

Machine Learning Methods for University Student Performance Prediction in Basic Skills based on Psychometric Profile

Glender Brás*, Samara Leal, Breno Sousa, Gabriel Paes, Cleberson Junior, João Souza, Rafael Assis, Tamires Marques, Thiago Teles Calazans Silva

UNA University Center, Belo Horizonte, 30640-070, Brazil

ARTICLE INFO

Article history:

Received: 15 March, 2025

Revised: 10 July, 2025

Accepted: 13 July, 2025

Online: 31 July, 2025

Keywords:

Psychometric Profile Assessment

Machine Learning

Student Performance

ABSTRACT

Ensuring the quality of higher education in Brazil presents a complex challenge, intensified by factors that directly affect students' academic performance. The pervasive influence of social media and the overconsumption of superficial digital content undermine students' ability to engage in deep comprehension, critical thinking, and the practical application of knowledge. Furthermore, inadequate preparation during the preceding educational years hinders students' ability to adapt to the academic demands of higher education, leading to difficulties in academic progression and increased dropout rates. In view of the above, this paper explores the potential of Machine Learning models (ML) in predicting the academic performance of higher education students within the *Ânima Educação* ecosystem, Brazil. The contribution of this work is the development of an artificial intelligence-based assessment tool called AILA that recommends personalized study content for fundamental skills such as Portuguese and Mathematics, based on the psychometric profile of each student. This approach aims to optimize the learning process by addressing individual needs, enhancing academic performance, and overcoming the challenges faced by students in the contemporary educational landscape. Psychometric profile data were collected from approximately 41,296 incoming students of the *Ânima Educação* universities on the following dimensions: learning, social intelligence, emotional management, socio-emotional skills, teaching method, and knowledge area of the students. The AILA ML models presented good results in predicting students' basic skills performance in the binary and regression approaches. Specifically, the CatBoost model showed an accuracy of 0.74 in predicting scores on the Portuguese and Mathematics and Logical Reasoning proficiency tests.

1. Introduction

This paper is an extension of the paper originally presented at the 2024 *IEEE 12th International Conference on Intelligent Systems* [1]. The 2024 Map of Higher Education in Brazil, published by the Semesp Institute, reveals that more than a half of university students (57.2%) drop out before completing their courses [2]. A potential contributor to the high student dropout rate in Brazil's higher education system is the deficient quality of public basic education, largely due to insufficient government investment in the sector. Furthermore, the growing influence of social media and the excessive consumption of superficial digital content have contributed to a reduced capacity for deep comprehension, critical analysis, and the practical application of knowledge. In [3], the authors argued that excessive daily internet use can lead to family conflicts, impaired communication, superficial relationships, learning difficulties, anxiety disorders, and attention deficits. Consequently, students who

enter university lacking essential competencies may struggle to maintain academic progress, ultimately leading some to withdraw.

A practical illustration of the substandard quality of education in Brazil is provided by the nation's performance in the Program for International Student Assessment (PISA), a study administered by the OECD (Organization for Economic Cooperation and Development). PISA evaluates the knowledge of 15-year-old students in the domains of reading, mathematics, and science. In the PISA 2022 study, which assessed 81 countries, Brazil ranked among the 20 lowest-performing nations in two of the three evaluated subjects [4]. Moreover, the country scored below the OECD average across all domains. The results indicate that approximately 50% of Brazilian students failed to reach the minimum proficiency level in the assessed areas.

Insufficiency in basic skills represents one of the main challenges faced by university students, directly impacting their academic per-

*Corresponding Author: Glender Brás, UNA University Center, Belo Horizonte, 30640-070, Brazil, glender.medeiros@ulife.com.br

formance. According to [5], difficulties in academic writing and text interpretation compromise the ability of students to organize ideas, write coherent texts, and adequately express their knowledge, thus hindering their performance in assessments and academic assignments. Similarly, the study of [6] reveals that gaps in fundamental mathematical skills, such as fractions and problem-solving, can hinder academic progress, especially in courses that require quantitative reasoning. These limitations often necessitate remedial coursework, which extends the duration of studies and contributes to higher dropout rates.

Furthermore, the basic skills play a critical role in the development of higher-order cognitive abilities, such as analytical thinking and problem-solving [7]. The absence of these fundamental skills can significantly impair a student's ability to engage with more complex academic content and tasks. Without a solid foundation in essential areas like literacy and numeracy, students often struggle to make connections between different concepts, analyze problems critically, and apply learned knowledge to real-world situations. As a result, their ability to develop academic autonomy is severely hindered. This lack of autonomy in turn affects their capacity to manage learning independently, which is vital for success in higher education. Students who are unable to independently navigate through academic challenges often rely heavily on external support, such as remedial courses or additional tutoring, which further extends their time in the education system.

In this way, the lack of basic skills not only directly impacts academic performance but also prevents students from developing the necessary self-regulation and problem-solving strategies that are essential for lifelong learning and success in the professional world. Consequently, addressing these deficiencies early in a student's academic career is essential for fostering both academic independence and long-term educational success [8].

Given this, a number of studies have explored the application of machine learning (ML) as a strategy to address the aforementioned issues. Some of the studies involve the use of data analysis and prediction techniques to assess students' academic performance and psychometric profiles. Psychometric profiling is a way to understand an individual's psychological and behavioral characteristics. This is done through psychometric tests that measure the cognitive abilities, personality traits, motivation, interests, and attitudes and can be used to understand how a person thinks, learns, and behaves in different situations [8].

Contemporary psychometrics has shown a significant potential in enhancing psychological assessments, particularly through the integration of ML algorithms [9]. These algorithms possess the ability to improve their performance over time, learning from new data and experiences. This capability allows the inclusion of probabilistic relationships within computer programs, enabling more nuanced and accurate predictions of individual behavior and characteristics. Unlike traditional educational technology tools that rely on predefined, rigidly programmed rules to process inputs and generate outputs, ML models offer a dynamic and adaptive approach. They can continuously evolve based on incoming data, providing greater flexibility and responsiveness in the analysis. This adaptability is a key advantage, as it enables psychometric tools to stay relevant and effective as new patterns and insights emerge, making them far more powerful in real-world applications where data and conditions

are constantly changing.

In [10], for instance, a hybrid regression model was proposed to enhance the accuracy of predicting student grades in various subjects. Additionally, an optimized multi-label classifier was developed to qualitatively predict the factors that influence student performance. The model employs three dynamic weighting techniques: collaborative filtering, fuzzy set rules, and Lasso linear regression. This integration of techniques enables a more flexible and adaptable analysis of the variables that impact learning.

Despite significant advances in recent literature on student performance prediction, several limitations persist, including challenges related to imbalanced datasets, unreliable data sources, and concerns regarding the transparency and quality of artificial intelligence (AI) models [11, 12]. These limitations pose substantial challenges in the practical application of predictive tools within real-world educational environments. Specifically, they hinder the ability to generate accurate and reliable insights into student performance, which are essential for formulating evidence-based learning strategies. In the absence of high-quality data and robust model transparency, the reliability of predictions is compromised, making it difficult for educators to make informed decisions. Consequently, this undermines the effectiveness of personalized learning interventions and hampers the creation of adaptive educational strategies that can cater to the diverse needs of students. Overcoming these challenges is crucial for ensuring that predictive analytics can be used to meaningfully enhance educational outcomes and improve the overall learning experience.

In view of the above, this paper presents the AILA (Artificial Intelligence for Learning Assistance) project, that aims to study and implement an AI-based algorithm that utilizes multiple ML models to evaluate students' performances and suggest appropriate academic content to assist the students of *Ânima Educação*. This tool enhances the accuracy of student performance predictions by integrating diverse data sources, enabling more effective management of unbalanced data sets and improving the transparency of its recommendations.

AILA was developed to enable a more accurate diagnosis of learning gaps, thereby supporting personalized interventions aimed at improving student performance. Leveraging machine learning models, AILA generates individualized learning plans, ensuring targeted support aligned with each student's specific needs. This innovative tool aligns with the institutional and academic objectives of higher education institutions by enhancing retention rates, minimizing the need for remedial instruction, and promoting a more efficient academic trajectory. Ultimately, the algorithm offers a data-driven approach to learning, fostering continuous improvement in student outcomes and advancing educational quality.

The case study of AILA was implemented among incoming students at *Ânima Educação* Group, with the objective of providing personalized recommendations based on each student's psychometric profile. The *Ânima Educação* Group is a prominent private educational organization in Brazil, operating 25 educational brands and managing over 500 educational centers nationwide, with a student population of approximately 400,000.

The data collection process for the mapping phase of this study occurred in two distinct methods. First, a series of questionnaires were answered by university students who have recently entered

their courses at one of Ânima's higher education institutions. The Likert scale [13], a prevalent instrument in questionnaire design, was used to employ a five-point scale ranging from "never" to "always." This scale is commonly utilized in the examination of attitudes, beliefs, and behaviors. The questionnaires are organized into three dimensions of knowledge: socio-demographic, socio-cognitive, and socio-emotional. In addition, the students completed Portuguese and Logical Reasoning tests.

These data were pre-processed and, based on the quantitative average of the scores obtained in the questionnaires/tests, the students are mapped as Naive, Beginner, Apprentice or Advanced and receive a recommendation based on this taxonomy. After this mapping, the ML models are then used to predict the students' scores in the Portuguese Language and Logical Reasoning diagnostic tests. This prediction is done considering three approaches:

- **Binary Classification:** The prediction is whether the students achieved high performance (1) or low performance (0) in these tests, based on the average scores of the population and;
- **Multi-class Classification:** The target variable is the classes of the taxonomy.
- **Regression:** The models predict the students' scores;

The case study demonstrates the effectiveness of AILA's machine learning models in predicting academic performance by integrating students' psychometric variables, thereby enhancing predictive accuracy. This approach underscores the importance of incorporating AI into teaching methodologies to support both students and educators. By identifying students' strengths, weaknesses, and potential academic difficulties in advance, AILA enables the provision of targeted resources to mitigate challenges and foster academic success

A total of 41,296 students completed the aforementioned questionnaires between September 2023 and October 2024 via a custom-developed web application. In addition to data collection, this application features a user-friendly interface that presents the learning content recommended by the models in a clear and accessible manner. The following ML models were employed: CatBoost, Decision Tree (DT), Random Forest (RF), XGBoost, Support Vector Classifier (SVC), and Support Vector Regressor (SVR), all of which demonstrated strong performance across regression, binary, and multi-class classification tasks. For example, the CatBoost model achieved an accuracy of 0.74 in predicting proficiency scores in Portuguese and logical-mathematical assessments using a binary classification approach. In contrast, under a multi-class configuration, the XGBoost and Decision Tree Classifier yielded better results. A comprehensive analysis and discussion of these findings are presented in Section 5.

The rest of this paper is organized as follows: Section 2 presents a summary of the literature, containing works that explore the use of ML models to assess student performance. In Section 3 concepts and methods relevant to this research are discussed. Section 4 explains the processes carried out to test the models and provides a comparative analysis of their performance. Finally, Section 6 presents our conclusions about the study.

2. Related Works

According to [14], predicting students' academic performance has become an increasingly complex task due to the growing volume and variety of data within educational systems. The authors argue that prevailing predictive methods remain insufficient for accurately identifying the most appropriate techniques to assess student performance in higher education institutions. Additionally, the identification of factors influencing student performance remains an underexplored area requiring further investigation, highlighting the need to determine which variables exert the most significant impact on academic outcomes.

Given this, a number of studies have been conducted in the literature to explore the use of ML models as a strategy to assess student performance in various domains. Among these studies, some are particularly noteworthy due to their relevance to the current research and the significant contributions they have made. In [14], the authors conducted a comprehensive review of 162 studies that utilized ML techniques to predict student performance between 2010 and 2022. The study of [15] proposes an intelligent system based on ML to predict students' academic performance, taking into account factors such as attendance, grades, and participation in activities. Algorithms such as Random Forest and Support Vector Machines (SVM), which have been shown to be effective in analyzing academic data, were used. The model developed showed an accuracy of 85%, standing out for its ability to accurately predict performance, with great potential for personalizing pedagogical interventions and optimizing educational outcomes. A key point observed in that research is that the appropriate choice of variables (features) can significantly influence the quality of the predictions.

The use of deep neural networks (DNN) to assess the quality of English language teaching is explored in [16], offering a more effective alternative to traditional methods. With an accuracy rate of 97%, the model is able to process large amounts of data and capture the semantic nuances present in texts, facilitating evaluation in a scalable and less subjective way. The research demonstrates how automating the feature extraction process can reduce cost and time, while improving the accuracy and consistency of scores, bringing an innovative solution to the field of language teaching.

The study of [17] utilizes ML algorithms to identify low-engagement students in a social science course at the Open University (OU) and assess how engagement affects performance. The analysis included variables such as education level, assessment scores, and interactions with virtual learning environment (VLE) activities. Several ML models, including decision trees and gradient-boosted classifiers, were tested, with the best performance in accuracy and recall. A dashboard was developed to help instructors monitor student engagement and provide timely interventions, further exploring the relationship between engagement and course assessment scores.

In [18], the use of ML models is proposed to predict the development of university students' skills over the course of their studies. By analyzing performance data in assessments and extracurricular activities, the authors were able to identify patterns that allow them to predict the evolution of students' cognitive and socio-emotional skills. The results show that deep learning models are effective, achieving 90% accuracy, and provide an agile way to tailor pedagogical approaches to students' individual needs.

The application of ML in the development of flexible learning environments is examined in [19], highlighting its potential to enable new forms of personalized instruction. The study shows how ML models can be used to adapt the content and pace of teaching to the needs of each student, resulting in greater engagement and better academic outcomes. In addition, automating the assessment process helps eliminate human bias, promoting a fairer and more accurate way of assessing students. Based on data collected in real time, the study suggests how curricula and pedagogical strategies can be continuously adjusted to promote more inclusive learning. [20] presents a comprehensive analysis of the literature on how ML has been applied to identify characteristics that affect students' academic performance. The review of 84 publications found that academic and demographic variables, such as grade history and attendance, are the most commonly studied. The study indicates that, although existing models yield satisfactory performance, incorporating additional factors—such as family dynamics and students' psychological characteristics—could enhance predictive accuracy. Moreover, it emphasizes that expanding educational databases is essential to optimizing personalized interventions.

A detailed review of the use of ML in online education is presented in [21], with a particular focus on enhancing student skill acquisition. The study shows that techniques such as content personalization, automatic correction, and progress prediction have been effective in optimizing learning. However, the authors also highlight important challenges, such as privacy issues and model accuracy, and suggest that more research should be done to overcome these limitations. Collaboration between educators, researchers and platform developers is also seen as essential to maximize the positive impact of ML in education. In [22], the authors reviews the main applications of AI and ML in digital education, covering topics such as intelligent tutors, dropout prediction, adaptive learning and process automation. The work shows that artificial neural networks and SVM are the most widely used algorithms, with an emphasis on predictive models aimed at preventing dropout and improving student performance.

The study [23] explores ML application to predict the development of university students' basic skills throughout their course. Using algorithms such as RF and SVM, the research analyzed academic performance data and practical activities to identify the students most likely to succeed or struggle. The ML model proved effective in identifying patterns, allowing for faster and more personalized interventions, which could be crucial in optimizing students' learning and academic development.

In [24], the author focus is on predicting which students are at risk of dropping out of courses on online learning platforms such as MOOCs and Learning Management Systems (LMS). Using machine learning and deep learning algorithms, the study analyzed variables such as performance in assessments, engagement, and online behavior to identify students at risk at an early stage. The model, based on the RF algorithm, achieved excellent results in terms of precision and recall, demonstrating how engagement data can significantly improve the effectiveness of predictions. This allows for timely intervention to prevent students from dropping out and improve their academic performance.

The work of [25] investigates the use of deep neural networks (DNN) to predict the academic performance of students in a data

structures course. The model achieved 89% accuracy when using the SMOTE oversampling technique based on students' grades from previous courses. In addition, DNN outperformed other ML algorithms, such as SVM and RFs, on several performance metrics. The research suggests that the model could be a valuable tool for identifying at-risk students early in the semester, enabling early intervention to improve academic outcomes.

Thus, several studies have been proposed to use ML models to improve the design of educational systems and create a more personalized and effective learning experience. However, challenges persist regarding feature selection, dataset size and balance, and the explanatory power and reliability of these models. Furthermore, there is a lack of studies that have applied the results of these models in real-world settings and presented the models' predictions and recommendations through a user-friendly interface.

3. Main Concepts

This section provides an overview of the main concepts and methods used in this study, aiming to make the reading smoother and the understanding more accessible. The goal is to clarify how ML techniques can be applied to assess academic performance, contributing to promoting student success. Section 3.1 outlines the fundamental concepts underlying student performance analysis and their application within the scope of this study. Section 3.2, in turn, presents and defines the ML models employed to predict deficiencies in essential academic skills.

3.1. Elements for Student Performance Analysis

As defined by [26], performance can be understood as the way someone or something acts or behaves, measured by its output. In the educational context, student performance refers to the assessment of students based on criteria that consider essential competencies for the current scenario [27]. In this study, performance analysis is conducted through three main approaches: the Likert Scale [13], the Item Response Theory (IRT) [28], and the principles of Psychometrics [8].

The Likert Scale is a widely used instrument for assessing perceptions and preferences, and it is classified as a summative assessment method [13]. It offers response options arranged in a progressive order, typically ranging from strong disagreement to strong agreement. In the context of this study, the Likert Scale is employed to measure an individual's self-perceived proficiency in a given skill, using a five-point scale with the following categories: "never," "rarely," "sometimes," "often," and "always".

The concept of psychometrics can be approached in various ways, one of which is its application in assessing an individual's psychological traits [29]. In this sense, psychometrics involves the development of measurement tools, such as tests, scales, and questionnaires, to perform a precise and valid analysis of different aspects of human behavior [8]. Moreover, psychometrics extends beyond simple measurement by emphasizing the quality and accuracy of assessment instruments. Its primary aim is to ensure that these instruments yield consistent results while effectively capturing the constructs they are intended to measure [30]. Guided by these

principles, this study applied psychometric concepts to construct the profile of the analyzed students, as detailed in Section 4.

3.2. Machine Learning (ML)

The field of machine learning focuses on developing algorithms capable of learning directly from data rather than following predefined commands [31]. The primary goal is to build computational systems that, when fed with a dataset, can generate models to make predictions, classifications, or identifications based on the acquired knowledge. Within this context, we applied various ML approaches to address the proposed problem, including Decision Tree, Random Forest, Neural Network, and Support Vector Machine. The ML techniques used in this study were:

- **Decision Tree (DT):** According to [32], a Decision Tree [33] is a hierarchical, branched structure composed of nodes and branches. At each internal node, a decision is made based on a test applied to input variables, guiding the flow along specific branches. The terminal nodes, or leaf nodes, provide the predicted values of the target variable or the associated probability distributions. According to [34], the Decision Tree, known for its quick understanding and ease of implementation, is often adopted in decision support systems in the healthcare field. Its versatility makes it applicable in various domains, offering benefits in terms of efficiency and simplified operation. With the ability to generate clear and easy-to-understand analyses, DT establishes itself as a valuable tool to optimize decision-making processes in different contexts, standing out for its accessibility and effectiveness in various areas of knowledge.
- **Random Forest (RF):** The Random Forest method, as described by [33], consists of a set of classifications based on decision trees, where each tree is influenced by a random vector. This vector is generated independently and follows a uniform distribution among the trees in the forest. Various approaches can be applied to construct these vectors, including bagging, estimated selection of splits, output randomization, and estimated attribute selection. The fundamental principle of this model lies in the independence of the generated vectors for each tree individually. By gathering a large number of trees and combining their decisions, the technique aims to increase accuracy in data classification for ML problems. This is achieved because, after building the trees, the final prediction is determined based on the most voted class by the ensemble [35].
- **Support Vector Machine (SVM):** Introduced by [36], SVM is a ML technique for binary classification problems. Its approach involves transforming input vectors through a non-linear mapping into a high-dimensional space, where a linear decision boundary is constructed with specific characteristics that ensure good generalization capability. Originally, this model was developed to handle perfectly separable datasets but was later improved to accommodate scenarios where data exhibits overlap and is not completely separable.
- **XGBoost:** As detailed by [37], XGBoost is a ML model based on decision trees, designed to maximize computational efficiency and predictive accuracy. This method is distinguished by using "boosting," an approach where each subsequent tree seeks to correct the errors of the previous ones. XGBoost differs from conventional methods by integrating regularization mechanisms that minimize the risk of overfitting, applying, among other strategies, data partitioning and parallelism in tree construction. Its ability to scale processing of large data volumes makes it a popular choice for a wide range of predictive problems. Additionally, the model enhances classification accuracy by coordinating the integration of multiple trees, focusing on reducing bias and variance. According to [37], XGBoost was developed to be an effective, versatile, and easy-to-implement tool, making it a predominant choice in competitions and practical applications.
- **CatBoost:** It is an ML algorithm that stands out for its ability to optimize performance in decision tree-based models, especially when dealing with categorical data. The algorithm's main innovation is the use of Ordered Target Statistics, a method that improves the encoding of categorical variables and thus reduces the risk of overfitting by preventing information from leaking improperly during training [38]. Additionally, CatBoost implements a symmetric boosting approach, which ensures greater training efficiency and enhances the model's ability to generalize to new data. These advancements make CatBoost an effective solution for classification and regression problems, particularly useful in contexts with large data volumes and challenging tasks, such as those in finance and marketing sectors [39].

4. Proposed Modeling

In this section, we present the models implemented for understanding the relationship between the psychometric profiles of the students and their performance in basic skills. The primary objective of this study is to develop the ML models, which aim to assess both the knowledge of the Portuguese Language and Mathematical Reasoning. The implementation of these methods will facilitate the determination of the relationship between different psychometric domains and the teaching and learning of these fundamental subjects, which can support the students with the recommendation of relevant content to improve their abilities. The information was modeled based on three output configurations:

- **Binary Classification:** categorizes students into two groups: "high performance" (1) or "low performance" (0).
- **Multi-class classification:** classifies students based on the four classes or taxonomies: Naive, Beginner, Apprentice, and Advanced, with Naive having the lowest scores and Advanced the highest.
- **Regression:** the regression analysis, in turn, is used to predict the scores in the assessments, showing the improvement of the students. In addition, logistic regression, based on the idea

of 'propensity scores', is used to identify the correlation between student grades and other factors, helping to understand the elements that influence their grades [40].

The development of the predictive models was guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [41], ensuring a structured, systematic, and replicable approach. The use of CRISP-DM enabled a logical progression from data understanding to model implementation, ensuring that each step effectively contributed to the quality and accuracy of the predictions. Figure 1 shows the process of this methodology.

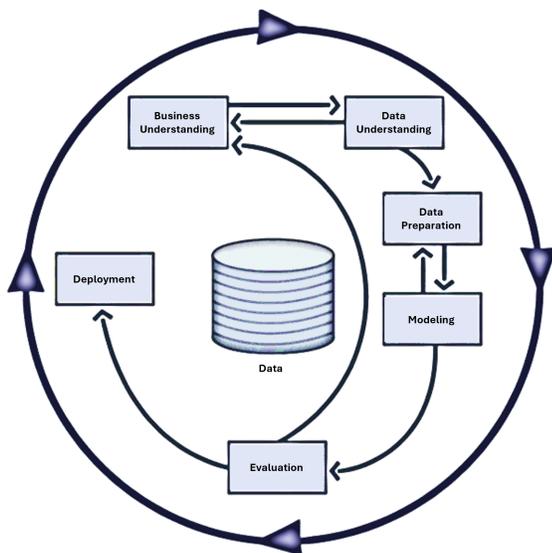


Figure 1: Phases of CRISP-DM applied to this study.

The application of CRISP-DM began with **business and data understanding**, during which the project objectives were defined and the characteristics of the data collected from student questionnaires were analyzed. These questionnaires — comprising emotional states, cognitive behaviors, and academic routines — underwent a rigorous preparation process conducted by educators from Ânima, which was essential to ensuring data quality and reliability.

During the **data preparation** phase, standardization and encoding techniques were applied to ensure data consistency and compatibility, optimizing it for subsequent modeling. This careful preparation was essential for generating a robust dataset that served as the foundation for building effective predictive models.

Following the data preparation stage, the process advanced to the **modeling** and **evaluation** phases, during the ML models were trained, fine-tuned, and rigorously assessed for performance. Finally, in the **deployment** phase, AILA's models and their artifacts were structured to ensure reusability and seamless integration into production environments. A web-based application was developed to make the models' personalized learning recommendations accessible to students in a user-friendly and practical manner, thereby completing the CRISP-DM cycle and establishing this study as a practical contribution to real-world educational contexts.

4.1. Data Understanding and Preparation

As previously mentioned, the data used to train and test the ML models were collected through psychometric questionnaires designed to assess students' emotional states, cognitive behaviors, and academic routines, along with diagnostic tests measuring proficiency in fundamental skills (Portuguese and Logical Reasoning). These instruments were administered to incoming students at Ânima Educação institutions, as illustrated in Figure 2.

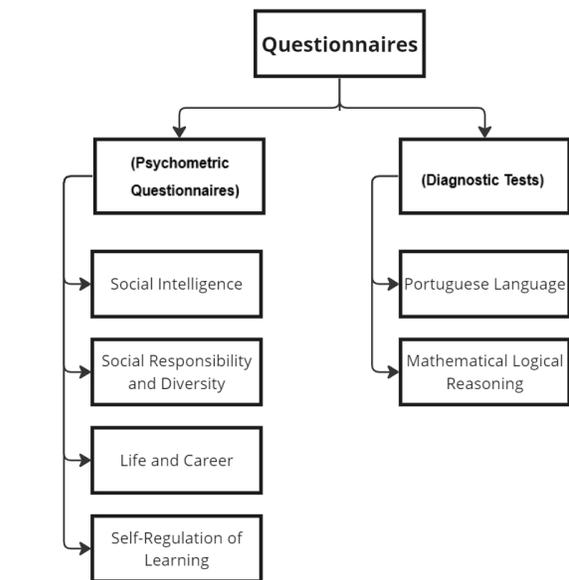


Figure 2: Questionnaire Flowchart

It is important to note that the aforementioned data were processed in accordance with Brazil's General Data Protection Law (Lei Geral de Proteção de Dados Pessoais – LGPD), which safeguards privacy and ensures information security. Participation in the study was limited to students who provided full consent, reinforcing ethical research practices and legal compliance established by Ânima Educação institutions. Although participation was voluntary, several measures were taken to mitigate self-report bias and preserve the diversity and representativeness of the sample. These included the use of validated psychometric instruments with clear, neutrally worded items; the assurance of anonymity and confidentiality to reduce social desirability effects; the incorporation of consistency checks across similar items; and the inclusion of control questions to identify inattentive responses. Moreover, behavioral frequency-based questions were prioritized over abstract self-assessments. Prior to responding, students were clearly informed about the purpose of the study and were encouraged to answer honestly.

All data collected in this study were securely stored on AWS infrastructure, through an account managed by Ânima. Access to both the source code and the MongoDB database was strictly limited to the project team, in compliance with privacy and data protection requirements. The application and database are hosted on an EC2 instance, ensuring controlled and secure access aligned with industry-standard data security practices.

The psychometric questionnaires were constructed using the Likert Scale, discussed in Section 3.1, and subjected to a statistical



Figure 3: ML Workflow

analysis to aim to validate their effectiveness in mapping the correct profiles based on population analysis. The diagnostics were developed by the specialized professionals in pedagogical intervention.

4.2. Modeling

Figure 3 provides an overview of the ML workflow designed for this study. This modeling framework encapsulates the main stages of the process, from the collection of psychometric and diagnostic data to the training and evaluation of predictive models. This representation provides a clear and systematic overview of the process, enhancing methodological transparency and supporting the reproducibility of the study.

4.2.1. Data Pre-processing

The preliminary processing of data is of the utmost importance for the construction of predictive models, as it ensures the quality of the information. To this end, we employed data derived from questionnaires described in Section 4.1. To ensure data standardization and compatibility, we implemented variable normalization and coding techniques.

We applied the numerical data to undergo *MinMaxScaler* normalization, which scales the values between 0 and 1. This normalization enhances model stability by mitigating distortions caused by features with varying scales and units, such as the number of clicks on VLE activities compared to assessment scores [42]. By standardizing the numerical data, the model can more effectively learn relevant patterns without being disproportionately influenced by any single feature.

For categorical variables, the *OneHotEncoder* method was used, converting each category into binary representations. This approach ensures that the model does not assign any hierarchical or ordinal relationships between categories, treating each one independently and without bias [42]. By encoding variables such as different VLE activities (e.g., forums, resources) separately, the model can better capture the impact of each activity on student engagement. These preprocessing techniques enhance the quality of the data and improve the ability of the ML models to accurately predict student engagement.

4.2.2. Used Models

A comparative analysis was conducted to evaluate the efficacy of several models to predict student's performance in this study. The analysis revealed that XGBoost exhibited superior performance in terms of efficiency in decision trees and the capacity to process

substantial volumes of data. CatBoost also demonstrated satisfactory results, particularly in scenarios involving disorganized data. Additionally, simple decision trees were utilized for comparative purposes, along with Random Forest, which was noteworthy for its stability and predictive capacity. For complex data, SVM was used, as it performs well on many variables [43].

4.2.3. Parameter Adjustment and Validation

The models were adjusted using the method *GridSearchCV*, an exhaustive search of various combinations of adjustments, with the aim of identifying the best option for prediction. To prevent the model from learning too much from the training data alone, k-fold cross-validation was used, ensuring a good forecast on new data. Studies demonstrate that this practice improves the prediction of academic performance by reducing statistical errors [44].

4.2.4. Evaluation Metrics

Accuracy is a widely used metric for evaluating the performance of a specific model, reflecting the ratio of correct predictions to total observations [45]. This metric can be used to further evaluate a model by measuring the ratio of correctly predicted positive cases to total predicted positive cases. This metric is advantageous in situations with high costs of false positive results [46].

The classification and regression methods were assessed using various methodologies, enabling a more comprehensive analysis [47]. Accuracy, which represents "hits," performs optimally with balanced data, while the area under the ROC curve (AUC) is more suitable for unbalanced data [45]. Additionally, the accuracy of positive predictions and the hit rate on positives were evaluated, which are fundamental in academic settings [46]. In the context of regression models, the mean squared error (MSE) was employed as a simplification, prioritizing significant errors while making predictions in a linear fashion [48]. The root mean squared error (RMSE) is the average error of the model expressed in the variable.

5. Case Study

This section details the case study developed to test the validity of the proposed methodologies, contextualizing the study and its results. The study involved 41,296 incoming students from various Ânima Educaç o institutions. The primary objective was to assist students in overcoming their challenges from the very beginning of their higher education journey, thereby optimizing their academic trajectory right from the outset. To achieve this, the selection of the

student sample for the algorithms was based on the availability of complete individual data. Specifically, only students with comprehensive information from all the applied tests were included in the sample.

The distribution of binary classes reveals a significant imbalance in both Portuguese and Math scores. In both subjects, the vast majority of students scored below the classification threshold, while a considerably smaller fraction achieved or exceeded this mark. This class imbalance demands attention in the development of predictive models, as it can negatively impact model performance and generalization capabilities (Figure 4).

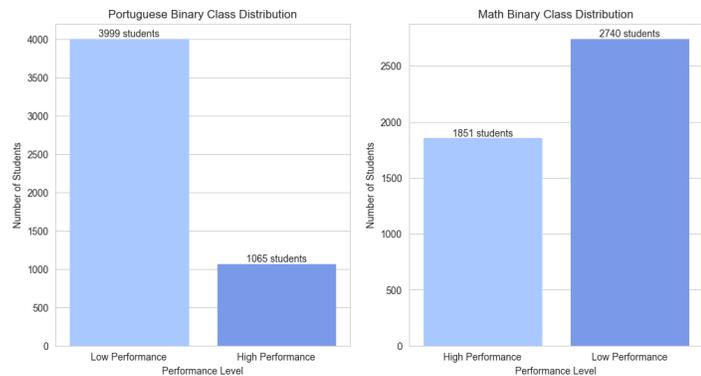


Figure 4: Class distribution Binary Classification

Furthermore, the distribution of learning taxonomies presents interesting patterns. In Portuguese, the "Beginner" category shows the highest concentration of students, while the "Advanced" category is the least represented. In Math, a similar pattern is observed, with the "Beginner" category predominating and "Advanced" being the least common. This uneven distribution among learning categories suggests a need for differentiated pedagogical approaches to address the specific needs of each group (Figure 5).

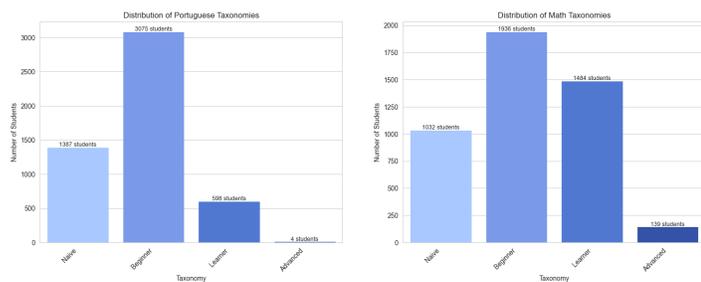


Figure 5: Class distribution Multi-class Classification

This approach was essential, as the supervised learning algorithms used in the study require a full set of labeled data to effectively learn and make accurate predictions. Incomplete data could lead to biased models or reduced prediction accuracy, making it crucial to ensure that only students with complete data were considered. By focusing on students with full datasets, the study aimed to maximize the reliability and validity of the predictions generated by the ML algorithms.

5.1. Experiments

To carry out the experiments, a process was structured that involved analyzing the students' answers to a series of questions, with the aim of measuring different aspects of learning. From these answers, a consolidated data set was generated in a CSV file, containing the features shown in Table 1. The implementation of the AILA algorithm was carried out in the Visual Studio Code environment, using the Jupyter extension to facilitate interactive code execution. The language chosen was Python, due to its wide range of specialized ML libraries.

Table 1: Features and Outputs

| Feature | Output |
|--------------------|------------------------------|
| Modality | Logical/Math Reasoning Tax. |
| Knowledge Area | Portuguese Lang. Tax. |
| Tax. Learning | Score Logical/Math Reasoning |
| Tax. Soc. Intell. | Score Portuguese Lang. |
| Tax. Life/Career | - |
| Tax. Emot. Mgmt. | - |
| Tax. Soc. Resp. | - |
| Score Learning | - |
| Score Soc. Intell. | - |
| Score Life/Career | - |
| Score Emot. Mgmt. | - |
| Score Soc. Resp. | - |

Initially, the data was subjected to a preparation pipeline, which included removing missing values and transforming the variables. To avoid bias in the models, incomplete entries were eliminated, resulting in 4,499 lines. Subsequently, the numerical attributes were normalized using *MinMaxScaler*, ensuring that all the variables were on the same scale, in the 0 to 1 range. *OneHotEncoder* was used to transform categorical variables into numerical representations suitable for machine learning algorithms. The least relevant variables, based on their predictive importance, were gradually removed by Recursive Feature Elimination (RFE), reducing the dimensionality of the data set and improving the performance of the models.

This study used the approach of splitting the data into training and test sets with the help of scikit-learn's train-test-split function. This technique makes it possible to split the data randomly, so that a fraction of it is used to train the model, while the other fraction is used to evaluate its performance.

After preparing the data, the model's hyperparameters were optimized using *GridSearchCV*, a method that systematically goes through a grid of predefined values to find the best combination of parameters. In the experiment, cross-validation was applied with three divisions (*cv=3*), and the accuracy metric was used as the evaluation criterion. The optimized set of hyperparameters was then selected, and the final model was adjusted based on this configuration, ensuring better predictive performance.

5.2. Results

The experiments were carried out with the aim of predicting the academic performance of the university's students and showed that

the models based on boosted decision trees (Boosting) and Random Forest obtained the best results, especially in binary classification. The selection of attributes included variables related to the dimensions of learning, social intelligence, emotional management, and other socio-emotional skills, as well as the teaching method and area of knowledge of the students.

The study also highlighted the influence of teaching methods and students' subject knowledge on academic performance. Certain pedagogical strategies appeared to promote better learning outcomes, while discipline-specific factors also played a role in shaping students' outcomes. These findings suggest that personalized interventions, tailored to students' academic and socio-emotional profiles, could be instrumental in improving educational outcomes. Future research could explore the impact of these interventions and further refine predictive models to support data-driven decision-making in educational settings.

5.2.1. Binary Classification

The *CatBoostClassifier* and *RandomForestClassifier* models performed better in predicting students' proficiency in Portuguese and logical and mathematical reasoning. In predicting Portuguese, the *CatBoost* model obtained the best results, while for logical and mathematical reasoning, the *RandomForest* model performed better, as shown in Table 2.

In addition, it was observed that the models showed high metrics for class 0, which encompasses the students with the greatest difficulty in the subjects assessed. This result is particularly relevant, as it indicates that the models are able to more accurately identify the students who need the most attention and pedagogical support. The high precision and recall for class 0 reinforce the ability of these approaches to correctly discriminate between students with difficulties, making them useful tools for targeted educational interventions.

The joint analysis of the confusion matrices from the *CatBoostClassifier*, shown in Figure 6, complements these findings. For Portuguese Language, there is a good performance in identifying students in class 0, but lower accuracy in identifying those in class 1. In the case of Logical and Mathematical Reasoning, the model produced more balanced results between the classes. This combined visualization highlights the models' focus on correctly identifying students with low performance, which is the central objective of this study.

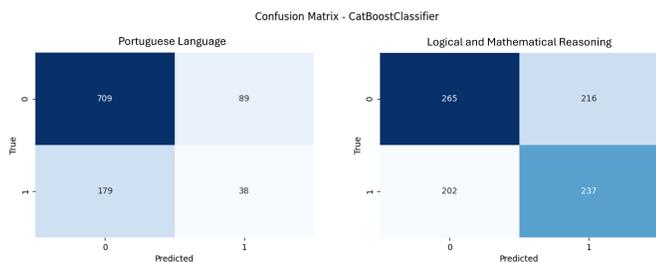


Figure 6: Confusion Matrix - CatBoostClassifier

The superior performance of the Boosting and Random Forest models can be justified by the fact that these techniques employ ensemble learning, reducing bias and variance [49]. *CatBoost*, in

particular, uses effective processing of categorical variables and reduces the impact of overfitting, while Random Forest benefits from the aggregation of multiple decision trees, promoting robustness to the model [38].

Table 2: Binary Classification Results. PT = Portuguese language; LR = Logical and Mathematical Reasoning. Acc = Accuracy; Prec 0 = Precision for class 0; Prec 1 = Precision for class 1; F1 = F1-Score (weighted).

| Model | Type | Acc | Prec 0 | Prec 1 | F1 |
|--------------|------|------|--------|--------|------|
| CatBoost | PT | 0,74 | 0,80 | 0,30 | 0,22 |
| XGB | PT | 0,71 | 0,80 | 0,26 | 0,22 |
| DecisionTree | PT | 0,53 | 0,82 | 0,25 | 0,35 |
| RandomForest | PT | 0,71 | 0,80 | 0,26 | 0,28 |
| SVC | PT | 0,53 | 0,82 | 0,25 | 0,35 |
| CatBoost | LR | 0,55 | 0,57 | 0,52 | 0,53 |
| XGB | LR | 0,52 | 0,55 | 0,50 | 0,51 |
| DecisionTree | LR | 0,56 | 0,57 | 0,54 | 0,53 |
| RandomForest | LR | 0,56 | 0,57 | 0,54 | 0,53 |
| SVC | LR | 0,55 | 0,56 | 0,53 | 0,51 |

5.2.2. Multi-class Classification

With regard to multi-class classification, the *XGBoost* and *DecisionTreeClassifier* models showed the best results for Portuguese language and logical and mathematical reasoning, respectively (Table 3). These results suggest that although Boosting models remain effective, the complexity of predicting multiple classes may have affected overall accuracy.

Table 3: Multi-class Classification Results. PT = Portuguese language; LR = Logical and Mathematical Reasoning. Acc = Accuracy; Prec = Macro-averaged Precision; F1 = Macro-averaged F1-Score.

| Model | Type | Acc | Prec | F1 |
|--------------|------|------|------|------|
| CatBoost | PT | 0,47 | 0,48 | 0,47 |
| XGB | PT | 0,56 | 0,48 | 0,49 |
| DecisionTree | PT | 0,56 | 0,49 | 0,51 |
| RandomForest | PT | 0,43 | 0,48 | 0,45 |
| SVC | PT | 0,35 | 0,47 | 0,38 |
| CatBoost | LR | 0,37 | 0,36 | 0,36 |
| XGB | LR | 0,40 | 0,38 | 0,38 |
| DecisionTree | LR | 0,42 | 0,40 | 0,39 |
| RandomForest | LR | 0,37 | 0,37 | 0,37 |
| SVC | LR | 0,35 | 0,37 | 0,35 |

As shown in Figure 5, the used dataset has a much larger number of students with low performance than students with high performance. This can bias the model's learning, especially when it involves multi-class classification, and make it difficult to identify different levels among the students. For this reason the multi-class models found a low accuracy, as shown in Table 3.

Despite these limitations, the multi-class classification models still provide complementary insights regarding the distribution of student performance levels. The scarcity of examples in intermediate and high-performance classes hinders the models' ability to learn discriminative patterns for these categories, contributing to the

overall lower accuracy. For the model to better classify student performance levels through the multi-class approach, more data would be needed in the other classes. However, since the critical task of identifying students with unsatisfactory performance is handled by the binary classification model—more robust and less affected by class imbalance—the limitations of the multi-class model do not compromise the practical objectives of this study. Thus, multi-class results should be viewed as a secondary analytical tool, useful for exploratory interpretation, while pedagogical decision-making is grounded on the binary model's outputs.

5.2.3. Regression

Although the *CatBoostRegressor* and *XGBRegressor* achieved the best results for Portuguese, and the *RandomForestRegressor* for logical and mathematical reasoning (Table 4), the regression models generally exhibited low predictive power. This limitation may be related to the fact that the domain scores used as predictors are not necessarily strong indicators of performance in another domain, as each one assesses distinct cognitive skills. Therefore, attempting to predict performance in a specific area based on performance in others may fail to adequately capture the unique characteristics of each evaluated competency.

Table 4: Regression Model Performance. PT = Portuguese language; LR = Logical and Mathematical Reasoning. RMSE = Root Mean Squared Error; MSE = Mean Squared Error; MAE = Mean Absolute Error; R² = Coefficient of Determination.

| Model | Type | RMSE | MSE | MAE | R ² |
|--------------|------|------|------|------|----------------|
| CatBoost | PT | 0.16 | 0.03 | 0.13 | 0.05 |
| XGB | PT | 0.16 | 0.03 | 0.14 | 0.04 |
| DecisionTree | PT | 0.17 | 0.03 | 0.14 | -0.02 |
| RandomForest | PT | 0.16 | 0.03 | 0.14 | 0.04 |
| SVR | PT | 0.17 | 0.03 | 0.13 | 0.03 |
| CatBoost | LR | 0.25 | 0.06 | 0.22 | 0.01 |
| XGB | LR | 0.25 | 0.06 | 0.22 | 0.00 |
| DecisionTree | LR | 0.25 | 0.06 | 0.22 | -0.04 |
| RandomForest | LR | 0.25 | 0.06 | 0.22 | 0.01 |
| SVR | LR | 0.25 | 0.06 | 0.22 | -0.02 |

5.3. Discussion of the Results

The results obtained confirm that the Boosting and Random Forest models are highly effective for binary classification problems, corroborating previous studies that highlight their ability to capture complex patterns and reduce overfitting through regularization [50] [37].

The analysis of variable importance in the CatBoost model, used to predict students' performance in the Portuguese language, reveals which features had the greatest influence on the predictions. Figure 7 presents a bar chart ranking the most relevant features based on their importance values.

It is observed that scores related to life and career and learning had the greatest impact, followed by social intelligence and emotional management. These results suggest that socioemotional skills play a crucial role in students' performance, not only in content mastery but also in their ability to apply such knowledge in the exam.

Additionally, categorical variables such as the field of study (Humanities, Biological and Health Sciences, among others) and the mode of instruction (in-person or not) also showed some influence, although to a lesser extent, in predicting the results. This analysis reinforces the need for educational policies that consider not only technical knowledge but also interpersonal and emotional skills, which directly impact students' academic success in the Portuguese language exam.

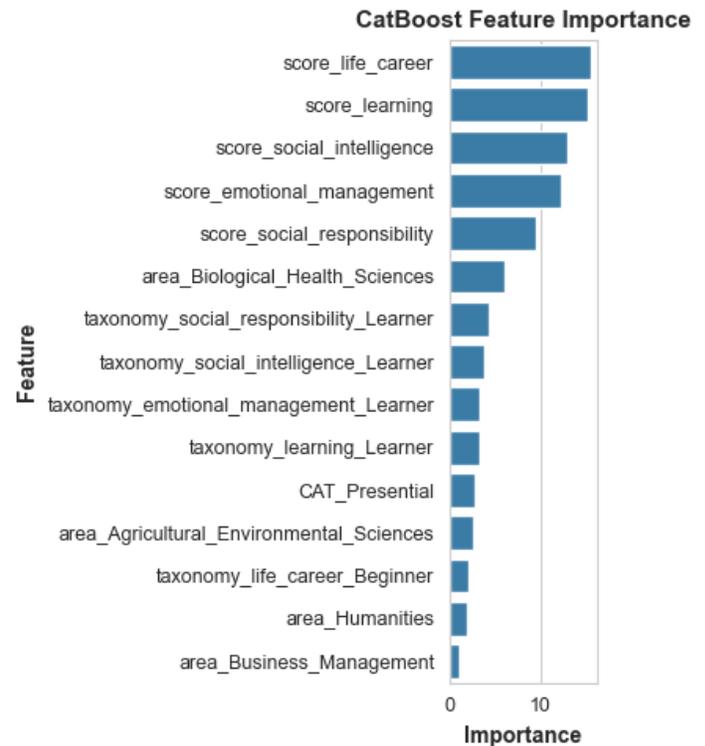


Figure 7: Feature Importances – CatBoost Model

However, multiclass prediction proved to be more challenging, possibly due to the imbalance in the data, which can guide the models learning to a good accuracy in one class but not in the others. Regarding regression, the relatively low R² values suggest that other factors, such as individual aspects of the students and unmeasured external factors, may have influenced academic performance.

5.4. Practical application of the recommendation

AILA employs the responses of students to psychometric questionnaires and diagnostic tests to categorize them according to a taxonomy comprising four levels: The designation of Naive, Beginner, Apprentice or Advanced is utilized to categorize individuals based on their level of expertise or proficiency in a particular domain. In accordance with the classification that has been presented, the platform provides recommendations that are customized to align with each student's unique profile.

Figure 8 presents a simulated example of the AILA's recommendation interface, which displays content suggestions categorized by dimension (e.g., logical reasoning, social intelligence, life and career). The recommendations presented are based on the student's

level in each domain, with the objective of strengthening specific skills.

The practical implementation in question establishes a connection between psychometric assessment and pedagogical guidance, promoting a more student-centered learning experience.

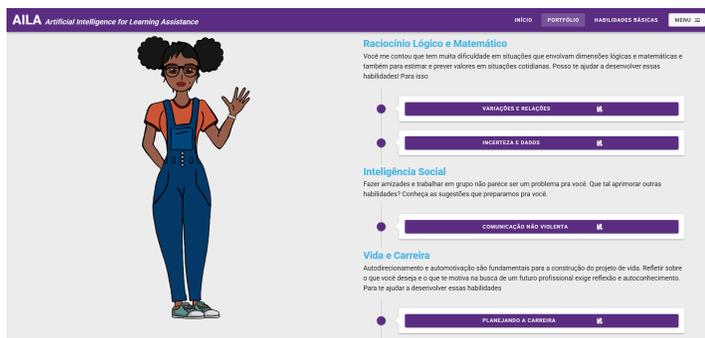


Figure 8: Simulated interface displaying personalized recommendations generated by AILA based on psychometric classification

6. Conclusion

This work explored ML models to university students performance in basic skills such as Portuguese language and Logical Reasoning. The study has been applied in a real case involving more than 40,000 students at several institutions of Ânima Educação, an educational group in Brazil.

The results reinforce the effectiveness of using ML to predict the academic performance of higher education students and show that the models implemented are effective for predicting student performance in basic skills, especially for predicting students who will have difficulty, which is the main objective of this study. The high accuracy in class zero, in the case of binary classification, reinforces this point and can provide support to intervention actions on the part of the university. Similarly, the regression models demonstrated effectiveness in predicting the student's score and can also serve as an important method for this purpose.

As for the multi-class classification task, the models presented a low accuracy and showed that the proposed models are still not sufficient to find several levels among the students satisfactory. This is justified by the imbalance of data in the used dataset and the low amount of data in classes that represent high student performance, that is, the low number of students with high performance in basic skills in the research carried out. This confirms that data balance is crucial for this type of task.

In addition, the models proved to be effective in finding a correlation between psychometric profile and performance in basic subjects, which suggests that emotional and psychosocial factors can influence the learning process of students. For future work, it will be important to conduct new tests with more advanced students in order to improve the performance of the multi-class model, thus generating advances in the learning of the model for these cases.

However, a central limitation of this study lies in the pronounced imbalance in class distribution, particularly in the context of multi-class classification. The scarcity of students with high performance in both Portuguese Language and Logical-Mathematical Reasoning restricted the models' ability to learn representative patterns across

all performance levels. This imbalance contributed to the lower accuracy observed in the multi-class models and compromised the regression results by reducing the diversity and richness of the training data. Moreover, the reliance on correlated data, such as psychometric profiles based on self-reported responses, may introduce potential bias, as subjective answers can reflect inconsistent or distorted perceptions. Future work should address these limitations through the collection of more balanced datasets, the application of techniques to mitigate self-report bias, and the use of data augmentation and resampling methods, aiming to improve the models' generalization capacity and robustness.

References

- [1] G. Brás, S. S. Leal, B. Sousa, C. Junior, J. Souza, Z. Mendes, G. Paes, "Machine Learning Models for Basic Skills Identification in University Students," in 2024 IEEE 12th International Conference on Intelligent Systems (IS), 1–9, IEEE, 2024, doi:10.1109/IS61756.2024.10705207.
- [2] I. Semesp, "Mapa do Ensino Superior no Brasil," in 14ª Edição, 39, 2024.
- [3] T. d. O. Silva, Os impactos sociais, cognitivos e afetivos sobre a geração de adolescentes conectados às tecnologias digitais, Undergraduate thesis (tcc), Universidade Federal da Paraíba, João Pessoa, Brazil, 2016, submitted June 14, 2016; bibliographic study; advisor: Leblam Tamar Gomes Silva.
- [4] D. Salinas, F. Avvisati, R. Castaneda Velle, "PISA 2022 Results: The State of Learning and Equity in Education – Volume I," Technical report, Organisation for Economic Co-operation and Development (OECD), Paris, France, 2023, doi:10.1787/53f23881-en.
- [5] R. Bailey, "Student writing and academic literacy development at university," Journal of Learning and Student Experience, 1, 7–7, 2018.
- [6] F. Ngo, "Fractions in college: How basic math remediation impacts community college students," Research in Higher Education, 60, 485–520, 2019, doi:10.1007/s11162-018-9519-x.
- [7] L. Aranda, E. Mena-Rodríguez, L. Rubio, "Basic skills in higher education: An analysis of attributed importance," Frontiers in Psychology, 13, 752248, 2022, doi:10.3389/fpsyg.2022.752248.
- [8] G. Kuan, Y. C. Kueh, N. Abdullah, E. L. M. Tai, "Psychometric properties of the health-promoting lifestyle profile II: cross-cultural validation of the Malay language version," BMC public health, 19(1), 1–10, 2019, doi:10.1186/s12889-019-7109-2.
- [9] V. R. Franco, "Aprendizado de Máquina e Psicometria: Inovações Analíticas na Avaliação Psicológica," PePisic Periodicos de Psicologia, 20, a.c, 2021, doi:10.15689/ap.2021.2003.ed.
- [10] A. Alshankiti, A. Namoun, "Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification," IEEE Access, 8, 203827–203844, 2020, doi:10.1109/ACCESS.2020.3036572.
- [11] E. P. F. Kelvin dos Santos, "A Previsão de Evasão em Cursos de Graduação Utilizando Machine Learning," Caderno Virtual, 1(58), 13–15, 2024.
- [12] A. da Silva Franqueira, B. F. C. Zanetti, Carlos A. L. Bitencourt, D. Z. Franco, E. H. B. de Oliveira, Érica Rafaela dos Santos Campos, H. G. M. Júnior, J. da Cruz Chagas, "Análise impulsada por IA para previsão de desempenho estudantil," Cuadernos de Educación y Desarrollo, 16(5), 14–17, 2024.
- [13] A. M. Feijó, E. F. R. Vicente, S. M. Petri, "O uso das escalas Likert nas pesquisas de contabilidade," Revista Gestão Organizacional, 13(1), 27–41, 2020, doi:10.22277/rgo.v13i1.5112.
- [14] A. Abu Saa, M. Al-Emran, K. Shaalan, "Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques," Technology, Knowledge and Learning, 24(4), 567–598, 2019, doi:10.1007/s10758-019-09408-7.

- [15] D. Petkovic, K. Okada, M. Sosnick, A. Iyer, S. Zhu, R. Todtenhoefer, S. Huang, "Work in progress: a machine learning approach for assessment and prediction of teamwork effectiveness in software engineering education," in 2012 frontiers in education conference proceedings, 1–3, IEEE, 2012, doi:[10.1109/FIE.2012.6462205](https://doi.org/10.1109/FIE.2012.6462205).
- [16] J. Zhu, C. Zhu, Z. Wang, "Application of Machine Learning in English Language Teaching Quality Assessment," in 2024 International Conference on Interactive Intelligent Systems and Techniques (IIST), 300–304, IEEE, 2024, doi:[10.1109/IIST62526.2024.00041](https://doi.org/10.1109/IIST62526.2024.00041).
- [17] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, "Student engagement predictions in an e-learning System and their impact on student course assessment scores," *Computational intelligence and neuroscience*, **2018**(1), 6347186, 2018, doi:[10.1155/2018/6347186](https://doi.org/10.1155/2018/6347186).
- [18] J. G. Valen-Dacanay, T. D. Palaoag, "Exploring The Learning Analytics Of Skill-Based Course Using Machine Learning Classification Models," in 2023 11th International Conference on Information and Education Technology (ICIET), 411–415, IEEE, 2023, doi:[10.1109/ICIET56899.2023.10111210](https://doi.org/10.1109/ICIET56899.2023.10111210).
- [19] A. Ravuri, M. Lourens, S. Aswini, G. Nijhawan, R. S. Zabibah, R. Chandrashekar, "Improving Personalized Education: A Machine Learning Method for Flexible Learning Environments," in 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 1715–1720, IEEE, 2023, doi:[10.1109/UPCON59197.2023.10434888](https://doi.org/10.1109/UPCON59197.2023.10434888).
- [20] I. Issah, O. Appiah, P. Appiahene, F. Inusah, "A systematic review of the literature on machine learning application of determining the attributes influencing academic performance," *Decision analytics journal*, **7**, 100204, 2023, doi:[10.1016/j.dajour.2023.100204](https://doi.org/10.1016/j.dajour.2023.100204).
- [21] Y. Christian, Y. Choo, N. Yusof, "Systematic literature review on the use of machine learning in online learning in the context of skill achievement," *Journal of Theoretical and Applied Information Technology*, **102**(6), 2466–2479, 2024.
- [22] H. Munir, B. Vogel, A. Jacobsson, "Artificial intelligence and machine learning approaches in digital education: A systematic revision," *Information*, **13**(4), 203, 2022, doi:[10.3390/info13040203](https://doi.org/10.3390/info13040203).
- [23] D. Petkovic, M. Sosnick-Pérez, S. Huang, R. Todtenhoefer, K. Okada, S. Arora, R. Sreenivasen, L. Flores, S. Dubey, "Setap: Software engineering teamwork assessment and prediction using machine learning," in 2014 IEEE frontiers in education conference (FIE) proceedings, 1–8, IEEE, 2014, doi:[10.1109/FIE.2014.7044199](https://doi.org/10.1109/FIE.2014.7044199).
- [24] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *Ieee Access*, **9**, 7519–7539, 2021, doi:[10.1109/ACCESS.2021.3049446](https://doi.org/10.1109/ACCESS.2021.3049446).
- [25] A. Nabil, M. Seyam, A. Abou-Elfetouh, "Prediction of students' academic performance based on courses' grades using deep neural networks," *IEEE Access*, **9**, 140731–140746, 2021, doi:[10.1109/ACCESS.2021.3119596](https://doi.org/10.1109/ACCESS.2021.3119596).
- [26] A. Namoun, A. Alshantqi, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, **11**(1), 237, 2020, doi:[10.3390/app11010237](https://doi.org/10.3390/app11010237).
- [27] A. M. J. d. Andrade, Desempenho acadêmico, permanência e desenvolvimento psicossocial de universitários: relação com indicadores da assistência estudantil, Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, 2014, doi:[10.1590/S1414-40772017000200014](https://doi.org/10.1590/S1414-40772017000200014).
- [28] E. A. C. d. Araujo, D. F. d. Andrade, S. L. V. Bortolotti, "Teoria da resposta ao item," *Revista da Escola de Enfermagem da USP*, **43**, 1000–1008, 2009, doi:[10.1590/S0080-62342009000500003](https://doi.org/10.1590/S0080-62342009000500003).
- [29] R. Primi, "Psicometria: fundamentos matemáticos da Teoria Clássica dos Testes," *Avaliação Psicológica*, **11**(2), 297–307, 2012.
- [30] C. F. Collares, W. L. P. Grec, J. L. M. Machado, "Psicometria na garantia de qualidade da educação médica: conceitos e aplicações," *Science in Health*, **3**(1), 33–49, 2012.
- [31] G. M. d. M. Paixão, B. C. Santos, R. M. d. Araujo, M. H. Ribeiro, J. L. d. Moraes, A. L. Ribeiro, "Machine learning na medicina: revisão e aplicabilidade," *Arquivos brasileiros de cardiologia*, **118**, 95–102, 2022, doi:[10.36660/abc.20200596](https://doi.org/10.36660/abc.20200596).
- [32] C. A. Meira, L. H. Rodrigues, S. A. Moraes, "Análise da epidemia da ferrugem do cafeeiro com árvore de decisão," *Tropical Plant Pathology*, **33**, 114–124, 2008, doi:[10.1590/S1982-56762008000200005](https://doi.org/10.1590/S1982-56762008000200005).
- [33] L. Breiman, "Random forests," *Machine learning*, **45**, 5–32, 2001.
- [34] J. P. da Silva Funchal, D. F. Adanatti, "Um Estudo Sobre a Classificação de Risco na Área da Saúde Utilizando Árvores de Decisão," *iSys-Brazilian Journal of Information Systems*, **9**(3), 89–111, 2016.
- [35] T. M. Oshiro, Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica, Ph.D. thesis, Universidade de São Paulo, 2013.
- [36] C. Cortes, V. Vapnik, "Support-vector networks," *Machine learning*, **20**, 273–297, 1995.
- [37] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794, 2016, doi:[10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).
- [38] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems*, **31**, 2018.
- [39] J. T. Hancock, T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *Journal of big data*, **7**(1), 94, 2020, doi:[10.1186/s40537-020-00369-8](https://doi.org/10.1186/s40537-020-00369-8).
- [40] P. J. L. Adeodato, "Data Mining Solution for Assessing Brazilian Secondary School Quality Based on ENEM and Census Data," 13th CONTECSI International Conference on Information Systems and Technology Management, online, 2016.
- [41] J. L. C. Ramos, R. L. Rodrigues, J. C. S. Silva, P. L. S. de Oliveira, "CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais," in Simpósio Brasileiro de Informática na Educação (SBIE), 1092–1101, SBC, 2020, doi:[10.5753/cbie.sbie.2020.1092](https://doi.org/10.5753/cbie.sbie.2020.1092).
- [42] K. Anguraj, B. Thiyaneswaran, G. Megashree, J. Preetha Shri, S. Navya, J. Jayanthi, "Crop recommendation on analyzing soil using machine learning," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, **12**(6), 1784–1791, 2021.
- [43] J. An, "Using CatBoost and Other Supervised Machine Learning Algorithms to Predict Alzheimer's Disease," in 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 1732–1739, 2022, doi:[10.1109/ICMLA55696.2022.00265](https://doi.org/10.1109/ICMLA55696.2022.00265).
- [44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, 2825–2830, 2011, doi:[10.5555/1953048.2078195](https://doi.org/10.5555/1953048.2078195).
- [45] C. Liu, M. White, G. Newell, "Measuring the Accuracy of Species Distribution Models: A Review," in Proceedings of the 18th World IMACS / MODSIM Congress, 4241–4246, Victoria, Australia, 2009, reviewed accuracy metrics like discrimination capacity and reliability in SDMs; introduced both threshold-dependent and threshold-independent indices.
- [46] M. Sokolova, G. Lapalme, "Learning Opinions in User-Generated Web Content," *Natural Language Engineering*, **17**(4), 541–567, 2011, doi:[10.1017/S135132491100012X](https://doi.org/10.1017/S135132491100012X), first published online 11 March 2011; Cambridge University Press.
- [47] M. Bekkar, H. K. Djemaa, T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, **3**(10), 27–38, 2013, open access; ISSN 2224-5782 (print), 2225-0506 (online).

- [48] R. L. Chambers, H. Chandra, N. Tzavidis, "On Bias-Robust Mean Squared Error Estimation for Pseudo-Linear Small Area Estimators," *Survey Methodology*, **37**(2), 153–170, 2011, develops a simpler, bias-robust MSE estimator for pseudo-linear small area estimators, including EBLUP, MBDE, and M-quantile predictors.
- [49] T. Hastie, R. Tibshirani, J. Friedman, "The elements of statistical learning. Springer series in statistics," New York, NY, USA, 2001.
- [50] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, **30**, 2017.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).