POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Early Warning System (EWS) in K-12 Italian Education

TESI MAGISTRALE IN MANAGEMENT ENGINEERING – INGEGNERIA GESTIONALE

AUTHORS: GIACOMO MARANI TASSINARI – 10661042
EVANGELIA MYLONOPOULOU – 10905670

SUPERVISOR: TOMMASO AGASISTI

CO-SUPERVISOR: MELISA LUCIA DIAZ LEMA

ACADEMIC YEAR: 2023 – 2024

1. Introduction

Early School Leaving (ESL), defined as the percentage of young people aged from 18-24 who leave education and training without completing upper secondary education, remains a significant issue in Europe, particularly in Italy, where the dropout rate is 10.5% and spikes to 14.6% in Southern regions (Eurostat, 2023; ISTAT, 2023). ESL not only hinders individual growth but also fuels unemployment rates and economic disparities, becoming a problem at a societal level. Research links dropout rates to socio-economic challenges, absenteeism, and academic performance, especially among foreign-born students and those in vocational tracks.

To address this, Early Warning Systems (EWS) have emerged as effective tools, leveraging real-time data on attendance and grades, as well as socio-emotional indicators to identify students at risk before they disengage entirely. In Italy, where regional and demographic disparities are significant, EWS could transform intervention strategies, anticipating risk factors and providing timely alerts to educators, parents, and policymakers to provide targeted support. With continuous data from electronic school registers and national datasets, EWS could proactively address dropout risks, helping students stay engaged and succeed in their educational journeys.

2. Research Questions

This thesis aims to develop an EWS specifically tailored to Italy's K-12 landscape, utilizing data from the Italian Ministry of Education (MIM) and INVALSI assessments to enable data-driven identification of students at risk of dropping out or underperforming academically. Our research will focus on the following questions.

Research Question 1: ***"Which factors influence admission in the following academic year?"***

This research question explores the various academic, socioeconomic, and institutional factors that affect student retention in Italian high schools.

*Research Question 2: "****What are the drivers of dropout during the transition from Middle School to High School?"***

By answering this question we expect to gain deeper insights into the dynamics of that critical time period.

*Research Question 3: "**What are the best Machine Learning Models to predict dropout and grade retention?**"*

By focusing on two critical outcomes—school dropout and grade retention—we aim to identify the most effective approach among the ones present at state-of-the-art literature for tackling these pressing issues in the Italian educational system.

*Research Question 4: "**Can real-time data enhance the predictive power of an Early Warning System?"***

We aim to understand how real-time data can enhance the predictive performance in the context of education.

*Research Question 5: "**How does the quality of educational data affect predictive power?"***

We aim at having a deeper understanding of the impact of the quality of the data on the predictions of the models.

3. Data

The datasets for this analysis combine information from the Italian Ministry of Education (MIM) and from INVALSI assessments, covering the academic years 2020 to 2023 and providing a longitudinal view of student performance, including grades, INVALSI test scores, attendance records, socioeconomic and demographic data, as well as dropout status for 2022. The data spans key educational milestones, capturing insights from middle school and INVALSI assessments conducted in both grade 8 and grade 10.

Table 1: Data Used to predict 10th Grade Outcomes

	2020/21	***2021/22***	***2022/23***
Grade	8th	9th	10th
Data Available	Esiti MM 2020/21 INVALSI G8 2021	Esiti SS 2021/22	Esiti SS 2022/23 INVALSI G10 2023
Filter Applied	cod_ann_cor = 3	cod_ann_cor = 1	cod_ann_cor = 2

By integrating the information summarized in Table 1, **434,836 students** have been identified. Similarly, in Table 2, **407,907** students were retained in total, **out of which 5,075 dropouts** were detected (1.24%).

To handle missing data, the K-Nearest Neighbors (KNN) imputation technique was applied, aiming to fill gaps with values that align closely with the dataset's patterns.

Table 2: Data Used to Predict 9th Grade Dropout

	2020/21	***2021/22***
Grade	8th	9th
Data Available	Esiti MM 2020/21 INVALSI G8 2021	Abbandono 2021/22
Filter Applied	cod_ann_cor = 3	cod_ann_cor = 1

4. Methodology

In order to build a robust Early Warning System (EWS) for predicting student outcomes and dropout risk, we employed 10 Machine Learning models across two datasets with distinct targets: annual school outcome prediction and dropout prediction. For outcome predictions, we paid attention to students' longitudinal academic path from 2020/21 to 2022/23, excluding 2022/23 grades to avoid circular logic, as these grades are finalized simultaneously with the outcome itself. To optimize model performance, we scaled the data when applicable, tuned hyperparameters through extensive grid searches with 3-fold and 5-fold cross-validation for outcome prediction and dropout prediction, respectively, and applied the synthetic minority oversampling technique, namely, SMOTE, to address the dataset's class imbalance in dropout prediction models that performed poorly. In addition to using imputed data, we also trained our outcome prediction

models considering a subset containing only complete cases to assess the influence of imputation on accuracy. Furthermore, we explored the potential of real-time data integration in the models by including 2022 grades in one version of the outcome dataset, allowing us to gauge how access to up-to-date information could enhance prediction effectiveness.

5. Results

Metrics such as macro-averaged sensitivity, specificity, AUC-ROC, AUC-PR and F1-score provided insights into model evaluation regarding outcome prediction. Logistic Regression and LASSO Regression excelled in terms of sensitivity (recall) (0.64), achieving one of the highest specificity metrics as well (0.84). Random Forest demonstrated a macro-averaged F1-score of 0.61, a specificity of 0.85 and an AUC-ROC of 0.88. XGBoost achieved an F1-score of 0.60, an AUC-ROC of 0.89 and an AUC-PR of 0.66 – the highest reported among all models. Support Vector Machines also performed reasonably well, balancing sensitivity (0.63) and specificity (0.85) while achieving an AUC-ROC of 0.87. A similar performance was reported by Neural Networks, which excelled in precision-oriented tasks with an AUC-PR of 0.64 but showed lower sensitivity.

.

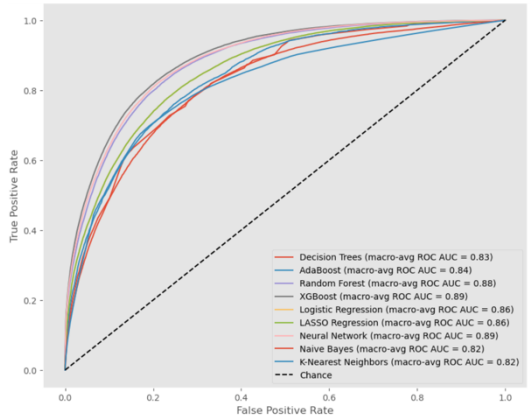

Figure 1: ROC-AUC of all outcome models

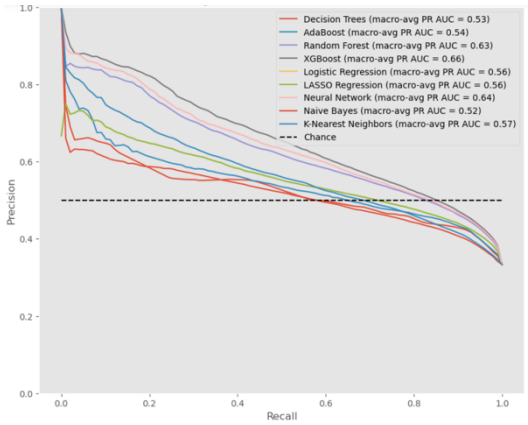

Figure 2: PR-AUC of all outcome models

Among the models, Random Forest and XGBoost emerged as the most effective for identifying at-risk students ("SOSPENSIONE DAL GIUDIZIO" and "NON AMMESSO/A"). Random Forest demonstrated a balanced performance, achieving a recall of 0.62 for identifying suspended students and a precision of 0.50 for the not admitted ones, respectively, while XGBoost, on the other hand, excelled in precision for "NON AMMESSO/A" (0.70) while maintaining competitive recall for "SOSPENSIONE DAL GIUDIZIO (0.44). Logistic and LASSO Regression reported a high recall (0.68) in predicting not admitted students, but a much lower precision (0.22). Overall, Random Forest and XGBoost stood out as the most robust and balanced models, making them well-suited for student classification.

Table 3: Evaluation Metrics for Outcome Prediction

	Accuracy	*Sensitivity*	*Specificity*	*Precision*	*Recall*	*AUC-ROC*	*AUC-PR*	*F1-score*
Random Forest	0.80	0.61	**0.85**	0.63	0.61	0.88	0.63	**0.61**
Decision Trees	0.72	0.61	0.83	0.51	0.61	0.83	0.52	0.53
AdaBoost	0.71	0.63	0.84	0.50	0.63	0.84	0.54	0.52
XGBoost	**0.83**	0.55	0.81	**0.71**	0.55	**0.89**	**0.66**	0.60
Support Vector Machines	0.73	0.63	**0.85**	0.53	0.63	0.87	0.58	0.55
Logistic Regression	0.72	**0.64**	0.84	0.51	**0.64**	0.86	0.56	0.53
LASSO Regression	0.72	**0.64**	0.84	0.51	**0.64**	0.86	0.56	0.53
Neural Networks	0.82	0.54	0.81	0.69	0.54	**0.89**	0.64	0.58
Naïve Bayes	0.73	0.59	0.83	0.51	0.59	0.82	0.52	0.52
K-Nearest Neighbors	0.81	0.49	0.77	0.62	0.49	0.82	0.57	0.54

With the inclusion of real-time predictors, model performance improved significantly, particularly in terms of accuracy, AUC-ROC, and AUC-PR. Random Forest and XGBoost stood out again, achieving AUC-ROC scores of 0.97 and AUC-PR scores of 0.86 and 0.91, respectively, demonstrating strong predictive ability across all outcome categories. Neural Networks also performed exceptionally well, with an AUC-PR of 0.98, furtherly highlighting their effectiveness. These findings highlight the importance of leveraging real-time data and demonstrate that ensemble methods like Random Forest and XGBoost, alongside Neural Networks, are particularly effective for identifying at-risk students.

Table 4: Evaluation Metrics for Outcome Prediction (including 2022 grades)

	Accuracy	*Sensitivity*	*Specificity*	*Precision*	*Recall*	*AUC-ROC*	*AUC-PR*	*F1-score*
Random Forest	0.92	0.81	0.91	0.86	0.81	**0.97**	0.86	0.83
Decision Trees	0.89	0.81	0.91	0.76	**0.82**	0.94	0.86	0.79
AdaBoost	0.88	0.76	0.90	0.77	0.76	0.84	0.79	0.77
XGBoost	**0.93**	0.79	0.91	**0.90**	0.79	0.97	0.91	**0.84**
Support Vector Machines	0.89	0.81	**0.92**	0.77	0.81	0.96	0.87	0.79
Logistic Regression	0.83	0.81	0.90	0.69	0.81	0.94	0.82	0.74
LASSO Regression	0.83	0.81	0.90	0.69	0.81	0.94	0.82	0.74
Neural Networks	**0.93**	0.78	0.91	0.89	0.78	**0.97**	0.98	0.83
Naïve Bayes	0.77	0.69	0.86	0.58	0.69	0.88	0.63	0.61
K-Nearest Neighbors	0.87	0.66	0.84	0.84	0.66	0.91	**0.94**	0.72

Regarding dropout prediction, the evaluation of models reveals a varied, and in cases poor performance, with Neural Networks reporting the highest sensitivity (0.81) in predicting dropout. Logistic and LASSO Regression appear to perform robustly, achieving a sensitivity of 0.77 and an AUC-ROC of 0.81, making them strong candidates for dropout identification. Decision Trees and XGBoost achieved a high sensitivity as well, with Decision Trees scoring 0.70 and XGBoost reaching 0.73. Decision Trees demonstrate a notable specificity (0.74) and AUC-ROC (0.79), balancing detection of dropouts and non-dropouts effectively. XGBoost reports a comparable specificity (0.75) and AUC-ROC (0.79).

Table 5: Evaluation Metrics for Dropout Prediction

	Accuracy	*Sensitivity*	*Specificity*	*Precision*	*Recall*	*AUC-ROC*	*AUC-PR*	*F1-score*
Random Forest	0.70	0.65	0.70	0.03	0.65	0.72	0.04	0.05
Decision Trees	0.73	0.73	0.74	0.03	0.73	0.79	0.13	0.06
AdaBoost	0.70	0.65	0.70	0.03	0.65	0.73	0.20	0.05
XGBoost	0.75	0.70	0.75	0.03	0.70	0.79	0.05	0.07
Support Vector Machines	0.83	0.50	0.83	0.04	0.50	0.73	0.04	0.07
Logistic Regression	0.74	0.77	0.74	0.04	0.77	0.81	0.05	0.07
Neural Networks	0.64	0.81	0.63	0.03	0.81	0.77	0.04	0.05
Naïve Bayes	0.93	0.30	0.93	0.05	0.30	0.72	0.03	0.09
K-Nearest Neighbors	0.93	0.21	0.94	0.04	0.21	0.74	0.08	0.07
LASSO Regression	0.74	0.77	0.74	0.04	0.77	0.81	0.05	0.07

6. Discussion

The outcome prediction models consistently highlighted as key predictors grades in Italian and Mathematics, INVALSI scores (particularly from grade 10), and socioeconomic indicators. These predictors proved instrumental, with standardized assessment scores and attendance data also serving as valuable markers for academic risk. The inclusion of historical INVALSI scores from grade 8, among the impactful variables, demonstrated the importance of longitudinal data, providing early signals for future academic challenges. Moreover, ESCS variables at the school level and the outcome of the previous year also demonstrated strong explanatory power.

A notable finding in dropout prediction is that demographic factors, such as having a foreign background and being older than the typical age for ones grade, are key contributors to dropout risk. Moreover, geographical location plays a highly significant role, with the Northwest showing the highest dropout rates, while the South exhibits the lowest.

Lastly, we evaluated the role of data completeness by comparing models trained on fully imputed data versus models restricted to complete cases. Results showed that imputing missing values preserved more at-risk indicators, as students with incomplete records often had higher dropout rates. Real-time data, while boosting predictive power, reduced interpretive clarity due to strong correlations with outcomes. However, the inclusion of such data improved model performance metrics significantly, underscoring its potential in refining EWS applications.

7. Limitations

Our study provided valuable insights but faced several limitations. Missing data affected model performance, as relying only on complete cases skewed results toward higher-performing students and reduced predictive accuracy. To address this, we used data imputation, which preserved a larger sample size but potentially masked indicators of risk, such as fluctuating performance or attendance issues. Additionally, excluding 2022 grades from the models avoided logical inconsistencies, though classifiers using these grades showed significantly stronger predictive power.

The absence of class-level data, such as average class performance, limited our understanding of peer effects and classroom dynamics that could influence individual student outcomes. Additionally, data on the root causes behind socio-economic influences on performance, as well as student motivation and engagement, were unavailable, hindering a deeper exploration of dropout behavior. Future research should incorporate these missing variables to provide a more comprehensive understanding of the factors affecting student success and dropout risk.

8. Recommendations

Our research highlights the need for broader predictors in dropout modeling, such as student engagement, family support, and peer effects, which are often overlooked in traditional datasets. Including historical data across multiple years and metrics like book borrowing and online activity could improve accuracy in identifying at-risk students.

Incorporating classmates' achievements could reveal valuable insights into the peer effects on academic performance. Better handling of missing data would also make models more inclusive, especially for vulnerable groups. Lastly, socio-economic and demographic factors remain essential but complex influences, emphasizing the need for a holistic view to capture the full range of challenges impacting at-risk students.

9. Conclusions

This study examines the potential of implementing an Early Warning System (EWS) in Italian K-12 education to detect students at-risk, boost educational success, and reduce dropout rates. The research focused on the feasibility of predicting at-risk students through machine learning, the effectiveness of different models, and the main predictors of academic risk.

Results showed that models like Random Forest, XGBoost, LASSO Regression, and Neural Networks are effective in identifying at-risk students. Key predictors include past grades in core subjects, INVALSI scores, attendance, and socioeconomic indicators. INVALSI scores from earlier school cycles proved particularly useful in identifying risks early and facilitating targeted interventions. Demographic and socioeconomic factors also played a role, pointing to the importance of context-sensitive support for students.

Implementing Early Warning Systems with these predictors can help schools detect students at-risk, reduce dropout rates, and enhance academic outcomes. Attention should be paid to data collection, standardized protocols, and integrated data systems for effective EWS use. Further research should refine these models and predictors, and improve data handling techniques, to make EWS more viable in real-world educational settings. Embracing a data-driven, individualized approach to student support holds great potential for strengthening resilience and academic success across diverse student groups.

10. Bibliography

[1] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.

[2] Bowers, A. J. (2021). Early warning systems and indicators of dropping out of upper secondary school: The emerging role of digital technologies. *OECD Digital Education Outlook 2021 Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, 173.

[3] Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124.