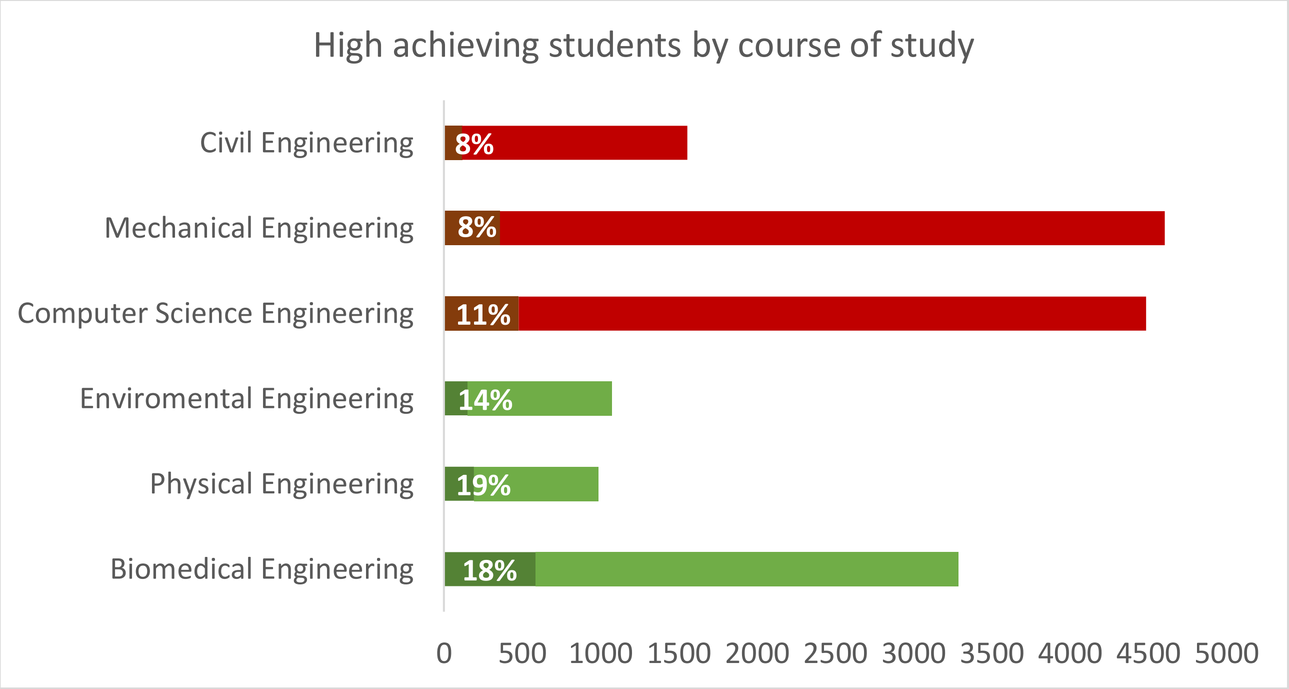


Figure 39. *The percentage in the graph highlights the portion of high achieving students according to the different degree courses.*

5.3 Research question 3: Which are the best models to predict high achieving students?

In this chapter further strategies have been employed in order to improve the models' performances. Then the following hypothesis have been tested:

- **H1**: Is it possible to improve performances with the balancing of dataset?
- **H2**: Is it possible to improve performances thanks to the combination of the best performing algorithms?
- **H3**: How do accuracy and sensitivity measures change according to the definition of an optimal threshold p_0 ?

The answer to these questions has been presented in the following paragraphs.

5.3.1 H1: Unbalanced data

Due to the **disparity of classes** in the response variable, the algorithm tends to classify the observations into the class with more instances, the majority class, while at the same time giving the false sense of a highly accurate model. Both the inability to predict rare events, the **minority class**, and the misleading accuracy weaken the predictive models we build. The following resampling methods have been employed in order to mitigate the inaccuracy in detecting the observations of the minority class:

- **Under-sampling**: we randomly select a subset of samples from the class with more instances to match the number of samples coming from each class. The main disadvantage of under-sampling is that we lose potentially relevant information from the left-out samples.
- **Oversampling**: we randomly duplicate samples from the class with fewer instances or we generate additional instances based on the data that we have, so as to match the number of samples in each class. While we avoid losing information with this approach, we also run the risk of overfitting our model as we are more likely to get the same samples in the training and in the test data, i.e. the test data is no longer independent from training data. This would lead to an overestimation of our model's performance and generalizability.

- **ROSE**: artificial balanced samples are generated according to a smoothed bootstrap approach and allow for aiding both the phases of estimation and accuracy evaluation of a binary classifier in the presence of a rare class.
- **SMOTE "Synthetic Minority Over-sampling Technique"** (Journal of Artificial Intelligence Research, 2002, Vol. 16, pp. 321–357): this paper shows that a combination of the SMOTE method of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space) than only under-sampling the majority class.

Using the package *caret* in R by means of the function *train* we applied the above-mentioned resampling techniques. The results have been compared in terms of accuracy, sensitivity and specificity. The resampling has been run with the Random Forest model.

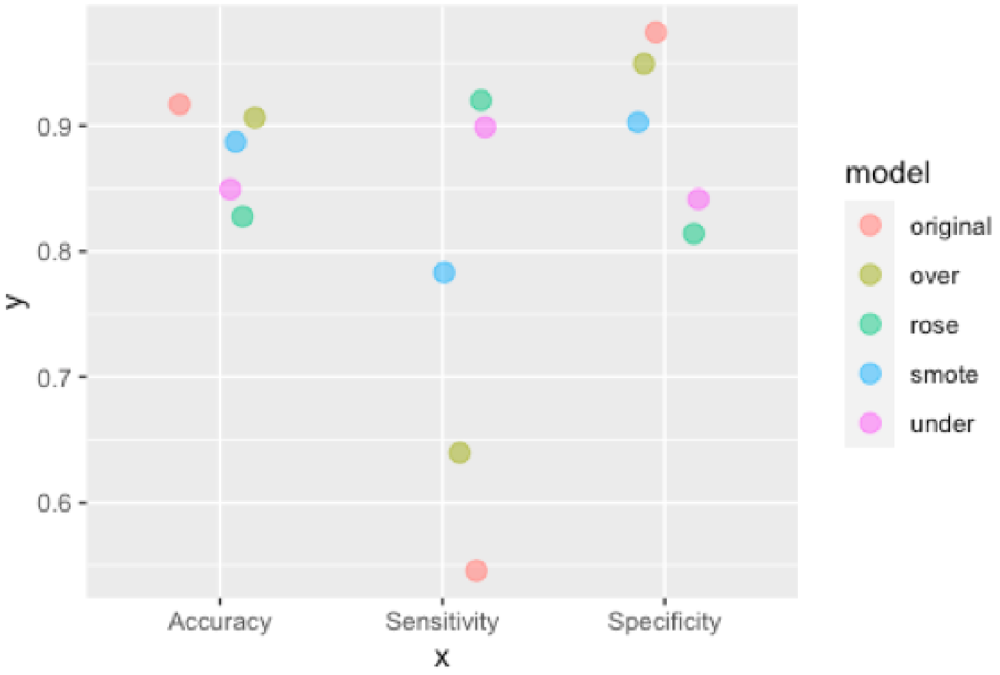


Figure 40. *Performance measures according to different sampling tecniques*

The original model (without sampling) has been compared with others and the results are shown in the previous graph. It can be seen that the accuracy of the prediction for all the models is quite high, it varies from 0.82 to 0.91, with a peak of 0.91 for the original model and followed by the oversampling (over) model. This result was expected, in fact in both the models the majority class prevails and so the algorithms tend to classify the records in the class with more instances.

While, we can see that the original model and the over model have the lowest level of sensitivity which is the percentage of positive instances (true Y=1) which are correctly classified. The highest value of prediction sensitivity is reached by the ROSE and Under models with a 64,5% 59,2% increases respectively compared to the original Random Forest model.

Table 14. *Summary of performance results after applying the sampling tecniques*

1st - semester			
	Accuracy	**Sensitivity**	**Specificity**
Random Forest	0.9153	0.5586	0.9708
RF - Under Sampling	0.8499	0.8997	0.8421
RF - ROSE Sampling	0.8285	0.9193	0.8143

5.3.2 H2: Ensemble methods – Stacking

Ensemble learning is a machine learning paradigm where multiple models (often called "weak learners") are trained to solve the same problem and combined to get better results. The main hypothesis is that when weak models are correctly combined, we can obtain more accurate and/or robust models[1].

The conventional ensemble methods include bagging, boosting and stacking based methods. I will use the stacking method in order to improve the performance of Logistic Regression and Random Forest algorithms. **Stacking** (sometimes called stacked generalization) involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs.

Stacking typically yields performance better than any single one of the trained models. It has been successfully used on both supervised learning tasks (regression, classification and distance learning) and unsupervised learning (density estimation).

[1] https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

But, in order to improve the performances some assumptions are needed. When we combine the predictions of different models using stacking, it is desirable that the predictions made by the sub-models have low correlation (< 0.75). This would suggest that the models are skillful but in different ways, allowing a new classifier to figure out how to get the best from each model for improving the score.

So, we analyzed the predictions for the sub-models (Logistic Regression and Random Forest) and it turns out that they are highly correlated (0.8506182) as shown in the graph and in the table below.

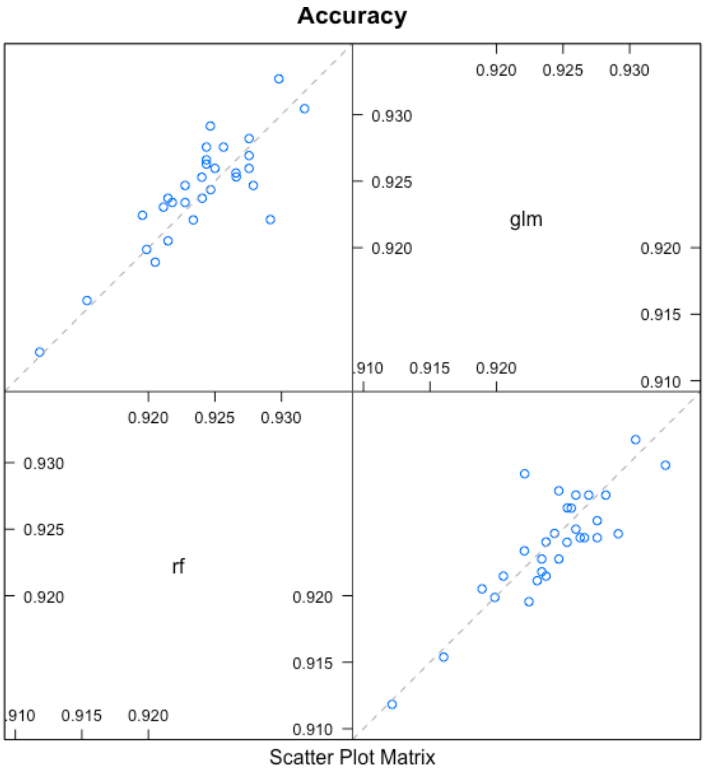


***Figure 41**. Scatter plot showing the correlation between Random Forest and Logistic Regression predictions*

***Table 15.** Matrix of correlations between Random Forest and Logistic regression models. Obtained with the function >modelCor(results)*

> modelCor(results)

	Random Forest	**Logistic Regression**
Random Forest	1.0000000	0.8506182
Logistic Regression	0.8506182	1.0000000

Employing stacking would not improve performances since the two models would be making the same or very similar predictions most of the time reducing the benefit of combining the predictions.

5.3.3 H3: Optimal threshold to minimize false negative

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

The following approach has been applied to the algorithms result more performing in the previous analysis: Logistic Regression and Random Forest.

It is important to highlight that I give more importance to the false negative rate since it measures the number of high achieving students that the algorithm was not able to predict and from our point of view, the misclassification “cost” could have been higher than the one related to the false positive rate. In fact, since our final goals is to provide additional resources to top performer students, it is important to identify them correctly as much as possible.

Consequently, my aim is to maximize the sensitivity in order to minimize the false negative rate. To reach this objective I tried to define a threshold p_0 that respected the following assumption:

- **H0**: The false negative must not exceed the 10% of the whole class Y=1;
- **H1**: The false positive should not exceed the 10% of the whole class Y=0.

The validation set has been used to choose the optimal value p_0, by looking at the indexes defined in the section 5.1 and the relative ROC curves.

The graph below highlights the distribution of students in each class for the validation dataset:

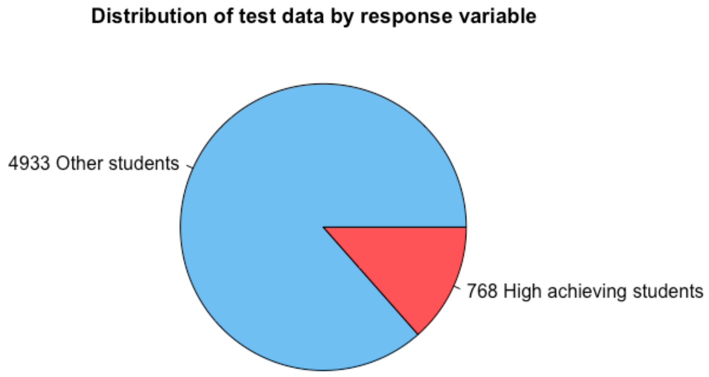


Figure 42. *Pie chart showing the proportion of observations by the response class in test set*

Firstly, I employed the Logistic Regression algorithm. So, by looking at prediction results through the trial and error strategy the optimal value turns out to be $p_0 = 0.18$ such that the H0 hypothesis is respected (FN lower than the 10% of positive observations). The related ROC curve and misclassification table are presented as follows:

***Table 16**. Misclassification table of test set predictions by Logistic Regression with p0 = 0.18*

	Actual Y = 1	**Actual Y = 0**
Predicted Y = 1	694	662
Predicted Y = 0	74	4271

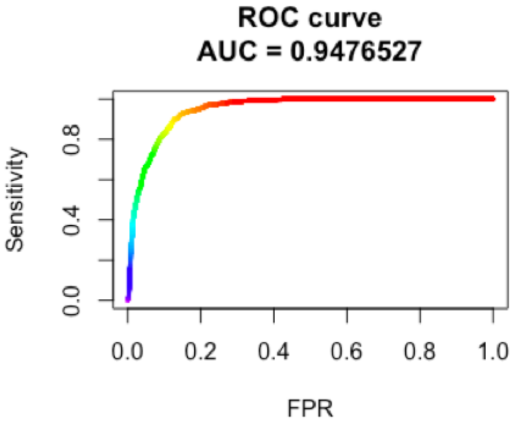


***Figure 43.** ROC curve resulting from the estimations on the test data by the Logistic Regression algorithm with p0 = 0.18*

It is also emerged as the hypothesis H0 and H1 are not suitable, since decreasing the p_0 would means a decrease in the false positive rate at the cost to an increase of the false negative rate, that is our final goal. The best result for FP is around the 13,4% of misclassified observations.

The same scenario of non-compatible hypothesis has been presented for the Random Forest model. The optimal value according to which is respected at least the H0 hypothesis is $p_0 = 0.12$. The following misclassification table represents the output of predictions with $p_0 = 0.12$

***Table 17.** Misclassification table of test set predictions by Random Forest with p0 = 0.12*

	Actual Y = 1	**Actual Y = 0**
Predicted Y = 1	694	740
Predicted Y = 0	74	4193

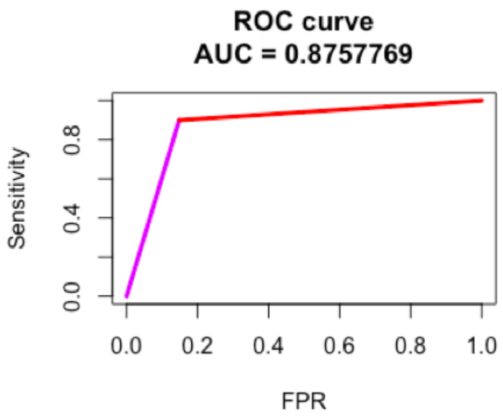


Figure 44. *ROC curve resulting from the estimations on the test data by the Random Forest algorithm with p0 = 0.22*

So, the number of FN with the employed Random Forest (RF) model is the same of the Logistic regression, but on equal terms RF performs slightly worst with a 15% of misclassified in the negative class.

Then considering the overall measures and comparing them with the original models, we can see that for both the algorithm there is an improvement as far as concern the sensitivity measure (as expected since the improvement of true positives was our goal).

Table 18. *Summary of the performance results after defining the optimal value $p_0$for minimizing the false negative rate*

	ACCURACY	**SENSITIVITY**	**SPECIFICITY**
LOGISTIC REGRESSION	0.8883	0.8229	0.8984
LOGISTIC REGRESSION P_0 = 0.18	0.8709	0.9036	0.8658
RANDOM FOREST	0.9153	0.5586	0.9708
RANDOM FOREST P_0 = 0.12	0.8752	0.9036	0.85

5.4 Summary of key results

The purpose of this chapter is to summarize the main takeaways that has emerged fromm the analysis ofdata. The goals of detecting high achieving student has been addressed though the identification of a binary response variable Y which indicates:

- Y = 1 if he/she is a top performer student in terms of academic performances and time to graduation;
- Y = 0 if he/she is not.

Proceeding in stages, the first phase of model implementation has regarded the comparison of two main datasets: the one consisting of academic data at the end of the first year and the other with academic data at the end of the first semester. Linear classification and tree-based classification models have been applied to both datasets and at the end, Logistic Regression and Random Forest have proved superior performances in terms of accuracy and sensitivity indexes. Applying the Logistic Regression model to the dataset gives a maximum value of 0.82 of sensitivity, while the Random Forest performs better in terms of accuracy (0.92).
Thus, from this first step it can be said that demographic, historical (high school studies) and 1st semester academic performance can be successful predictors for detecting high talented students.

Then, the importance of each attributes on the response variable has been analyzed introducing also multilevel classification models. As expected, the most impacting variables were related to the academic performance and history data about high school. In particular, it has been shows that:

- The high school score, the admission score and the gender (being female) are positively correlated with the likelihood to be a talented student;
- The average attempt per exam, attending a technical high school and surprisingly also the number of CFU earned at the end of the first semester have a negative impact on the response variable Y. This means that the increase in the number of attempts per exam, for example, leads to a decrease in the probability to be a talented student.

Similar results have been obtained with multilevel classification models, except for the attribute related to gender that in the GLMER and GMET models seems to not have any relevance. This could be given by the fact that in the single classification model the variance due to different degree courses has been misguided and explained by differences in gender category.

In fact, knowing that educational data are nested, it has been opportune to test the hypothesis according to which the Degree Course (21 groups) can affect the likelihood of being a high achieving student. The GLMER improves the performance in terms of sensitivity, thus increasing the proportion of positive observation correctly classified. Actually, the results show that the random effect given by the degree course clusters can explain the 10% of variance in our predictions. Notably:

- Environmental Engineering, Physical Engineering and Biomedical Engineering are positively correlated with the probability to be in the positive class;
- Mechanical Engineering, Computer Science Engineering, Building, Architectural Engineering and Civil Engineering are negative correlated. It means that the probability to be a talented student if him/she has undertaken one of these degree courses is lower.

At the end of this first phase was clear that the best performing algorithm were those summarized in the below table:

Table 19. *Summary of the best performing algorithms at the end of the first phase*

	Accuracy	Sensitivity	Specificity
Logistic Regression	0.8883	0.8229	0.8984
Random Forest	0.9153	0.5586	0.9708
GLMER	0.9202	0.9076	0.9222

Indeed, the last step was that of implementing additional strategies, to understand if it was possible to improve the performance of the Logistic Regression Model and Random Forest without considering the random effect of the degree courses. The following hypothesis (H*n*) have been tested:

- **H1**: the performance is affected by unbalanced dataset. It has been shown that applying the Random Forest model with *under* or *ROSE sampling* improves by 50-60% the sensitivity of the predictions;
- **H2**: a combination of the two algorithms can lead to better results. H2 was not supported due to the fact that the predictions of the Logistic Regression and the Random Forest were highly correlated, so their combination would not lead to additional benefits;

- **H3**: the definition of an optimal threshold p_0 to minimize the false negative rate can improve the performances. Since the key objective of this analysis is the identification of the positive observations, choosing the p_0 that minimizes the false negative rate turned out to be a precious insight. Indeed, the initial hypothesis H3 has been confirmed: for both the employed models the choice of a common strategy to define the optimal p_0 brought to an improvement in predictions, reaching just the 10% of positive observations misclassified on the overall class.

6. Discussion, policy and managerial implications

As reveled from the literature review, a gap exists in the present-day literature as regards the prediction of students' performances. The majority of the studies are focused on the detection of at-risk students, but also the high achieving ones could have a great impact on the improvement of learning efficacy and the overall success of faculty.

Also, from educational experts' point of view is emerged the necessity to address the task of talents' detection towards more multidimensional measurement method and classificatory approaches to data analysis. Indeed, due to the complexity of the task, it could be not sufficient addressing this issue just with traditional one-dimension measures (i.e QI).

The purpose of this chapter is to answer to the research questions identified at the beginning of the present work. The most relevant findings coming from the data analysis made will be provided, highlighting what innovation has come out compared to the existing literature.

Finally, also policy implications and managerial issues have been considered when implementing analytics at the higher education level.

6.1 Discussion

By elaborating what the literature says about talented students, both from learning analytics and educational stand points, some recurring attributes come out. They have been classified into five main groups: behavioral characteristics, social and personality characteristics, cognitive abilities, time related characteristics, academic achievement.

These categories and the related variables are shown in the following table:

Table 20. *Summary of the main classes of typical attributes for a high achieving student*

	High achieving student typical attribute
Behavioral characteristics	Level of participation in class (traditional and virtual) Level of participation in discussion (traditional and virtual) Participation in extra activities/courses Prone to conducting scientific researches Report studying abroad experience

Social and personality characteristics	Conscientious
	Open to experience
	Emotionally instable
	Introverted
Cognitive abilities	Processing speed
	Working memory
	Intelligence (QI)
Time related characteristics	Time to graduation
	The participation in class is constant since the beginning
Academic achievement	First term GPA
	GPA at the end of the first year
	Final course grade
	Internship
	High school education
	Test scores

The objective of the analysis of data was to test the relevance of these attributes thanks to the application of machine learning models. In accordance with the data available from “Politecnico di Milano” database, the relevance of some of these variables has been tested employing binary classification algorithm. Only time-related characteristics and academic achievement attributes have been used in the present work. Then some additional variables, derived from the initial dataset received, have been added to those in reported in the table. As for example the average attempt per exam or the type of course study that from the explanatory analysis comes out to be relevant in relation with the response variable.

Since there was not a clear definition of what a high achieving student is, in the present work it has been defined as a combination of great academic scores and speed in time to graduation. A binary response variable **Y** has been created, where Y = 1 is the positive condition and represents the high achieving students’ class. According to this classification, a high achieving student has a graduation score higher than 101 and he/she graduates within 3 years. The class of students that belong to Y = 0 includes all other students.

In the following sections the identified research questions have been answered, highlighting the novelty coming from the analysis I made.

6.1.1 Research question 1: Is it possible to predict high achieving students since the earliest stage of their academic career?

To answer to the first research question, two different models have been employed distinguished by time. One model consists of the status of academic variables at the end of the first semester and the second one at the end of the first year.

As expected, the 1st year model comes out with better performances in terms of predictions accuracy and sensitivity since the information are closer to the time to graduation, that is one of the parameters used to define a high achieving student. But it must be taken in account the value of timing. The outcomes of the 1st semester model performed well and the increase in performance' predictions due to the upgraded information at the end of the 1st year would not be worthy of the wait, since that period of time could be employed for exploiting the talent of students.

In fact, it would be advantageous since the beginning of the career to involve the student classified as high achieving in tailored initiatives, as for example involving them in university researches or honor programs. This can lead to the creation of a community with high-profile bachelor's students that combine multi-disciplinary knowledge from different course of study and faculty.

So, following the above considerations and according to the results obtained from the analysis it is possible to predict high achieving students at the end of first semester with a prediction accuracy and sensitivity in a range between 88-91%.

A future development in this sense could be given by an integration of previous high school studies, as for example including the INVALSI score.

6.1.2 Research question 2: Which are the relevant attributes to predict a talented student?

After having understood that is possible to predict high achieving students since the first semester with prediction accuracy higher than 88%, the focus is on the on the most impacting variables for predicting these students.

It seems from the analysis that there is not only a matter of academic scores, but also of their past results at the high school and of the course of study they chose to enroll in.

The results highlight the following predominant variables among high achieving students:

- The higher the number of attempts per exam, the lower will be the probability to be a high achieving student;
- Female have higher probability to be top performer students;
- The number of total CFU gained is negative correlated to the response variable if it is considered stand-alone. While its correlation with the weighted average evaluation increases the probability to be a high achiever. As expected, this means that to be classified as a top student is a matter also of the quality of the performance and not just quantity;
- The high school score and the admission score are positively correlated with the response variable;
- Students who were enrolled in a technical high school have lower probability to be top performers.

Contrary to what has been expected, the mobility (that was one of the behavioral characteristics emerged from the literature) has no impact on the dependent variable. The same applies for Residence variables and previous studies. It is not verified that who was enrolled in classic or scientific high school has more chance to have success at the university, the only state deriving from our results is that who was enrolled in a technical high school can find more difficulties compared to others.

Then, a multilevel classification has been employed, in order to understand if the enrollment in a study course over another has some impact in the predictions of talents. The following findings were made:

- The courses of Environmental Engineering, Physical Engineering and Biomedical Engineering show a positive correlation with the dependent class Y. Meaning that being enrolled in one of these courses of study improves the probability to be a high achieving student;
- Mechanical Engineering, Computer Science Engineering, Building, Architectural Engineering and Civil Engineering have a negative impact on the probability to be a top performer.

The other courses of study not mentioned were found to be not relevant for predicting high achieving students.

The idea for a future work that goes in the same direction of this project could be the integration of data coming from virtual learning environments (i.e. MOOCs, digital platforms, online forums) that could explain also the impact of behavioral attributes as for example the participation in class and in forum discussions.

6.1.3 Research question 3: Which are the best models to predict high achieving students?

From the analysis made is emerged that the best performing algorithms to predict high achieving students were:

- Multilevel logistic regression that reach accuracy and sensitivity higher than 90%;
- Good results were obtained with Logistic regression algorithm but with prediction accuracy and sensitivity lower than 90%;
- Random Forest performs very well in terms of accuracy (around the 92%) but it is worse in terms of sensitivity (around the 55%).

In order to improve the results of Logistic Regression and Random Forest models, some techniques have been applied.

The first one has been the **balancing of data**. Due to the fact that the class of high achieving students constitute only the 12% of the total sample, when the algorithms make predictions there could be bias in the final result. In fact, the algorithms tend to classify the observation as belonging to the class with the greatest number of records. In order to avoid this, several balancing techniques have been applied and it is emerged that the results in terms of sensitivity are considerably higher with parity of accuracy when applying the ROSE technique and the under-sampling.

Then the combination of the two algorithms has been taken into account through the use of the ensemble method: **Stacking**. But this hypothesis was then discarded since the two algorithms produces highly correlated results and then their combination would not bring significant improvement to classification performance.

With the third strategy a common approach has been applied to define the **optimal threshold p_0** when predicting high achieving students with Logistic Regression and Random Forest models in order to minimize the false negative rate. The choice of minimizing the false negative rate was given by the objective of the project itself: the misclassified talented students are more critical regarding our purpose. Since if the talents are not detected would not be possible to address them with additional resources.

After the employment of these techniques all the algorithms used reach prediction accuracy and sensitivity around the 90% as for the multilevel logistic regression model.

6.2 Managerial and policy implications

The EU Joint Research Center (JRC) report described the use of LA in Europe as rapidly developing, but fragmented. The use of LA in Europe to improve the level of teaching and learning is still in its infancy, and the main researches focus on the collection and interpretation of data without using them to improve teaching and learning achievement (*Tsai et al*.,2018).

The advantage given by the earlier identification of high achieving students can be spread across different actors in the faculty. The present work establishes the foundation about the theme of LA for predicting talents at higher education level; some hints are presented as follows to introduce the opportunities deriving from this study that can be helpful for management and policymakers.

The first point of interest can be the selection of talented students for honors programs. As highlighted by *János Szabó* (2019) the aspect of talent-management programs, in the European universities are underrepresented compared to American ones; much less universities have talent program in Europe, and the majority of these can be find in the Netherlands.

Politecnico di Milano has already made some progresses in this sense when it founded **Alta Scuola Politecnica (ASP)** in 2004. It restricted to 150 young and exceptionally talented students, selected solely on the basis of merit among the applicants to the Master of Science in Engineering, Architecture and Design. These students follow a curriculum additional to their degree programmes, completely in English, based on ad-hoc courses, including the development of multidisciplinary projects.

The prediction of high achieving students can be useful in this sense according to several aspects. The first benefit could be given by the introduction of a data-driven decision in the selection of the applicants for the ASP, not anymore based only on academic merit but driven by a multi-factor analysis that has learned from historical data. Then, it could be thought to adopt the same system also for bachelor’s students.

The second point regards the theme of university research. The possibility to identify top performers students can help the faculty to select the most promising talent to include in their research teams, generating then a double benefit:

- If talent students are involved in project research, the quality of the research projects will improve, thus improving the prestigious of the university;

- These students could discover a passion for the research career that maybe they would not consider at all if they had not the chance to be involved;

The third point is related to the point of contact between university and the employment industry. For those students who are not suited for the research career path, the prediction of high achieving students can be useful as well. In a global learning environment, this type of information not only can facilitate better educational and post-education vocational planning, but also may prove useful to organizations as they make hiring and budgeting decisions for college graduates in different disciplines (Avela et al.,2016).

Companies are always seeking of talents, and if the university already knows who the brilliant students are since the beginning of their career, preferential and tailored job opportunities or events could be provided for high achieving students in order to get acquainted about the existent proposals in the job market, to understand their interests and moreover to get in touch with successful persons who can became role models for these students, generating also the creation of a network outside the university. This can have a positive impact when the time of applying for a job will come.

One more point of reflection is given by acceleration of studies. It has been pointed out the practical utility of the *time-to-graduation* criterion, since it is of interest also to policy makers because it can be a benefit in terms of efficiency and reduction in costs. Educational acceleration, for example content or grade-based, has been claimed to be advantageous for high abilities students, because it helps to increase academic achievement of those students who were accelerated, and it saves time and frees up other resources (*Steenbergen-Hu et al.*, 2012). So, for those classified as high achieving, the application of accelerated courses could be beneficial not only for them but also for the faculty.

6.3 Conclusions

The present work aims at contributing to the field of learning analytics in higher education, focusing on the performance prediction of high achieving students.

Filling the gap existing in the present literature regarding the prediction of students' performance, the first step was giving a definition for high achieving students. They have been described in terms of academic achievement, so the final graduation score has been taken into account, and in terms of speed, so the time to graduation has been considered. One of the novelties of this work lies on the definition of the response variable that is a combination of the above-mentioned measures: a high achieving student is who graduates within 3 years at bachelor's level with a final score higher than 101.

The findings revealed that using academic data that represent the situation at the end of the first semester, could lead to prediction accuracy around the 90% when applying Logistic Regression, Random Forest and Multilevel Logistic Regression algorithms. Moreover, applying the ROSE or under-sampling balancing techniques leads to significant improvements of the results when a minority class as that of high achieving students is present (talent students constitute only the 12.4% of the total sample).

In this specific case, the number of misclassified top performer students were more critical, so in order to improve the prediction sensitivity of the employed algorithms an optimal p_0 value has been defined that minimizes the number of false negative in the test set predictions, thus improving the performances.

It has been shown that the most impacting variables for predicting talented students are, as expected, the academic performances, previous high school studies, the gender and the admission score.

In particular:

- Academic performances: the higher the average of attempts per exam and the higher the number of CFU gained at the end of the first semester, the lower the probability to be a high achiever.
- Gender: being a female is positive correlated with the probability to be a top performer.
- Previous high school studies: having attended the technical high school decrease the probability to be a high achiever. While the diploma score shows positive correlation with the response variable.

- Admission score: the higher the score of the university admission test, the higher the probability to be a top performer.

From a multilevel analysis emerged then that graduating in a degree course over another can improve or decrease the probability to be a high achieving student. In fact, the course of Environmental Engineering, Physical Engineering and Biomedical Engineering show a positive correlation with the response variable, while Mechanical Engineering, Computer Science Engineering, Building, Architectural Engineering and Civil Engineering have a negative impact on the probability to be a top performer.

Several benefits might be derived from the identification of talented students. From a practical standpoint, the extent to which a campus can attract the most academically talented students speaks directly to the campus' ability to successfully navigate legislative and public demands for accountability and assessment outcome that center on student success. In short, high achieving students increase the local faculty and national prestige, which directly and indirectly leads to an increase in funding opportunities (*Bradshaw et al.*, 2001). This is the reason why offering valuable opportunities for high achieving student can impact the whole community. Some proposals have been made on this direction, as the reinforcement of honors program also at the bachelor level, the involvement of talents within university research group, an earlier point of contact with the labor industry or accelerated education.

To conclude, the innovative aspects deriving from the present work are several and interesting for various stakeholders. The results coming from this research are fruit of a study about the learning analytics topic combined with the point of view of pedagogical experts, who discuss the theme of talents from decades. The main contribution to the literature is given by the introduction of a definition for talented students in the learning analytics field. While in the previous studies the aim to predict students', performance focused mainly on students failing or in danger of failing, here the purpose is to offer something also to those students already succeeding.

The path to follow is still long because the confirmation of such results needs to be supported by further researchers. It is needed to quantify the benefits coming from the implementation of the above-mentioned initiatives in favor of talented students. Anyway, having defined such classification is something new for the literature.

The present work represents the first footsteps to address high achieving students. Their identification from the beginning of their bachelor’s degree will allow to take actions which could enhance their skills and generate benefits for the whole community: institution management, university research field, employment industry and policymakers. It prepares the ground for further researches that will apply similar models with enlarged datasets, for example introducing VLEs data can help to characterize the behavior of talented students considering the interaction in virtual class or in forum discussion. In the light of the most recent events, the faculty has put in place virtual classes that can generate vast amount of data. Future studies could also investigate in deep the actions to be undertaken for talents, evaluating the benefits of their implementation.

References

Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *Mining educational data to improve students' performance: a case study*, *2*(2).

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, *37*, 13-49.

Ali, A., Ali, S., Khan, S. A., Khan, D. M., Abbas, K., Khalil, A., ... & Khalil, U. (2019). Sample size issues in multilevel logistic regression models. *PloS one*, *14*(11).

Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270).

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. Statistics in medicine, 36(20), 3257-3277.

Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning*, *20*(2), 13-29.

Azcona, D., Hsiao, I. H., & Smeaton, A. F. (2018). PredictCS: Personalizing Programming learning by leveraging learning analytics.

Azcona, D., Hsiao, I. H., & Smeaton, A. F. (2018, October). Personalizing computer science education by leveraging multimodal learning analytics. In *2018 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE.

Belloc, F., Maruotti, A., & Petrella, L. (2010). University drop-out: an Italian experience. *Higher education*, *60*(2), 127-138.

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220-238.

Brohi, S. N., Pillai, T. R., Kaur, S., Kaur, H., Sukumaran, S., & Asirvatham, D. (2019, August). Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education. In *International Conference for Emerging Technologies in Computing* (pp. 254-261). Springer, Cham.

Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability*, *11*(10), 2833.

Ciolacu, M., Tehrani, A. F., Beer, R., & Popp, H. (2017, October). Education 4.0—Fostering student's performance with machine learning methods. In *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)*(pp. 438-443). IEEE.

Cognard-Black, A. J., & Spisak, A. L. (2019). Creating a Profile of an Honors Student: A Comparison of Honors and Non-Honors Students at Public Research Universities in the United States.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017, April). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion* (pp. 415-421).

De Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., ... & Arnab, S. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *British journal of educational technology*, *46*(6), 1175-1188.

Dietz-Uhler, B., & Hurn, J. E. (2013). Using learning analytics to predict (and improve) student success: A faculty perspective. *Journal of interactive online learning*, *12*(1), 17-26.

Dziuban, C., Moskal, P., Cavanagh, T., & Watts, A. (2012). Analytics that Inform the University: Using Data You Already Have. *Journal of Asynchronous Learning Networks*, *16*(3), 21-38.

El George Bradshaw, S. E., & Hausman, C. (2001). The college decision-making of high achieving students. *College and University*, *77*, 2.

Fontana, L., Masci, C., Ieva, F., & Paganoni, A. M. (2018). Performing Learning Analytics via Generalized Mixed-Effects Trees.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (pp. 175-199).

Heller, K. A., & Hany, E. (2004). Identification of gifted and talented students. *Psychology Science*, *46*(3), 302-323.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3-7). New York: springer.

Jha, S., Jha, M., & O'Brien, L. (2018, December). A Step towards Big Data Architecture for Higher Education Analytics. In *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (pp. 178-183). IEEE.

Kappe, R., & van der Flier, H. (2012). Predicting academic success in higher education: what’s more important than being smart?. *European Journal of Psychology of Education*, *27*(4), 605-619.

Koshy, V., Ernest, P., & Casey, R. (2009). Mathematically gifted and talented learners: theory and practice. *International Journal of Mathematical Education in Science and Technology*, *40*(2), 213-228.

Kruse, A. N. N. A., & Pongsajapan, R. (2012). Student-centered learning analytics. *CNDLS Thought Papers*, 1-9.

Kumar, S. R., & Hamid, S. (2017, November). Analysis of Learning Analytics in Higher Educational Institutions: A Review. In *International Visual Informatics Conference* (pp. 185-196). Springer, Cham.

Larrabee Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*(5), 2594-2618.

Lee, L. K., & Cheung, S. K. (2020). Learning analytics: current trends and innovative practices. *Journal of Computers in Education*, *7*(1), 1-6.

Manning, C. (2007). Generalized Linear Mixed Models (illustrated with R on Bresnan et al.'s datives data). *línea:< nlp. stanford. edu/manning/courses/ling289/GLMM. pdf>(consultado el 8 de junio de 2016)*.

McCuaig, J., & Baldwin, J. (2012). Identifying Successful Learners from Interaction Behaviour. *International Educational Data Mining Society*.

Merceron, A. (2015, September). Educational Data Mining/Learning Analytics: Methods, Tasks and Current Trends. In *DeLFI WOrkshops* (pp. 101-109).

Mustafina, J., Galiullin, L., Al-Jumeily, D., Petrov, E., Alloghani, M., & Kaky, A. (2018, September). Application of learning analytics in higher educational institutions. In *2018 11th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 163-168). IEEE.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438-450.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences*, *10*(3), 1042.

Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational intelligence magazine*, *11*(1), 41-53.

Román-González, M., Pérez-González, J. C., Moreno-León, J., & Robles, G. (2018). Can computational talent be detected? Predictive validity of the Computational Thinking Test. *International Journal of Child-Computer Interaction*, *18*, 47-58.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Saqr, M., Nouri, J., & Fors, U. (2019). Time to focus on the temporal dimension of learning: A learning analytics study of the temporal patterns of students' interactions and self-regulation. *International Journal of Technology Enhanced Learning*, *11*(4), 398-412.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).

Sin, K., & Muthu, L. (2015). APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW. *ICTACT journal on soft computing*, *5*(4).

Steenbergen-Hu, S., & Moon, S. M. (2011). The effects of acceleration on high-ability learners: A meta-analysis. *Gifted Child Quarterly*, *55*(1), 39-53.

Szabó, J. (2019). How Can Be Academic Talent Measured During Higher Education Studies?-An Exploratory Study. *Higher Education*, *9*(4).

Vercellis, C. (2009). *Business intelligence: data mining and optimization for decision making* (pp. 1-420). New York: Wiley.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98-110.

Yildirim, M. C. (2014). Developing a scale for constructivist learning environment management skills. *Eurasian Journal of Educational Research*, *54*, 1-18.