NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Using Machine Learning Models to predict high school student's Academic Achievement

Afonso João Mendes Quintino

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Using Machine Learning Models to predict high school student's Academic Achievement

by

Afonso João Mendes Quintino

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Frederico Cruz Jesus, PhD, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 15, 2024

ABSTRACT

Understanding student dropout has become increasingly relevant given the growing importance of educated people in today's workforce. Therefore, predicting a student's academic achievement (AA), whether he/she passes the academic year or not, can prove crucial to assisting teachers and competent decision-makers to create measures to help retain and eventually reduce academic abandonment. To address such issues, this paper utilizes Machine Learning (ML) models to obtain accurate predictions of essentially every student's AA in Portuguese public high schools using data from the Portuguese Ministry of Education and understand what are the drivers of AA that most affect the predictive abilities of said models. Our results show that Random Forest and XGBoost have similar levels of accuracy, however, the latter displayed slightly better predictions. Regarding the most influential AA drivers, previous retention, gender, and the location of the student's school were the ones that showed the greatest effect on the XGBoost model's ability to accurately predict the student's success. Several suggestions are made to educational stakeholders on the results of this study.

KEYWORDS

Education; Academic Achievement; Artificial Intelligence; Machine Learning

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AA	Academic Achievement
AI	Artificial Intelligence
ML	Machine Learning
NN	Neural Networks
DT	Decision Trees
ERT	Extremely Randomized Trees
RF	Random Forest
SVM	Support Vector Machines
KNN	k-Nearest Neighbors
LR	Logistic Regression
XGB	Extreme Gradient Boosting
BT	Bagged Trees
ABT	Adaptive Boosting Trees
MLP	Multi-Layer Perceptron
SMOTE	Synthetic Minority Over-Sampling Technique
RFE	Recursive Feature Elimination
LASSO	Least Absolute Shrinkage and Selection Operator
SL	Supervised Learning
UL	Unsupervised Learning
ANN	Artificial Neural Networks
NBC	Naïve Bayes Classifier
GBDT	Gradient Boosting Decision Tree
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
PFI	Permutation Feature Importance

1. INTRODUCTION

Education has been a topic of discussion since ancient Greece (Griffith, 2015), and its benefits are evident on both personal and societal levels. Tilak (2002) emphasized that education broadens understanding, strengthens democratic processes, and promotes sustainable development through a better understanding of environmental and ecological relationships. Studies across various countries and periods have shown that education enhances the quality of life, work opportunities, and professional prospects (Dumciuviene, 2015).

A positive correlation exists between education and earnings, with educated individuals earning higher wages and experiencing lower unemployment rates. In Portugal, the unemployment rate for those with only basic education ranges from 7% to 12.8% according to Pordata. Education also serves as a crucial tool for inclusion (UNESCO, 2008). Marginalized and vulnerable groups, such as minorities and those in poverty, often face educational exclusion, which can lead to lower academic performance and early dropout rates if not addressed promptly (Jahnukainen, 2001). Understanding the factors that affect academic achievement (AA) is therefore critical.

AA, the measure of a student's success in achieving certain goals in an academic environment, is a subject that has inspired much interest from researchers from different backgrounds. Scholars have approached this matter from a behavioral standpoint (Amrai et al., 2011), and through a more economic view since education has relevant personal, social, and political implications. Education impacts more than just the classroom, as it has an immensely positive role in such aspects as public health, the development of democratic values, culture enrichment, and the economic fabric of nations, more or less developed (Astakhova et al., 2016). With the acquisition of knowledge and skills, education paves the way for individuals to obtain improved career opportunities and quality of life. This leads to a more competitive labor market. Given the pivotal role education plays, decision-makers must be aware of what influences students' performance at school so they can promote the best strategies to enhance educational performance. Hence, it is important to study and understand AA drivers.

This study is different from previous ones made using AI and Machine Learning models to predict AA as the data used to train, and test said models include practically every student in Portugal since the data source is the Ministry of Education.

In this paper, we use a large sample of students who are then scored with ML models, to quantify how much each AA driver (independent variables) changes the predictive capability of the models. The objective of the application of these methods is to understand what impacts a student's academic success and by how much. The insights achieved with this study should allow schools and upper policymakers to understand better how to introduce more effective strategies to lower retention rates and improve AA. By following this approach, this paper looks to answer the following research questions:

- What are the most accurate ML models to predict AA?

- What are the main drivers for early dropout among high school students in Portugal? How do they affect the predictive capability of the ML models?

To address this question, this paper is outlined as follows: Section 2 includes a Theoretical Background; Section 3 addresses the Methodology; Section 4 presents the Results; Section 5 shows a Discussion and the limitations/implications of this work; Section 6 closes with the conclusions.

2. LITERATURE REVIEW

2.1. EDUCATION AND ACADEMIC ACHIEVEMENT

Academic Achievement (AA), the measure of a student's academic success, can be measured by standardized test scores, a student's grade point average (GPA), and whether the student fails, or not, to pass the year. When students are provided with proper career education interventions, a somewhat positive gain in AA has been noticed (Evans & Burck, 1992). Therefore, understanding AA drivers is fundamental for educational policymakers as these are the ones who can set measures in place to prevent academic underachievement and ensure everything is done to reduce dropout rates. However, measuring AA can be challenging as it is a combination of socioeconomic, psychological, and environmental factors (Pandey & Thapa, 2017).

It has been established long ago that an early introduction of children into the educational scene improves their future AA, and the longer that preliminary exposure to education is, the better (Fergusson et al., 1994). However, adolescence, the stage of life focused on in this study, is the most critical period in a person's life when it comes to developing one's personality and ideals, strongly influenced by their surroundings (peer groups, adult role models). It is in the adolescence stage of life that individuals start experiencing autonomy and begin to make their own decisions (Hernandez Jozefowicz-Simbeni, 2008). There is evidence of the existence of a link between impatience, self-control problems, drop-out rates, and overall low AA (Koch et al., 2015). At times, it may be in this period that students start feeling unsure about future possibilities especially academically, thus educational intervenients need to be aware of these situations so they can act upon them. Studies like this one aim to help those responsible for ensuring academic solid success prevent these situations.

Upper-secondary education is mandatory for academic completion in Portugal, but there are many cases where students abandon school earlier. In 2022, according to INE, the dropout rate in Portugal was 6.5% whereas the European average was 9.6%. Over the last two decades, a decreasing tendency in dropout rates is clear. Despite this, the responsible authorities still need to be active in ensuring these numbers do not increase again.

2.2. PREVIOUSLY RESEARCH ON AA DRIVERS

Academic achievement has long been a subject of discussion within the academic world and so have the drivers affecting it. The students' characteristics, in particular, have been a major focus point in previous literature for understanding AA. Gender is one of those characteristics. It has been shown that female students present lower levels of academic procrastination, leading to higher academic performance and satisfaction per Balkis & Duru (2017). According to these authors, while male students show higher levels of procrastination, this factor does not strongly influence their academic success. However, they make better use of concentration and information processing. On the other hand, girls use their attitude,

motivation, and time management skills more extensively as they show a stronger drive towards school (Ghazvini & Khajehpour, 2011).

The level of academic adjustment, particularly academic integration, strongly shapes one's academic success (Rienties et al., 2012). A student's nationality is an important aspect when studying AA because certain nationalities have more of an aptitude to develop better social skills which are strongly correlated with academic performance (Fernández-Leyva et al., 2021).

Grade retention has also been subject to studies, with some authors concluding that retaining students may represent a decrease in motivation, and self-esteem and may increase the probability of absence from classes (Martin, 2011). However, according to Lorence (2006), who gathered conclusions from other studies, stated that "there is no overwhelming body of scientifically sound evidence that making academically challenged students repeat a grade is ineffective or harmful". In fact, Klapproth et al. (2016) concluded that students who were retained showed better marks than the rest of their peers and that they achieved better academic success than if they had continued on the usual path.

Parents have also been the subject of many studies regarding their children's academic success. The higher the parents' educational status, the more likely a student is to pass (Crede et al., 2015). Higher-educated parents can help lead their children to greater academic success since they have already experienced at least part of the same process. Through their guidance, parents help mold a child's beliefs and views on education, and overall life (Idris et al., 2020). The parents make the most relevant educational decisions made pre-university. Decisions such as whether to attend public or private school or even what area of study the adolescent will study can be influenced by the parents.

In previous studies, another common finding is that the parent's income also affects their children's AA especially when taking into consideration other factors such as education and parental union status (Chevalier et al., 2013). Parents who are strongly involved in their children's lives also tend to lead to higher academic achievement (Machebe et al., 2017). As parents are the first ones to communicate the importance of education to their children, it is normal for them to absorb their way of thinking. Schmuck & Cho (2011) found that most students whose parents help them through tough moments or school problems, live less stressful lives and become more motivated towards their academic activities.

Given that schools are the learning environment students spend the most time on, it is only natural that they have been the subject of several studies. Research conducted by Hahn et al. (2014) in Japan, observed that private school students are more likely to attend university than public school students. It is important to notice that students are randomly assigned to any school regardless of its type. This means that the academic abilities or their families' economic situation is not a factor taken into account, which leads to the conclusion that the environment and conditions created by private schools are essential for students' AA.

Another aspect to consider is the school size. Over the last decades, there have been many studies to uncover whether the size of a school has any effect on students' success (Lee &

Smith, 1997). Most reach the same conclusion: larger schools show a negative relation with AA. Weiss et al. (2010) showed that schools with a range of 600-1000 students are the ones that show the best student engagement which, as mentioned before, can lead to better AA. However, smaller schools (below 600 students) have demonstrated worse levels of academic success (Lee & Smith, 1997).

There is an inverse relationship between class size and AA, meaning the smaller a class is the higher the success of a student (Ehrenberg et al., 2001). With fewer students to manage per teacher, these can devote more time and attention to each student. Some particular groups of students may benefit even more from smaller classes, such as those with disadvantaged backgrounds, special needs students, and those learning the country's main language. This may lead to a heightened sense of self-confidence and belonging which ultimately will help with academic success (Nye et al., 2000).

2.3. MACHINE LEARNING USED FOR AA

Data Mining and Machine Learning (ML) techniques have, time and time again, proven useful in analyzing student's trends and behaviors toward education (GE, 2020). A large number of studies have been made to predict and analyze AA using ML in the last few years which only proves the potential these methods present to the education world.

Before this surge of ML in the educational setting, more traditional ways were used to predict a student's success or what drives said success. Survey-based studies were performed regularly (Caison, 2007). However, papers like Delen (2010) have proven that these surveys are harder to apply to all institutions and may become costly to implement larger surveys. This is why ML models appear to be the solution as they can handle larger amounts of data and extract better insights from them (Cruz-Jesus et al., 2020).

The most common ML models used to predict student success are, among others, Neural Networks (NN), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), K-nearest neighbors (kNN), Logistic Regression (LR), and XGBoost (XGB). Overall, the best-performing models seem to be Random Forests (Asif et al., 2017). Decision Trees are also widely successful, as they are the genesis of Random Forests (Hussain & Khan, 2023). In this study, we attempted to use some of these models and others to obtain the best possible prediction of AA.

However, some of these models have limitations such as their practical interpretability (Costa-Mendes et al., 2021).

Table 1 - Previous use of ML models for AA prediction

Reference	Title	ML methods
(Nunes et al., 2022)	Mathematics and Mother Tongue Academic Achievement: A Machine Learning Approach	NN
(Cruz-Jesus et al., 2020)	Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country	NN, DT, ERT, RF, SVM, KNN, LR
(Costa-Mendes et al., 2021)	A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach	LR, RF, SVM, NN, XGB
(Miguéis et al., 2018)	Early segmentation of students according to their academic performance: a predictive modeling approach	NB, SVM, DT, RF, BT, ABT
(Şen et al., 2012)	Predicting and analyzing secondary education placement-test scores: A data mining approach	NN, SVM, DT, LR
(Asif et al., 2017)	Analyzing undergraduate students' performance using educational data mining	DT, KNN, NB, NN, RF
(Costa-Mendes et al., 2022)	Academic achievement critical factors and the bias and variance decomposition: evidence from high school students' grades	NN
(Hoffait & Schyns, 2017)	Early Detection of University Students with Potential Difficulties	LR, NN, DT, RF
(Marbouti, 2016)	Models for early prediction of at-risk students in a course using standards-based grading	LR, SVM, DT, MLP, NB, KNN
(Costa, 2017)	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses	NN, DT, SVM, NB

3. METHODOLOGY

3.1. DATA

This study was part of a project featuring the collaboration between NOVA IMS and the Directorate-General of Statistics for Education and Science of the Portuguese Ministry of Education (DGEEC). The dataset used in this study consists of 182.060 public high-school students attending the four major areas of the secondary school system in Portugal in 2022/23. This amount of data allows us to reach the best conclusions possible since it includes virtually every high-school student in this system.

Initially, the data was spread across several SQL tables containing information regarding the student's personal information, schedules, evaluations, etc. Besides this, it was also comprised of information about the schools and the teachers. The evaluations extracted were the ones the students achieved in the prior year (2021/22), while the rest of the information is from 2022/23. This will help the prediction be more accurate. After selecting the relevant data from numerous tables, a CSV file was created to hold the students' and school data deemed important to build the models to predict their AA using Python. In Tables 2,3 & 4, we can see the original features in the dataset, with numerical, binary, and categorical variables, with the latter being transformed later in the process.

Table 2 – Original Dataset: Numerical Features

Variable	***Description***
SchoolYear	School year the student will be attending
ClassSize	Size of the student's class
YearsRetained	Number of years the student was previously retained
RatioFemaleStudentsScool	Ratio of female students within the student's school (0-1)
SchoolCode	Code of the student's school
SchoolSize	Size of the student's school
RatioPortugueseGuardianSchool	Ratio of Portuguese guardians within the student's school (0-1)
Grade-PointAverage	Average of all the student's grades
MathPortugueseAverageGrade	Average of the student's Math and Portuguese grades
PortugueseGrade	Student's grade in Portuguese class
HistoryGrade	Student's grade in History class
EnglishGrade	Student's grade in English class
MathematicsGrade	Student's grade in Mathematics class
PhilosophyGrade	Student's grade in Philosophy class
PhysicalEducationGrade	Student's grade in Physical Education class

Table 3 – Original Dataset: Categorical Features

Variable	*Description*
Cycle	Cycle of the student (in this case, high school)
GuardianNationality	Nationality of the student's guardian
Class	Name of the student's class
ASE_Level	Level of financial help provided to the student
City	Name of the student's city
GuardianEducation	Level of education of the student's guardian

Table 4 – Original Dataset: Binary Features

Variables	*Description*
FemaleStudent	If the student is female
PortugueseGuardian	If the student's guardian is Portuguese
Internet	If the student has access to the Internet at home
Computer	If the student has access to a computer at home
Retention	If the student is retained

3.2. TREATMENT OF DATA

3.2.1. DATA TRANSFORMATION

Data transformation is a major step in data analysis since it converts the raw data into relevant information to help with the analysis. By improving the quality of the data available, the accuracy of the predictive models will increase as well as the better interpretability of the results (Ferketich & Verran, 1994). The following variables were the ones that suffered transformations:

Table 5 – Transformations on original variables

Variables	*Transformation/Reason*
Grades	Standardizing the grades to a numeric scale (0-20)
GuardianEducation	Assigning a number to each education level
FinancialSituation	Returns “1” if true and “0” if false (Financial_Situation: “1” if helped)

3.2.2. Feature Engineering

The creation of new variables, a part of feature engineering, is crucial in machine learning as it can refine models' efficiency and learning performance (Wang et al., 2022). New variables can help better capture patterns and relationships between one another and improve the models' interpretability in such a way the old ones cannot. In this work, the following variables were created from the existing ones:

Table 6 – Newly created variables

New Variables	Meaning
LargeSchool	Whether a school is considered large or small (large > 910 students); 1 if large
GuardiansHigherEduc	Whether or not a student's guardian has Upper-Level Education ("1" if True)

3.2.3. Descriptive Analysis

As mentioned above, the original dataset from DGEEC contained information from 182060 students in the Portuguese high school system. In Table 7, we can find the descriptive statistics of some of the variables of the original dataset, after the transformations mentioned in previous sections:

Table 7 – Summary Statistics

	count	mean	min	25%	50%	75%	max
SchoolSize	182060	588.64	1.00	346.00	556.00	782.00	1493.00
PortugueseGrade	152709	10.19	0.00	4.00	11.00	14.00	20.00
ClassSize	178748	23.55	1.00	21.00	24.00	27.00	44.00
PhilosophyGrade	98194	14.11	0.00	12.00	14.00	16.00	20.00
HistoryGrade	86838	7.70	0.00	4.00	5.00	12.00	20.00
Computer	181675	0.57	0.00	0.00	1.00	1.00	1.00
FinancialSituation	182060	0.24	0.00	0.00	0.00	0.00	1.00
FemaleStudent	182060	0.55	0.00	0.00	1.00	1.00	1.00
Mathematics Grade	135990	9.53	0.00	4.00	10.00	14.00	20.00

Portuguese Guardian	182060	0.84	0.00	1.00	1.00	1.00	1.00
PhysicalEducation Grade	151923	12.3	0.00	5.00	15.00	17.00	20.00
Internet	181818	0.64	0.00	0.00	1.00	1.00	1.00
Grade-PointAverage	155542	10.64	0.00	4.36	12.17	14.88	20.00
YearsRetained	182060	0.30	-1.00	0.00	0.00	0.00	10.00
EnglishGrade	142245	10.99	0.00	4.00	12.00	16.00	20.00
RatioPortuguese GuardianSchool	182060	0.83	0.02	0.79	0.85	0.90	1.00
MathPortuguese AverageGrade	154019	10.03	0.00	4.00	11.00	14.00	20.00
RatioFemale StudentsSchool	182060	0.53	0.42	0.51	0.53	0.55	0.69
LargeSchool	182060	1.00	0.00	1.00	1.00	1.00	1.00
Retention	182060	0.16	0.00	0.00	0.00	0.00	1.00

Based on this table, several observations can be made. Out of the 182060 students, 55% are women, 64% have access to the internet at home, but only 57% possess a computer. 24% of the students require financial help from the state. Finally, in 2022/23, 16% of all students ended the year retained.

3.2.4. Data Cleaning

Data cleaning is a vital part of the data analysis process since errors and inconsistencies, such as missing values, can lead to different results when building predictive models and analyzing feature importance (Ridzuan & Wan Zainon, 2019).

To deal with these inconsistencies, several steps were taken. In the first place, variables with over 70% missing values were eliminated. This allowed the analysis to proceed with less noise in the data, which is important due to the large dataset. To handle the remaining missing values, the K-NN Imputer was used. It has been known to be used in works related to predicting AA since it captures patterns and relationships to, more effectively, fill the missing data (Almonteros et al., 2024).

3.2.5. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) addresses instances of class imbalance in datasets. When the minority class is significantly smaller than the majority class, SMOTE proposes oversampling the minority class by creating synthetic examples to balance the class distribution (Chawla et al., 2002). These examples are created by generating new instances between the minority class instance and its nearest minority class neighbors. SMOTE is highly regarded in ML and data mining since it presents a more balanced training dataset thus leading to better generalization and more accurate predictions of the minority class.

In the case of this paper's dataset, a significant discrepancy between the students' approvals and disapprovals can be found. Only 28674 students out of 182060 available did not succeed (approximately 16%). This contrast between the two classes can lead to inferior performances of the predictive models since they might tend to favor the majority class, therefore SMOTE was implemented.

3.3. FEATURE SELECTION

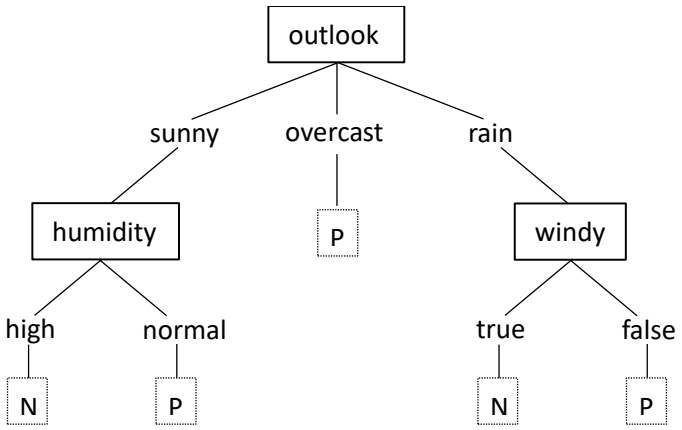
Feature selection is essential in studies like this since it enhances the accuracy, interpretability, and efficiency of the predictive models built after. It involves selecting informative and influential variables, eliminating irrelevant ones, and refining the model's predictive power (Owusu-Adjei et al., 2023). In high-dimensional data, selecting relevant features reduces model complexity, managing computational demands and overfitting risks (Pudjihartono et al., 2022).

Various feature selection methods were used in this study to obtain the best for the models. These methods were Decision Trees, Feature Correlation, Recursive Feature Elimination, and Lasso Regression. After analyzing which features were the most relevant for predicting the target variable for each method, the most important ones were those selected at least three times.

3.3.1. Decision Trees

A decision tree (Breiman et al., 1984) is a supervised machine-learning algorithm that can either produce an accurate classifier or uncover the predictive structure of the problem. This classifier has a tree-like structure where each internal node represents a test on an attribute, each branch represents an outcome, and each leaf node (terminal node) holds a class label.

Figure 1 – Decision Tree



The process of using a decision tree to predict a query instance starts by testing the value of the descriptive feature at the root node of the tree (Kelleher & Mac Namee, 2020). The result of this test determines which of the root node's children the process should then descend to. After that, it repeatedly splits the training data into subsets based on the values of the attributes, based on a metric such as Entropy or Gini Impurity, until the subset at a node all has the same value as the target variable, or when splitting no longer adds value to the predictions.

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

Some academic success prediction studies use Decision Trees since it is easier to understand the results than most models (Costa, 2017; Marbouti, 2016; Strecht et al., 2015).

3.3.2. Correlation-Based Feature Selection

Correlation-based feature selection is used to select highly correlated features with the target variable to identify the ones that contribute the most to the best predictive model. This method also finds redundant information, variables that are highly correlated with one or more other features, which should then be eliminated (Hall, 1999). Besides the aforementioned advantages of feature selection methods, Correlation-based feature selection allows for a simpler and easier interpretation of the importance of the features.

3.3.3. Recursive Feature Elimination (RFE)

This method is used to identify the most relevant features when creating a predictive model. It starts by training a model, ranking the features in terms of importance, and then removing the least important one. Since this is an iterative process, a subset of variables will ultimately be selected and will not impact the model's error if not removed (Ramírez-Hernández & Fernandez, 2007). The RFE method is easy to use and improves the model's predictive

accuracy (Senan et al., 2021). Given that it eliminates less important features, it reduces overfitting, making the models easier to interpret and understand.

3.3.4. Lasso Regression

LASSO (Least Absolute Shrinkage and Selection Operator) regression is a type of linear regression that performs feature selection while also acting as a regularization technique to improve the predictive power of the model it is training by identifying the variable's regression coefficients to minimize the error (Ranstam & Cook, 2018).

LASSO introduces a penalty to the linear regression objective function equal to the absolute value of the magnitude of coefficients which is then multiplied by λ (lambda) to regulate the strength of the penalty.

$$\hat{\beta}(\lambda) = \operatorname{argmin}_{\beta} \left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right) \quad (2)$$

where $\|Y - X\beta\|_2^2 = \sum_{i=0}^{n} (y_i - (X\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^{k} |\beta_j|$, and $\lambda \geq 0$.

3.4. ML Models

3.4.1. Random Forest

Random Forests (Ho, 1995) is one of many ensemble techniques methods as they combine differently built decision trees to improve models' performance. It constructs numerous decision trees during the training of the model and aggregates their predictions to reach final decisions.

To ensure more accurate results, and to minimize the correlation between individual trees, this method, when training each tree with a random subset of data, it only considers a random combination of features at each node split to grow every tree (Breiman, 2001). The randomness employed in this method makes it less prone to overfitting when compared with individual decision trees.

Along with this random feature selection, this model also uses bagging (Bootstrap Aggregating) which is a technique that draws new training subsets, with replacement, from the original dataset to then grow trees using randomly selected features.

3.4.2. Neural Networks

Neural networks (NN) employ both Supervised Learning (SL), utilizing labeled datasets for classification, and Unsupervised Learning (UL), analyzing unlabeled data to uncover patterns like clusters and associations. In early ideas of NN architectures (Mcculloch & Pitts, 1943), these did not learn. In the next few years, newer, simpler NNs trained by SL (Widrow & Hoff, 1962; Narendra et al., 1974) and UL (Grossberg, 1969; Kohonen, 1972) were presented.

Standard artificial neural networks (ANNs) consist of interconnected neurons with real-valued activations, where input neurons are activated by environmental sensors, while connections

from previously active neurons activate others (Schmidhuber, 2015); if a node's output surpasses a specified threshold, it activates, forwarding data to the next layer, yet the model's interpretability is limited due to it being a "black box" of connection weights, and their development often demands extensive data and time.

Many studies addressing student's academic performance have used NN techniques to help predict students' success (Costa, 2017; Hoffait & Schyns, 2017; Marbouti, 2016).

3.4.3. Naïve Bayes Classifier

The naive Bayesian classifier (NBC) provides a simple approach for representing, using, and learning probabilistic knowledge (John, 1995). This method's performance goal is to accurately predict the class of test instances. NBC calculates the probability of an event by multiplying the probabilities of each independent feature given the class and then multiplying the result by the prior probability of the class. NBC was built upon the simple probability theorem known as the Bayes' theorem:

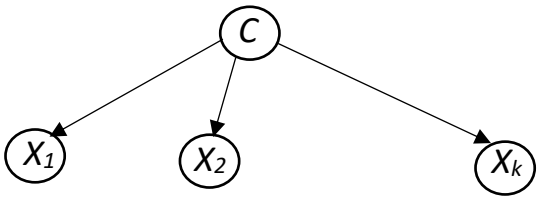
$$P(C = c_k|X = x) = P(C = c_k) * \frac{P(X = x|C = c_k)}{P(x)} \qquad (3)$$

where

$$P(X = x) = \sum_{k'=1}^{eC} P(X = x|C = c_{k'}) * P(C = c_{k'}) \qquad (4)$$

This specialized form of a Bayesian network relies on two important assumptions: it assumes that the predictive attributes are independent of one another given the class (as shown in the image below), hence "naïve"; and that it posits that no hidden or latent attributes influence the prediction process.

Figure 2 – Bayesian Network



Despite having unrealistic assumptions, the resulting classifier known is remarkably successful in practice, often competing with much more sophisticated techniques (Rish, 2001). It has also been proven effective in many practical applications, including text classification, medical diagnosis, and systems performance management (Domingos & Pazzani, 1997; T. Mitchell, 1997). In addition, in previous academic performance studies, NBC has also been tested to predict students' success (Costa, 2017; Marbouti, 2016; Romero et al., 2013; Strecht et al., 2015).

3.4.4. AdaBoost

AdaBoost is a popular method in machine learning where weak learners are combined to form a strong model. Introduced by Freund & Schapire (1997), any machine learning algorithm that accepts weights in the training dataset can be used as the initial weak learner, though decision trees are commonly chosen.

The core idea of AdaBoost is to assign more weight to difficult-to-classify instances and less weight to those already well-classified. Each weak learner, in successive iterations, corrects the errors made by its predecessor. This iterative process focuses on instances misclassified in previous iterations, leading to improved overall predictive performance.

AdaBoost has found widespread application in computer science, financial, and artificial intelligence domains (Alfaro et al., 2008; Hu et al., 2008) due to its ability to enhance weak learners' performance and handle diverse data types. Notably, it excels in handling high-dimensional data and complex classification tasks. Despite its size, the generated classifier typically maintains low test errors, showcasing its robustness and generalization capabilities.

3.4.5. XGBoost

XGBoost (eXtreme Gradient Boosting) first appeared in 2016 by Chen & Guestrin. Based on the Gradient Boosting Decision Tree (GBDT) algorithm, XGBoost operates through an ensemble learning method, iteratively refining weak learners to enhance predictive accuracy. It optimizes a loss function by minimizing residuals in each iteration, utilizing techniques like gradient descent and regularization to prevent overfitting.

There are several benefits of using XGBoost: its scalability allows for efficient handling of large datasets, while its speed enables faster model training and prediction. Besides Academic Achievement, this model has also been successfully used in other domains. In finance, XGBoost has been employed for stock price prediction (Somkunwar et al., 2024). In healthcare, it has facilitated disease diagnosis and prognosis modeling (Jang et al., 2020).

3.4.6. Support Vector Machines

Support Vector Machines (SVM) emerged as a powerful machine learning algorithm in the 1990s, introduced by (Cortes & Vapnik, 1995). SVM is a supervised learning algorithm and is mainly used for classification tasks.

The goal of SVM is to find the optimal hyperplane that best separates different classes in the feature space. This hyperplane is chosen to maximize the margin, i.e., the distance between the hyperplane and the nearest data points from each class, known as support vectors.

One of the key benefits of SVM is their ability to handle high-dimensional data efficiently and effectively. Furthermore, SVM are resilient against overfitting, especially in cases where the number of features exceeds the number of samples.

SVM have been applied across various fields with notable success. In bioinformatics, SVM have been used for protein structure prediction (Huang et al., 2024). In finance, this model has been used to foresee financial trends and bankruptcy prediction (Hamdi et al., 2024). Additionally, SVM have been employed in the creation of climate warning systems to predict future temperatures (Yang & Li, 2024).

4. RESULTS

4.1. MODEL RESULTS

This project was designed to create predictive models to identify students at high risk of failing before the start of the new school year. This paper focuses on the 10th, 11th, and 12th grade students. Each student's grades are relative to their previous year's grades and were used to predict if the student is likely to pass in the following year. Other socio-demographic features were also included in the dataset. This project aims to create a tool that helps teachers and others properly support the most vulnerable students and help policymakers adjust their policies considering the characteristics of the students.

In order to obtain the best possible models, the data from DGEEC was treated by converting every variable existing and created (Table 2-6) into numeric formats for uniformity. After this, we proceeded to remove the columns with more than 70% of missing data. Additionally, students with missing grades for all subjects were removed since the models heavily involved grades to predict the student's academic success. This led to the removal of 26,744 students.

After the data cleaning, the dataset was split between training and testing sets to then create the models, resulting in 80% for training data and 20% for testing. To respond to the class imbalance mentioned before, the dataset used to train the models underwent SMOTE resulting in a balanced dataset with 107,925 of positive and negative classes.

With the dataset ready after the application of SMOTE, feature selection techniques were applied to choose the most appropriate variables in terms of positive influence in the predictive models. The feature selection methods used were Correlation-based selection, Decision Tree Feature Selection, Lasso Regression, and Recursive Feature Elimination using Random Forest. The following features in Table 8 were the ones selected by at least 3 of these methods.

Table 8 – Selected Features

Grade-Point Average
EnglishGrade
YearsRetained
HistoryGrade
MathematicsGrade
PortugueseGrade
City
MathPortugueseAverageGrade
RatioPortugueseGuardianSchool
FemaleStudent
LargeSchool

Once the most relevant features were chosen, the training data was split into train and validation datasets (70%/30%). To make sure each feature contributes equally to the models, a feature scaling mechanism, StandardScaler, was used. Through Z-score normalization, this process divides the value by its mean and then divides it by the standard deviation for each feature. This way, every feature is on a comparable scale, allowing for potentially more reliable models. As mentioned in a previous section, several machine learning models were contemplated, including traditional and more advanced, ensemble methods: Logistic Regression (LR), Decision Trees (DT), Naïve Bayes Classifier (NB), AdaBoost, Random Forests (RF), XGBoost, Support Vector Machines (SVM), and Neural Networks (NN). A comprehensive search for hyperparameters was carried out using a cross-validated grid search. This involved setting up a grid of hyperparameter values and assessing all possible combinations with cross-validation on the training data (see Annex).

To guarantee that the best possible results are achieved, a cross-validation method was used to preserve the models' ability to generalize new, unseen datasets. Stratified K-Fold Cross-Validation assesses models' performance by repeatedly training various subsets of data and then evaluating its performance metrics.

After tuning the models according to the hyperparameters generated by the gridsearch, each one was evaluated on train, validation, and test datasets. The following metrics were applied to assess the models' performance: train accuracy, validation accuracy, validation recall, test accuracy, test recall, and validation misclassification rate. Tables 9 & 10 show the performance metrics for each model given the best hyperparameters.

Table 9 – Model Results for LR, DT, NB, AdaBoost and RF

	LR	**DT**	**NB**	**AdaBoost**	**RF**
Training Accuracy	0.7347	0.9493	0.5390	0.8543	0.9699
Validation Accuracy	0.7348	0.9057	0.5390	0.8542	0.9280
Validation Recall	0.7348	0.9057	0.5390	0.8542	0.9280
Test Accuracy	0.6957	0.8601	0.8710	0.8370	0.8786
Test Recall	0.6957	0.8601	0.8710	0.8370	0.8786
Misclassification Rate	0.2652	0.0943	0.4610	0.1458	0.0720

Table 10 – Model Results for XGBoost, SVM and NN

	XGBoost	SVM	NN
Training Accuracy	0.9622	0.7710	0.8768
Validation Accuracy	0.9285	0.7708	0.8686
Validation Recall	0.9285	0.7708	0.8686
Test Accuracy	0.8808	0.7656	0.8599
Test Recall	0.8808	0.7656	0.8599
Misclassification Rate	0.0715	0.2292	0.1314

The best model was chosen based on its performance of the train, validation, and test sets according to previously mentioned metrics. Additionally, the model's complexity was weighed alongside its performance metrics to certify that the most complete model was picked based on its ability to generalize new data and keep its sophistication.

NB was the worst-performing model compared with the others in this research. NB returned training and validation accuracies of 53.90% despite showing significantly better results in the test dataset with 87.10% accuracy. LR displayed better results with 73.47% and 73.48% for training and validation accuracy, respectively. However, the test accuracy dropped slightly to 69.57%. Given the complex dataset used in this study, NB and LR are not the most appropriate to use since these are linear models.

SVM showed good results with a training and validation accuracy of approximately 77%, while slightly dropping to 76.56% in the test sample. Neural Networks and AdaBoost registered similar test values to those of the better models with 85.99% and 83.70%, respectively. However, these two models returned underwhelming training values, 85.43% for AdaBoost and 87.68% for NN.

Decision Trees, Random Forests, and XGBoost proved to be the strongest models for the data used in this research. DT returned a training and test accuracy of 94.93% and 86.01%, similar to the other two models. However, DT is a simpler model for straightforward analysis and may not be as appropriate for this study, therefore RF's and XGBoost's capacity to capture intricate patterns and relationships in data, especially with an imbalanced dataset, were the final two considered. Their results were very similar: training accuracy of around 96%-97%, validation accuracy of 93%, and test accuracy of 88%.

In order to respond to the class imbalance mentioned before, the dataset used to train the models underwent SMOTE resulting in a balanced dataset with 107,925 of positive and

negative classes. In Table 11, we can see SMOTE applied to the selected models. The Random Forest model returned a training accuracy of 96.99% and a validation accuracy of 92.80%. While using the data without the SMOTE pre-processing, these values instantly decreased with 94.34% training accuracy and 88.45% validation accuracy. On the other hand, XGBoost also registered lower values of accuracy for both training and test without SMOTE (93.94% and 88.52%). This proves the value of SMOTE as a tool to improve the performance of predictive models by balancing the class distribution and, given the large and unbalanced dataset used in this analysis, this method demonstrated its relevance in the Machine Learning scene.

Table 11: RF and XGBoost model with and without SMOTE treatment

Random Forest | XGBoost	***With SMOTE***	***Without SMOTE***
Training Accuracy	0.9699 | 0.9622	0.9434 | 0,9394
Test Accuracy	0.8786| 0.8808	0.8845 | 0.8852

Despite XGBoost responding slightly better to unseen data, it requires more computational resources than RF given its complexity which can be intensive and slower to train. Therefore, further analysis should be conducted to reach a conclusion on which is the best model to predict AA.

4.2. MODEL QUALITY

Several methods were developed to evaluate each model's quality and strength further when predicting a student's outcome in terms of AA in the following year. In the first place, in Figure 3, we can see the ROC curve for the Random Forest and XGBoost models. It is a graphical representation that illustrates how well a binary classifier system, in this case, predicts academic achievement as its discrimination threshold is adjusted. It plots the true positive rate (sensitivity), which indicates the proportion of correctly predicted successful students, against the false positive rate (FRP), representing the proportion of incorrectly predicted high achievers, across various threshold settings.

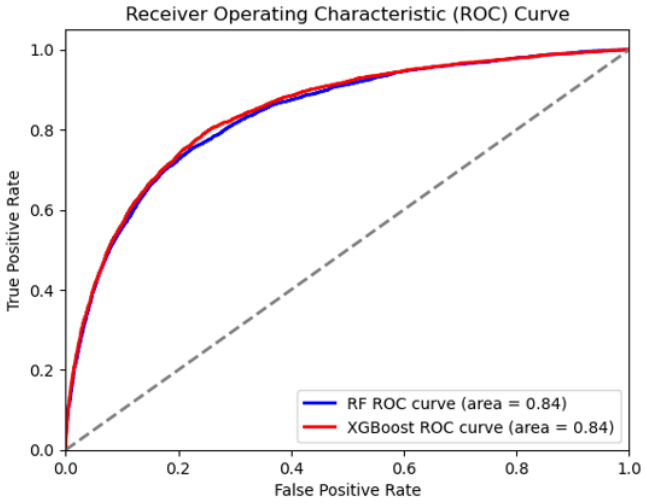
$$Sensitivity = \frac{TP}{TP + FN} = \frac{TP}{P} \qquad (5)$$

$$FRP = \frac{FP}{FP + TN} = 1 - Specificity \qquad (6)$$

$$AUC = \int_0^1 TPR(fpr)d(fpr) \qquad (7)$$

Where: TPR(fpr) represents the Sensitivity as a function of FPR

Figure 3 – ROC and AUC: Random Forest & XGBoost



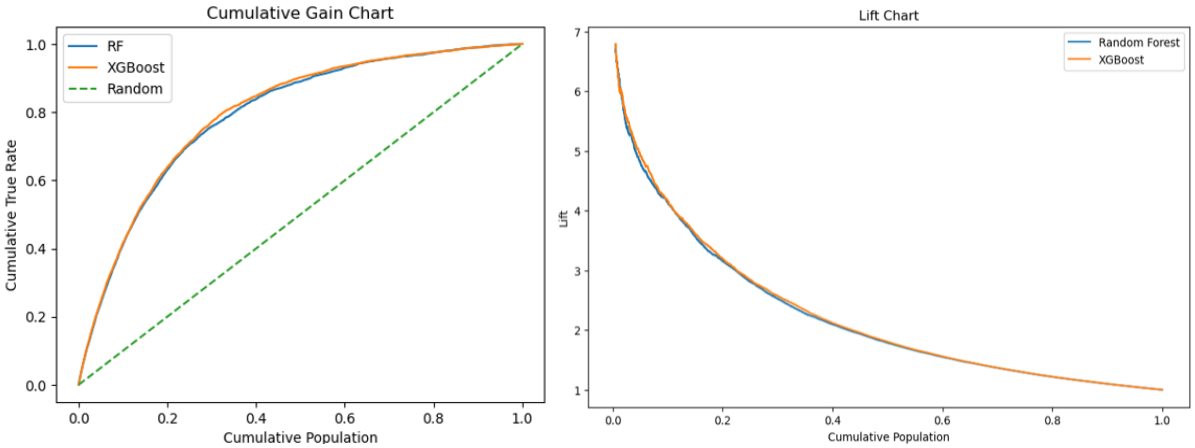
As shown in this graph, the models' AUC (Area Under the Curve) equals 0.84. This means that if we were to randomly select one student who will fail and one who will pass, both models have an 84% chance of assigning a higher probability of failing the student who will actually fail.

In order to assess the model's effectiveness in ranking and identifying students at risk of academic failure, a method called Precision at K was deployed. This method measures the proportion of correctly identified failures among the top K predictions made by the model. In this case, for the RF model, the precision at 10% is 0.5389, meaning that among the top 10% of students predicted to fail, approximately 53.9% do end up failing. As for XGBoost, the precision at 10% is 0.5473. These values suggest both models are fairly good at ranking the most at-risk students early on, thus helping stakeholders prioritize those who are likely to require additional support or intervention. In this case, XGBoost has a slight advantage of around 1 percentage point.

The previous method is dependent on the probability of a student failing, calculated by the model. Therefore, we must understand how accurate these probabilities are. The Brier score helps with this purpose. Lower values indicate better calibration, and, in this case, the RF model returned a Brier score of 0.0892. At the same time, XGBoost displayed a score of 0.0881, demonstrating that both models' predicted probabilities closely reflect the actual outcomes, with XGBoost being slightly better given its lower score.

Finally, to confirm the strength of the models selected, cumulative gain and lift charts (Figure 4) provide a visual representation of the effectiveness compared to random selection. Lift quantifies how much better a model performs compared to a baseline, such as random selection or no model. As we can see, both models show considerably good Lift values which means they effectively identify a larger proportion of failures early on.

Figure 4 – Cumulative Gain and Lift Charts: RF & XGBoost



After the analysis conducted in this section, we can conclude that the XGBoost model provides better predictions than the RF model, despite its elevated computational costs. As it is widely known, educational resources are scarce in comparison to the needs of the students. Therefore, this analysis ensures that the students who are more likely to fail the following year are provided with the most appropriate and higher-quality support.

4.3. FEATURE IMPORTANCE

This section examines feature importance using various AI methods to identify the key features crucial for effectively predicting students' academic outcome. Multiple feature selection methods were used to obtain the features used in the models: Correlation-based Feature Selection, Decision Tree selection, Lasso Regression, and Recursive Feature Elimination via Random Forest Classifier.

The most influential features were identified based on their overall significance in predicting the target variable using permutation feature importance (Altmann et al., 2010). Permutation Feature Importance (PFI) involves shuffling the values of each feature one at a time and measuring the decrease in model performance, in this case, its accuracy. If shuffling a feature's values significantly lowers the model's performance, the feature is considered important. The advantage of this method is that it provides a direct measure of the feature's impact on the model's predictive performance, making it easier to understand the actual contribution of each feature in the context of the model's performance. Unlike the regular Feature Importance method which is specific to tree-based models, like Decision Trees and Random Forest, PFI is model-agnostic, meaning it can be applied to any machine learning model (Fisher et al., 2018). This makes it a versatile tool for feature importance analysis across different types of models, despite being more expensive, computationally speaking. However, this cost is often justified by the more reliable and interpretable results. Another advantage of PFI is that it is more robust to collinearity among features since it takes into account the interactions between all. This comprehensive approach helped to explain the critical elements that drive the model's accuracy and effectiveness, highlighting the key contributors to the predictive performance. In this case, a cross-validated PFI method was used because it provides a robust and reliable way of evaluating the importance of features in an ML model (Kaneko, 2022).

Cross-validation ensures that the model's performance is evaluated on multiple train-test splits, therefore providing a more accurate estimate of the feature importances.

Table 12 shows a ranking of the importance of each feature to the XGBoost model created using Cross-Validated PFI. The number of years a student has been retained proved to be the most significant factor. The importance of demographic factors is evident, particularly the gender of the student and the city they are from. Academic performance indicators such as the mathematics grade, grade-point average, and the average of mathematics and Portuguese grades also showed notable importance. Additionally, specific subject grades like English and History contributed to the model's predictions. Conversely, the size of the school demonstrated minimal relevance.

Table 12: Cross-Validated Permutation Feature Importance for XGBoost model

Rank	**Variable**	**PFI**
#1	YearsRetained	0.087
#2	FemaleStudent	0.082
#3	City	0.079
#4	MathematicsGrade	0.056
#5	Grade-PointAverage	0.043
#6	HistoryGrade	0.035
#7	EnglishGrade	0.033
#8	MathPortugueseAverageGrade	0.029
#9	PortugueseGrade	0.026
#10	RatioPortugueseGuardianSchool	0.012
#11	LargeSchool	0.001

5. DISCUSSION

5.1. DISCUSSION OF FINDINGS

This project aimed to develop predictive models to identify students at high risk of failing before the new school year begins. The study specifically targets the 10th, 11th, and 12th grade students. Each student's performance, measured by their grades from the previous year and socio-demographic variables, was used to predict their likelihood of passing in the upcoming year and to recognize at-risk students. The results from this study confirm the recent trend of deploying AI and ML models to forecast AA over traditional empirical models given the predicting power of ML models (Golino et al., 2014; Musso et al., 2020).

In order to demonstrate this superiority, several ML models were trained and tested to see which one revealed the best accuracy scores for predicting the student's AA. Two models, Random Forest and XGBoost, showed superior predictive performance. Both models achieved high accuracy during training and substantial accuracy in testing. Although Random Forest is more commonly referred to as the superior model to predict AA (Asif et al., 2017; Miguéis et al., 2018), XGBoost was ultimately chosen as the optimal model due to its stronger performance on unseen data and predictions closer to the real values. Unlike the papers mentioned before, this study not only uses this model but reveals it as the best for predicting academic success.

It is of the utmost importance to perform an analysis of the accuracy and quality of the models since it allows the educational stakeholders to better allocate their resources by identifying the groups of students at higher risk of retention. This can enable a more efficient and targeted implementation of strategies to combat retention and ultimately prevent high percentages of high school dropouts.

Following these results of the predictive RF model, the variables used, i.e., drivers of AA, were ranked in terms of importance to the prediction. The most important driver of AA included in the development of the XGBoost model is the number of years a student has been previously retained. The high importance score indicates that altering the retention feature significantly impacts model accuracy, emphasizing its critical role in predicting the target variable which is in line with prior research (Cruz-Jesus et al., 2020). This underscores the importance of implementing early intervention strategies to support students at risk of retention, mainly those with a similar past (Martin, 2011).

The gender of the students proved to be one of the most influential drivers for the predictive model. These results clearly contradict studies like Ebenuwa-Okoh (2010) which states that gender has no significant influence on AA. The current study, however, does not specify which gender demonstrates higher academic performance. Therefore, it is suggested that further studies are made to confirm the positions of previous research claims female students outperform their male counterparts like Sheard (2009). Further analysis could also explore potential factors that might influence the relationship between gender and academic

performance, such as socioeconomic background, study habits, and access to resources, to provide both male and female students with proper academic support.

Previous academic records also showed significant importance in impacting the model's performance, particularly in mathematics and the overall grade-point average (GPA), play a crucial role in predicting academic outcomes. This underscores the importance of prior academic performance as a predictor of student success in the Portuguese educational context, aligning with previous research findings (Etemadpour et al., 2020; van Rooij et al., 2018).

The city of residence also demonstrated significant importance in impacting the model's predictions. This finding underscores the regional differences that influence student outcomes within the study context, emphasizing the need for localized educational strategies. It has been found, through previous research, that AA can be different depending on whether a student attends a school in an urban area or a rural area, with the first registering better results (Lounkaew, 2013). Further studies could be conducted on the data from this study to understand how the region of the school impacts AA.

The size of the school also showed some importance in impacting the model's predictions, although to a lesser extent. This suggests that changes in school size have a minimal impact on model accuracy, indicating that other factors may have a more significant influence on the target variable. While not as influential as other variables, such as city or academic performance metrics, understanding the role of school size contributes to a comprehensive analysis of factors affecting student outcomes.

5.2. THEORETICAL IMPLICATIONS

This study enhances the current understanding of AA prediction models and offers significant insights into the elements affecting the academic performance of Portuguese students transitioning into the 10th, 11th, and 12th grades.

The research performed for this paper confirms the superiority of ML models in predicting AA given its multiple possibilities for analyzing data and building patterns (Nazir et al., 2023). Among the various ML models used to predict the outcome of the student's performance, Random Forest proved one of the most accurate, with results aligning with previous studies thus confirming its effectiveness in academic performance prediction (Ananna et al., 2023; Yağcı, 2022). However, XGBoost also displayed bright results, even displaying higher quality in its predicting ability. Therefore, this study highlights XGBoost as a reliable option for obtaining proper AA predictions.

Previous grade retention has been the subject of many studies mainly due to its negative influence, both on short- and long-term academic performance for male and female students. According to Pagani et al. (2001), students' anxiety and inattentiveness problems not only persisted but also worsened after grade retention, with stronger effects on boys. Despite some studies claiming there is no clear relationship between retention and academic success

(Lorence, 2006), others like Mariano et al. (2024) show no indication that retention benefits students as dropouts seem to increase as well as retention levels. The results from this paper are in line with previous ones since grade retention affects the probability of students failing the following academic year given it compromises the effectiveness of the model.

According to the results from the present study, students' prior grades appear to be one of the strongest predictors of academic achievement. These findings are sustained by studies like Casillas et al. (2012) and Brookshire & Palocsay (2005). The latter further suggests that when mathematics is one of the main features, previous mathematics performance is the strongest predictor for students' success. This perfectly aligns with the results of this study since previous academic performance illustrates the progressive nature of education, highlighting the lasting influence of earlier learning experiences on subsequent achievements.

Gender is one of the drivers of AA that has been exhaustively studied. However, the findings from other papers tend to be contradictory (Yu, 2021). According to Dayioğlu & Türüt-Aşik (2007), female students outperform their male counterparts in terms of GPA, on average. This is the common trend observed across various studies. On the other hand, some studies state that male students have better academic performances than the opposite gender (Joseph et al., 2015). While definitive conclusions cannot be drawn at this stage, one clear finding from this study is that gender significantly influences predicting AA.

Regarding school size, preceding literature shows that smaller schools lead to a decline in student's achievements, particularly in mathematics (Kuziemko, 2006). There is also evidence of a stronger negative effect of larger schools in higher grades, just as schools are larger (Egalite & Kisida, 2016). Further studies should be made to understand this effect given that this paper suggests a lesser importance of school size in predicting the success of students.

This study examines factors affecting academic success among Portuguese students in the 10th, 11th, and 12th grades, offering detailed insights into this specific group. By aligning theoretical models with practical data, the research highlights the key predictors of academic performance and bridges the gap between educational theory and real-world challenges. The findings not only deepen the understanding of academic achievement but also provide valuable information for educators, policymakers, and researchers to improve educational practices and policies in Portugal.

5.3. PRACTICAL IMPLICATIONS

Besides the theoretical implications mentioned above, this study has numerous practical inferences that are highly significant for stakeholders such as teachers, school boards, policymakers, and others. By analyzing the many provided drivers of academic achievement, this research can provide valuable answers that may lead to better support for students in jeopardy of failing based on their age, gender, previously obtained grades, and guardians' background, among others. Given its data-driven approach, this work may encourage educational stakeholders to have a further analytical mindset and use studies like this one to drive their future interventions. For instance, teachers can adapt strategies aimed at

answering the needs of a diverse group of students, while school boards can create programs to address the factors of academic achievement identified through this study. Policymakers could also use these insights to establish education policies that effectively allocate resources to the fields with the greatest need.

Previous retention was one of the main AA drivers found. According to DGEEC, in 2021/22, the retention rate for the four scientific-humanistic courses in Portugal was 8.3%. Despite these values seeing a continuous decrease over the last few years, more can be done to prevent students' failure to succeed academically. By identifying the previously retained students, and others at higher risk of unsuccess through the use of the created models, teachers can provide them with additional academic support through tutoring or even mentorship programs. Moreover, schools can implement personalized plans particularly made to reach each student's specific needs and to closely monitor their progress. Oftentimes, grade retention does not happen solely due to poor reception of the subjects on the part of students. Therefore, schools could extend the guidance from school counselors to offer emotional and psychological support to help students overcome other barriers to academic success.

Mathematics and Portuguese were found to be the most relevant subjects for predicting academic success among high school students. Those with lower grades in these subjects tend to be flagged as "at-risk". Besides the proper assistance the teachers can provide them with, schools can go further with this approach by offering extracurricular activities related to the student's overall weaker subjects, such as book and math clubs. This would encourage the students to develop a deeper interest in these matters, thus potentially helping them achieve greater academic success. Implementing peer tutoring programs can also create a more collaborative learning environment since students who excel in certain subjects help their colleagues achieve better results.

Parental involvement has been deemed essential for the student's academic success as it enhances their motivation, engagement, and overall performance. Given that parents are the first major influence in a child's perspective of life, they can reinforce the importance of learning, and create at home a favorable environment for academic achievement. Schools can cultivate this environment through various strategies like organizing regular parent-teacher meetings to keep parents up to date on the school's curriculum and their children's progress. Furthermore, teachers can maintain open channels of communication with parents to ensure they are aware of their children's regular needs. Beyond the academic aspect, schools can also invite parents to participate in various extracurricular activities or school events, further integrating them into the educational community. This involvement not only aids the students' learning but also strengthens the relationship between families and schools, creating a more supportive environment.

By understanding how the drivers of academic achievement evidenced by this study affect the creation the predictive models and previous ones on the same topic, policymakers can implement data-driven policies, ensuring that the evolving needs of students are met. Despite the relative importance of the model created, the size of schools, and consequently the size

of classes, is a factor to be considered in shaping effective educational environments. Reduced class sizes lead to more individualized attention from the teachers which may ultimately improve at-risk students' chances to succeed. In addition to these measures, diverse classrooms in terms of gender and previous academic achievement can create more dynamic learning opportunities. For instance, since female students are proven to achieve better results than their male counterparts, policymakers can incentivize schools to mix female and previously retained students to address the various needs of different students and promote peer learning.

This study provides several relevant insights that can prove to be essential for educators, school boards, and policymakers in their efforts to reduce retention and dropout rates. Using the information gathered, stakeholders can fashion interventions to support at-risk students and create better, and more inclusive learning environments.

5.4. LIMITATIONS AND FUTURE WORK

Despite returning important insights on the factors impacting AA among high school students in Portugal, several improvements can be made to enhance the quality of this research. Certain machine learning models used for predicting students' success in this study require well-organized and complete data to ensure the effectiveness of said models. One limitation this study presented was the large amount of missing data, especially in the different grade variables. Even though imputing techniques like k-NN were used to fill these gaps in the information, real and updated data will certainly optimize the predictive ability of the ML models. Furthermore, incorporating qualitative methods, such as interviews, and questionnaires, along with quantitative analysis could provide even more insights into the drivers of academic achievement, since it complements the predictive capabilities of machine learning models.

Additionally, the quality of data may also be related to the different features presented which may provide valuable information to help understand drivers of AA. Variables like the parents' income level and involvement can be added to future studies to analyze their impact on students' success. Besides this, in the present study, all Portuguese high school students from the four main areas of study were used to predict their academic progress, future studies could enhance accuracy by segmenting each academic area. Exploring the differences between regions within Portugal could also uncover disparities useful to then tailor effective strategies to specific contexts.

Given the importance of the school's typology (public or private) studied in other papers, it would be relevant to add a variable to confirm or disprove the hypothesis stated above in the Literature Review section, that private schools tend to display higher levels of academic success.

Another interesting approach would also be to understand how much each driver actually affects the outcome. In this paper we only study the effect of each driver on the accuracy of

the XGBoost model, however, it could be interesting to find out how much the variables change the student's outcome.

This paper presents relevant results regarding some of the most important variables that affect academic achievement among high school students in Portugal, however, more can be done. By addressing the limitations and possibilities posed above, new studies can continuously improve the quality of resources provided to help students with a higher chance of retention.

6. CONCLUSIONS

This paper gives the academic community key findings regarding the use of ML and AI to predict academic achievement and improve the quality of the Portuguese education system. ML models showed their superior predictive power, and, in this study, XGBoost proved to be the most effective model to predict the student's AA. This is an uncommon result given the prior research on predicting AA, thus offering yet another, strong alternative to traditional methods. Several factors that influence AA were evaluated in this paper with some revealing interesting findings. Previous retention strongly influences predictions, as students may experience various psychological and educational effects from repeating school years. Gender also proved to be one of the most influential drivers of AA. Further research should be conducted to understand which gender is the most affected and why. Previous academic records were deemed to be relevant for the chosen model since students may experience difficulties in specific subjects, which may hurt their probability of successfully passing the year. Finally, the region where the student attends school was an important variable for predicting AA as well. More densely populated regions may have access to better infrastructures, more qualified educators, and better overall conditions which is why these findings may help overturn this scenario in the long-term. By evaluating these factors, educators and other stakeholders will be able to adapt previous ways of approaching the needs of each student and the overall academic society. Ultimately, the objective of this research is to reduce the dropout rates of high-school students in Portugal.

BIBLIOGRAPHICAL REFERENCES

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decision Support Systems*, *45*(1), 110–122. https://doi.org/10.1016/j.dss.2007.12.002

Almonteros, J. R., Matias, J. B., & Pitao, J. V. S. (2024). Forecasting Students' Success To Graduate Using Predictive Analytics. *International Journal of Computing and Digital Systems*, *15*(1), 713–722. https://doi.org/10.12785/ijcds/150151

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Amrai, K., Motlagh, S. E., Zalani, H. A., & Parhon, H. (2011). The relationship between academic motivation and academic achievement students. *Procedia - Social and Behavioral Sciences*, *15*, 399–402. https://doi.org/10.1016/j.sbspro.2011.03.111

Ananna, F. F., Nowreen, R., Al Jahwari, S. S. R., Costa, E. A., Angeline, L., & Sindiramutty, S. R. (2023). Analysing Influential Factors in Student Academic Achievement: Prediction Modelling and Insight. *International Journal of Emerging Multidisciplinaries: Computer Science & Artificial Intelligence*, *2*(1). https://doi.org/10.54938/ijemdcsai.2023.02.1.254

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Astakhova, K. V, Korobeev, A. I., Prokhorova, V. V, Kolupaev, A. A., Vorotnoy, M. V, & Kucheryavaya, E. R. (2016). International Review of Management and Marketing The Role of Education in Economic and Social Development of the Country. *International Review of Management and Marketing*, *6*(S1), 53–58. http:www.econjournals.com

Balkis, M., & Duru, E. (2017). Gender differences in the relationship between academic procrastination, satifaction with academic life and academic performance. *Electronic Journal of Research in Educational Psychology*, *15*(1), 105–125. https://doi.org/10.14204/ejrep.41.16042

Breiman, L. (2001). *Random Forests* (Vol. 45, pp. 5–32). Kluwer Academic Publishers. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees* (1st ed.). https://doi.org/10.1201/9781315139470

Brookshire, R. G., & Palocsay, S. W. (2005). Factors Contributing to the Success of Undergraduate Business Students in Management Science Courses. In *Decision Sciences*

Journal of Innovative Education (Vol. 3). https://doi.org/10.1111/j.1540-4609.2005.00054.x

Caison, A. L. (2007). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education*, *48*(4), 435–451. https://doi.org/10.1007/s11162-006-9032-5

Casillas, A., Robbins, S., Allen, J., Kuo, Y.-L., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology*, *104*(2), 407–420. https://doi.org/10.1037/a0027180

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16). https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13*, 785–794. https://doi.org/10.1145/2939672.2939785

Chevalier, A., Harmon, C., O' Sullivan, V., & Walker, I. (2013). The impact of parental income and education on the schooling of their children. *IZA Journal of Labor Economics*, *2*(8). https://doi.org/10.1186/2193-8997-2-8

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

Costa, E. B. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

Costa-Mendes, R., Cruz-Jesus, F., Oliveira, T., & Castelli, M. (2022). Academic achievement critical factors and the bias and variance decomposition: evidence from high school students' grades. *Paper Proceedings*, 54–62. http://www.imrjournal.info/2022/ProceedingsEduTeach2022.pdf#page=54

Costa-Mendes, R., Oliveira, T., Castelli, M., & Cruz-Jesus, F. (2021). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies*, *26*(2), 1527–1547. https://doi.org/10.1007/s10639-020-10316-y

Crede, J., Wirthwein, L., McElvany, N., & Steinmayr, R. (2015). Adolescents' academic achievement and life satisfaction: The role of parents' education. *Frontiers in Psychology*, *6*(52). https://doi.org/10.3389/fpsyg.2015.00052

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public

high schools of a European Union country. *Heliyon*, *6*(6). https://doi.org/10.1016/j.heliyon.2020.e04081

Dayioğlu, M., & Türüt-Aşik, S. (2007). Gender differences in academic performance in a large public university in Turkey. *Higher Education*, *53*(2), 255–277. https://doi.org/10.1007/s10734-005-2464-6

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*, 103–130. https://doi.org/10.1023/A:1007413511361

Dumciuviene, D. (2015). The Impact of Education Policy to Country Economic Development. *Procedia - Social and Behavioral Sciences*, *191*, 2427–2436. https://doi.org/10.1016/j.sbspro.2015.04.302

Ebenuwa-Okoh, E. E. (2010). Influence of Age, Financial Status, and Gender on Academic Performance among Undergraduates. *Journal of Psychology*, *1*(2), 99–103. https://doi.org/10.1080/09764224.2010.11885451

Egalite, A. J., & Kisida, B. (2016). School size and student achievement: a longitudinal analysis. *School Effectiveness and School Improvement*, *27*(3), 406–417. https://doi.org/10.1080/09243453.2016.1190385

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Douglas Willms, J. (2001). Class Size and Student Achievement. *Psychological Science in the Public Interest*, *2*(1), 1–30. https://doi.org/10.1111/1529-1006

Etemadpour, R., Zhu, Y., Zhao, Q., Hu, Y., Chen, B., Sharier, M. A., Zheng, S., & Jose, J. G. (2020). Role of absence in academic success: an analysis using visualization tools. *Smart Learning Environments*, *7*(1). https://doi.org/10.1186/s40561-019-0112-3

Evans, J. H., & Burck, H. D. (1992). The Effects of Career Education Interventions on Academic Achievement: A Meta-Analysis. *Journal of Counseling & Development*, *71*(1), 63–68. https://doi.org/10.1002/j.1556-6676.1992.tb02173.x

Fergusson, D. M., Horwood, L. J., & Lynskey, M. T. (1994). A Longitudinal Study of Early Childhood Education and Subsequent Academic Achievement. *Australian Psychologist*, *29*(2), 110–115. https://doi.org/10.1080/00050069408257333

Ferketich, S., & Verran, J. (1994). Focus on Psychometrics An Overview of Data Transformation. *Research in Nursing and Health*, *17*(5), 393–396. https://doi.org/10.1002/nur.4770170510

Fernández-Leyva, C., Tomé-Fernández, M., & Ortiz-Marcos, J. M. (2021). Nationality as an influential variable with regard to the social skills and academic success of immigrant students. *Education Sciences*, *11*(10). https://doi.org/10.3390/educsci11100605

Fisher, A., Rudin, C., & Dominici, F. (2018). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. http://arxiv.org/abs/1801.01489

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

GE, O. (2020). A Machine Learning Based Framework for Predicting Student's Academic Performance. *Physical Science & Biophysics Journal*, *4*(2). https://doi.org/10.23880/psbj-16000145

Ghazvini, S. D., & Khajehpour, M. (2011). Gender differences in factors affecting academic performance of high school students. *Procedia - Social and Behavioral Sciences*, *15*, 1040–1045. https://doi.org/10.1016/j.sbspro.2011.03.236

Golino, H. F., Gomes, C. M. A., & Andrade, D. (2014). Predicting Academic Achievement of High-School Students Using Machine Learning. *Psychology*, *05*(18), 2046–2057. https://doi.org/10.4236/psych.2014.518207

Griffith, M. (2015). The Earliest Greek Systems of Education. In *A Companion to Ancient Education*. Wiley-Blackwell. https://doi.org/https://doi.org/10.1002/9781119023913.ch2

Grossberg, S. (1969). Some Networks That Can Learn, Remember, and Reproduce Any Number of Complicated Space-Time Patterns, I. In *Source: Journal of Mathematics and Mechanics* (Vol. 19, Issue 1).

Hahn, S., Kim, T.-H., & Seo, B. (2014). Effects of Public and Private Schools on Academic Achievement. *Seoul Journal of Economics*, *27*(2), 137–147. http://ssrn.com/abstract=2466238

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning* [The University of Waikato]. https://hdl.handle.net/10289/15043

Hamdi, M., Mestiri, S., & Arbi, A. (2024). Artificial Intelligence Techniques for Bankruptcy Prediction of Tunisian Companies: An Application of Machine Learning and Deep Learning-Based Models. *Journal of Risk and Financial Management*, *17*(4), 132. https://doi.org/10.3390/jrfm17040132

Hernandez Jozefowicz-Simbeni, D. M. (2008). An Ecological Perspective on Dropout Risk Factors in Early Adolescence An Ecological and Developmental Perspective on Dropout

Risk Factors in Early Adolescence: Role of School Social Workers in Dropout Prevention Efforts. *Children & Schools*, *30*(1), 49–62. https://doi.org/10.1093/cs/30.1.49

Ho, T. K. (1995). Random Decision Forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Hu, W., Hu, W., & Maybank, S. (2008). AdaBoost-based algorithm for network intrusion detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *38*(2), 577–583. https://doi.org/10.1109/TSMCB.2007.914695

Huang, J., Osthushenrich, T., MacNamara, A., Mälarstig, A., Brocchetti, S., Bradberry, S., Scarabottolo, L., Ferrada, E., Sosnin, S., Digles, D., Superti-Furga, G., & Ecker, G. F. (2024). ProteoMutaMetrics: machine learning approaches for solute carrier family 6 mutation pathogenicity prediction. *RSC Advances*, *14*(19), 13083–13094. https://doi.org/10.1039/d4ra00748d

Hussain, S., & Khan, M. Q. (2023). Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning. *Annals of Data Science*, *10*(3), 637–655. https://doi.org/10.1007/s40745-021-00341-0

Idris, M., Hussain, S., & Ahmad, N. (2020). Relationship between Parents' Education and their children's Academic Achievement. *Journal of Arts and Social Sciences*, *7*(2), 2020. https://doi.org/10.46662/jass-vol7-iss2-2020(82-92)

Jahnukainen, M. (2001). Social exclusion and dropping out of education. *International Perspectives on Inclusive Education*, *1*, 1–12. https://doi.org/10.1016/s1479-3636(01)80003-9

Jang, S. K., Chang, J. Y., Lee, J. S., Lee, E. J., Kim, Y. H., Han, J. H., Chang, D. Il, Cho, H. J., Cha, J. K., Yu, K. H., Jung, J. M., Ahn, S. H., Kim, D. E., Sohn, S. Il, Lee, J. H., Park, K. P., Kwon, S. U., Kim, J. S., & Kang, D. W. (2020). Reliability and clinical utility of machine learning to predict stroke prognosis: Comparison with logistic regression. In *Journal of Stroke* (Vol. 22, Issue 3, pp. 403–406). Korean Stroke Society. https://doi.org/10.5853/jos.2020.02537

John, G. H. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 338–345. https://doi.org/10.48550/arXiv.1302.4964

Joseph, A., John, O., Eric, I., Yusuf, S., & Olubunmi, A. (2015). Effect of Gender on Students' Academic Performance in Computer Studies in Secondary Schools in New Bussa, Borgu

Local Government of Niger State. *Journal of Education and Practice*, *6*(33). https://eric.ed.gov/?id=EJ1083613

Kaneko, H. (2022). Cross-validated permutation feature importance considering correlation between features. *Analytical Science Advances*, *3*(9–10), 278–287. https://doi.org/10.1002/ansa.202200018

Klapproth, F., Schaltz, P., Brunner, M., Keller, U., Fischbach, A., Ugen, S., & Martin, R. (2016). Short-term and medium-term effects of grade retention in secondary school on academic achievement and psychosocial outcome variables. *Learning and Individual Differences*, *50*, 182–194. https://doi.org/10.1016/j.lindif.2016.08.014

Koch, A., Nafziger, J., & Nielsen, H. S. (2015). Behavioral economics of education. *Journal of Economic Behavior and Organization*, *115*, 3–17. https://doi.org/10.1016/j.jebo.2014.09.005

Kohonen, T. (1972). Correlation Matrix Memories. *IEEE TRANSACTIONS ON COMPUTERS*, *C–21*(4), 353–359. https://doi.org/10.1109/TC.1972.5008975

Kuziemko, I. (2006). Using shocks to school enrollment to estimate the effect of school size on student achievement. *Economics of Education Review*, *25*(1), 63–75. https://doi.org/10.1016/j.econedurev.2004.10.003

Lee, V. E., & Smith, J. B. (1997). High School Size: Which Works Best and for Whom? In *Educational Evaluation and Policy Analysis Fall* (Vol. 19, Issue 3). Lee & Smith. http://eepa.aera.net

Lorence, J. (2006). Retention and academic achievement research revisited from a United States perspective. *International Education Journal*, *7*(5), 731–777. http://iej.com.au

Lounkaew, K. (2013). Explaining urban-rural differences in educational achievement in Thailand: Evidence from PISA literacy data. *Economics of Education Review*, *37*, 213–225. https://doi.org/10.1016/j.econedurev.2013.09.003

Machebe, C. H., Ezegbe, B. N., & Onuoha, J. (2017). The Impact of Parental Level of Income on Students' Academic Performance in High School in Japan. *Universal Journal of Educational Research*, *5*(9), 1614–1620. https://doi.org/10.13189/ujer.2017.050919

Marbouti, F. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers and Education*, *103*, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Mariano, L. T., Martorell, P., & Berglund, T. (2024). The Effects of Grade Retention on High School Outcomes: Evidence from New York City Schools. *Journal of Research on Educational Effectiveness*, 1–31. https://doi.org/10.1080/19345747.2023.2287607

Martin, A. J. (2011). Holding back and holding behind: Grade retention and students' non-academic and academic outcomes. *British Educational Research Journal*, *37*(5), 739–763. https://doi.org/10.1080/01411926.2010.490874

Mcculloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133. https://doi.org/10.1007/BF02478259

Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, *115*, 36–51. https://doi.org/10.1016/j.dss.2018.09.001

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education*, *80*(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

Narendra, K. S., Member, S., & L Thathachar, M. A. (1974). Learning Automata - A Survey. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-4*(4), 323. https://doi.org/10.1109/TSMC.1974.5408453

Nazir, M., Noraziah, A., Rahmah, M., & Sharma, A. (2023). Examining the potential of machine learning for predicting academic achievement: A systematic review. *Fusion: Practice and Applications*, *13*(2), 71–90. https://doi.org/10.54216/FPA.130207

Nunes, C., Beatriz-Afonso, A., Cruz-Jesus, F., Oliveira, T., & Castelli, M. (2022). Mathematics and Mother Tongue Academic Achievement: A Machine Learning Approach. *Emerging Science Journal*, *6*(special issue), 137–149. https://doi.org/10.28991/ESJ-2022-SIED-010

Nye, B., Hedges, L. V, & Konstantopoulos, S. (2000). The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment. *American Educational Research Fournal Spring*, *37*(1), 123–151. https://doi.org/10.3102/00028312037001123

Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., Abdul-Salaam, G., Twum, F., & Abdul-Salaam, G. (2023). *Machine learning prediction accuracy score: The use of Feature selection techniques*. https://doi.org/10.21203/rs.3.rs-1799571/v1

Pagani, L., Tremblay, R. E., Vitaro, F., Boulerice, B., & Mcduff, P. (2001). Effects of grade retention on academic performance and behavioral development. *Development and Psychopathology*, *13*(2), 297–315. https://doi.org/10.1017/S0954579401002061

Pandey, P., & Thapa, K. (2017). Parental influences in academic performance of school going students. *Indian Journal of Positive Psychology*, *8*(2), 132–137. https://www.researchgate.net/profile/Priyanka-Pandey-4/publication/324755784_Parental_Influences_in_Academic_Performance_of_School_

Going_Students/links/5b5098b945851507a7b1a532/Parental-Influences-in-Academic-Performance-of-School-Going-Students.pdf

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. In *Frontiers in Bioinformatics* (Vol. 2). Frontiers Media SA. https://doi.org/10.3389/fbinf.2022.927312

Ramírez-Hernández, J. A., & Fernandez, E. (2007). Enhanced Recursive Feature Elimination. *Proceedings - 6th International Conference on Machine Learning and Applications, ICMLA 2007*, 330–335. https://doi.org/10.1109/ICMLA.2007.35

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895

Ridzuan, F., & Wan Zainon, W. M. N. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, *161*, 731–738. https://doi.org/10.1016/j.procs.2019.11.177

Rienties, B., Beausaert, S., Grohnert, T., Niemantsverdriet, S., & Kommers, P. (2012). Understanding academic performance of international students: The role of ethnicity, academic and social integration. *Higher Education*, *63*(6), 685–700. https://doi.org/10.1007/s10734-011-9468-1

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46. https://www.dors.it/documentazione/testo/201911/10.1.1.330.2788.pdf

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135–146. https://doi.org/10.1002/cae.20456

Schmuck, J., & Cho, S. H. (2011). Parental Influence on Adolescent's Academic Performance. *The Journal of Undergraduate Research*, *9*(11), 77–84. http://openprairie.sdstate.edu/jur/vol9/iss1/11

Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications*, *39*(10), 9468–9476. https://doi.org/10.1016/j.eswa.2012.02.112

Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., Aldhyani, T. H. H., Alqarni, A. A., Alsharif, N., Uddin, M. I., Alahmadi, A. H., Jadhav, M. E., & Alzahrani, M. Y. (2021). Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. *Journal of Healthcare Engineering*, *2021*(1). https://doi.org/10.1155/2021/1004767

Sheard, M. (2009). Hardiness commitment, gender, and age differentiate university academic performance. *British Journal of Educational Psychology*, *79*(1), 189–204. https://doi.org/10.1348/000709908X304406

Somkunwar, R. K., Pimpalkar, A., & Srivastava, V. (2024). A Novel Approach for Accurate Stock Market Forecasting by Integrating ARIMA and XGBoost. *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science*. https://doi.org/10.1109/SCEECS61402.2024.10481891

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*. https://eric.ed.gov/?id=ED560769

Tilak, J. B. G. (2002). Education and Poverty. *Journal of Human Development*, *3*(2), 191–207. https://doi.org/10.1080/14649880220147301

UNESCO. (2008). *Inclusive education: the way of the future; conclusions and recommendations of the 48th session of the International Conference on Education (ICE)*. https://unesdoc.unesco.org/ark:/48223/pf0000180629

van Rooij, E. C. M., Jansen, E. P. W. A., & van de Grift, W. J. C. M. (2018). First-year university students' academic success: the importance of academic adjustment. *European Journal of Psychology of Education*, *33*(4), 749–767. https://doi.org/10.1007/s10212-017-0347-8

Wang, Z., Xia, L., Yuan, H., Srinivasan, R. S., & Song, X. (2022). Principles, research status, and prospects of feature engineering for data-driven building energy prediction: A comprehensive review. *Journal of Building Engineering*, *58*. https://doi.org/10.1016/j.jobe.2022.105028

Weiss, C. C., Carolan, B. V., & Baker-Smith, E. C. (2010). Big school, small school: (Re)testing assumptions about high school size, school engagement and mathematics achievement. *Journal of Youth and Adolescence*, *39*(2), 163–176. https://doi.org/10.1007/s10964-009-9402-3

Widrow, B., & Hoff, M. (1962). Associative Storage and Retrieval of Digital Information in Networksof Adaptive "Neurons." In *Biological Prototypes and Synthetic Systems* (Vol. 1). https://doi.org/10.1007/978-1-4684-1716-6_25

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z

Yang, J., & Li, Z. (2024). Construction of a Climate Early Warning System: Predicting Future Temperatures and Climate Security Using BiLSTM. *Frontiers in Computing and Intelligent Systems*, *7*(2), 11–20. https://doi.org/10.54097/zscep661

Yu, Z. (2021). The effects of gender, educational level, and personality on online learning outcomes during the COVID-19 pandemic. *International Journal of Educational Technology in Higher Education*, *18*(14). https://doi.org/10.1186/s41239-021-00252-3

APPENDIX A

Hyperparameters tested for each model (XGBoost chosen hyperparameters highlighted)

Model	Hyperparameters	Definition	Values Tested
LR	C	The inverse of regularization strength; smaller values specify stronger regularization	[0.1, 1, 10]
	Penalty	Specifies the norm of the penalty	['l1', 'l2']
DT	Max_Depth	The maximum depth of the tree; If None, then nodes are expanded until all leaves are pure	[None, 10, 20]
	min_samples_split	The minimum number of samples required to split an internal node	[2, 5, 10]
NB	-	Naïve Bayes has no hyperparameters	-
AdaBoost	n_estimators	The maximum number of estimators at which boosting is terminated	[50, 100, 200]
	learning_rate	Weight applied to each classifier at each boosting iteration	[0.01, 0.1, 0.2]
RF	n_estimators	The number of trees in the forest.	[100, 200, 300]
	max_depth	The maximum depth of the tree; If None, then nodes are expanded until all leaves are pure	[None, 10, 15, 20]
	min_samples_split	The minimum number of samples required to be at a leaf node.	[2, 5, 10]
	min_samples_leaf	The minimum number of samples required to be at a leaf node	[1, 2, 4, 8]
	max_features	The number of features to consider when looking for the best split	[‘auto’, ‘sqrt’, ‘log2’]
XGBoost	n_estimators	Number of boosting rounds (the number of trees to fit)	[100, 200, **300**]
	max_depth	Maximum depth of each tree in the boosting process	[3, 6, **9**]
	learning_rate	Step size shrinkage	[0.01, 0.1, **0.2**]
SVM	C	Regularization parameter. The strength of the regularization is inversely proportional to C	[0.1, 1, 10]
	kernel	Specifies the kernel type to be used in the algorithm.	['linear', 'rbf']
NN	hidden_layer_sizes	Size of the hidden layers in the neural network	[(50,50), (100,100)]
	max_iter	Maximum number of iterations	[500, 1000]
	activation	Activation function for the hidden layer	['relu', 'tanh']

APPENDIX B

This is to certify that
Project No.: **DSCI2024-7-151396**
Project Title: **Using Machine Learning Models to predict high school student's Academic Achievement**
Principal Researcher: **Afonso Quintino**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 7/15/2024.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of
- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 7/15/2024

NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa