Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 31/10/2024; Camera ready: 26/03/2025; 1st round notif.: 27/01/2025; Edition review: 20/05/2025; New version: 14/02/2025; Available online: 20/05/2025; 2nd round notif.: 20/03/2025; Published: 20/05/2025;

# Predictive system for failure of basic education student

*Bruna Damaris Ramos dos Santos*
*Universidade Federal de Alagoas, Alagoas, Brazil*
*ORCID: 0000-0001-9819-201X*
*bdrs@ic.ufal.br*

*Daniel Alisson Feitosa Lopes*
*Confederação Nacional da Indústria, Brasília, Brazil*
*ORCID: 0000-0001-6489-2720*
*daniel.lopes@sesicni.com.br*

*Baldoino Fonseca dos Santos Neto*
*Universidade Federal de Alagoas, Alagoas, Brazil*
*ORCID: 0000-0002-0730-0319*
*baldoino@ic.ufal.br*

*Marcelo Costa Oliveira*
*Universidade Federal de Alagoas, Alagoas, Brazil*
*ORCID: 0000-0002-0825-6217*
*oliveiramc@ic.ufal.br*

*Rafael Sampaio de Melo Fragoso*
*Federação das Indústrias do Estado de Alagoas, Alagoas, Brazil*
*ORCID: 0009-0002-1807-1758*
*rafael.fragoso@sistemafiea.com.br*

## Abstract

*Academic performance measures how students are doing according to the study plan proposed by schools. Educational performance is a crucial factor in education, making it important to develop a system that can indicate if a student is at risk of failing. This work discusses the development of a student failure prediction system for basic education using a dataset of about three million records of students' assessment grades from schools in Brazil. The system's primary objective is to identify students at risk of failing and provide indications to educators in search of preventing students' failure. The system architecture uses historical data from Middle and Basic Secondary Education students' assessments to predict approval or failure outcomes. Classification algorithms, including K-Nearest Neighbors, Decision Tree, Random Forest, and eXtreme Gradient Boosting, were applied for the system development, creating machine learning models. Results of model evaluations indicated that Random Forest and eXtreme Gradient Boosting showed the best performances. The final system utilizes eXtreme Gradient Boosting to create the predictive models due to its consistent performance and computational efficiency, achieving approximately 86% accuracy and F1-Score. These findings are implemented in a Business Intelligence Panel to empower educators to identify and assist students at risk of academic failure proactively.*

***Keywords:*** *Artificial Intelligence; Predictive Models; Educational Data Mining; Failure Prediction; Supervised Classification;*

## Resumo

*O desempenho acadêmico mede o desempenho dos alunos conforme o plano de estudos proposto pelas escolas. O desempenho educacional é um fator crucial na educação, tornando importante o desenvolvimento de um sistema que possa indicar se um aluno está em risco de reprovação. Este trabalho discute o desenvolvimento de um sistema de previsão de reprovação de alunos para a educação básica usando um conjunto de dados de cerca de três milhões de registros de notas de avaliação de alunos de escolas no Brasil. O objetivo principal do sistema é identificar alunos em risco de reprovação e fornecer indicações aos educadores em busca de prevenção da reprovação dos alunos. A arquitetura do sistema usa dados históricos de avaliações de alunos do Ensino Básico para prever resultados de aprovação ou reprovação. Algoritmos de classificação, incluindo K-Nearest Neighbors, Decision Tree, Random Forest e eXtreme Gradient Boosting, foram aplicados para o desenvolvimento do sistema, criando modelos de aprendizado de máquina. Os resultados das avaliações dos modelos indicaram que Random Forest e eXtreme Gradient Boosting apresentaram os melhores desempenhos. O sistema final utiliza o eXtreme Gradient Boosting para criar os modelos*

*Cite as:* *Santos, B. D. R., Oliveira, M. C., Lopes, D. A. F., Fragoso, R. S. M. & Santos Neto, B. F. Predictive system for failure of basic education student. Revista Brasileira de Informática na Educação, vol. 33, 394–416. https://doi.org/10.5753/rbie.2025.5127.*

*preditivos devido ao seu desempenho consistente e eficiência computacional, alcançando aproximadamente 86% de precisão e F1-Score. Essas descobertas são implementadas em um Painel de Business Intelligence para capacitar educadores a identificar e auxiliar alunos em risco de fracasso acadêmico proativamente.*

***Palavras-chave:*** *Inteligência Artificial; Modelos Preditivos; Mineração de Dados Educacionais; Previsão de Reprovação; Classificação Supervisionada;*

# 1 Introduction

Repetition often indicates a lack of progress in school. It can result from various factors, including academic failure, unsatisfactory performance, insufficient exam scores needed to advance, age restrictions, poor attendance, or simply a lack of available educational opportunities. It is estimated that in the World, approximately 10% to 15% of children of school age experience repetition or failure grades (Byrd, 2005) (Sunny et al., 2017).

According to a recent UNICEF report (UNICEF Brasil, Instituto Claro & Cenpec, 2021), about 2.1 million students in public schools of Basic Education in Brazil were considered school failures, totaling about 8% of the students. Failures in education can profoundly impact students' lives, leading to problematic behaviors, increasing the likelihood of dropping out of school, and compromising educational and professional paths into adulthood. These failures can cause feelings of discouragement and loss of confidence, which may result in diminished efforts to seek improvement. Consequently, students who experience these setbacks are more likely to encounter further failures (Crosnoe, 2002) (Kamal & Bener, 2009). Moreover, student failures have significant social and economic consequences, reinforcing cycles of disadvantage and social exclusion. This situation increases the likelihood of involvement in criminal activities, as these individuals often lack the skills and qualifications needed for good job opportunities. Consequently, this contributes to higher unemployment rates and reduced productivity (Giavrimis & Papanis, 2008). Identifying and supporting at-risk students is crucial in education, as it aims to promote individual academic success and improve the educational system as a whole (Al-Tameemi et al., 2023)(Silva et al., 2020).

Educational institutions can use information collected from failures to recognize problems affecting students' performance (Singh & Pal, 2020). Moreover, collected data can be used for early detection of the student's failure risk situation for early identification of learning difficulties, promoting actions that can help the student improve academic performance (Chanlekha & Niramitranon, 2018). Early detection involves analyzing student performance and predicting whether the student will fail. Some works tried to make an early detection using the recorded data in many ways(Simanca et al., 2019). Due to the teacher's work routine and the large volume of data to be analyzed, the task of making predictions through manual data analysis is both intractable and costly (Fayyad et al., 1996). Hence, the high cost of manual analysis and the need to promote actions is the motivation for creating computational predictive systems that can assist educators with the needed insights into the risk of student failure based on the analysis of predictive results, the innovative and beneficial application of data mining techniques has great significance in the progress of learning environments, making possible the adoption of initiatives aimed at improving the learning process in the classroom (Albreiki et al., 2021).

Recent studies such as (Amal Asselman & Aammou, 2021), (Siddique et al., 2021), (Singh & Pal, 2020), (Ouatik et al., 2022) used machine learning techniques with data mining of educational data to provide prediction models of student performance. In (Amal Asselman & Aammou, 2021), the aim is to assess whether ensemble methods can improve student performance prediction. Using three different databases containing information about students, activities performed, and associated skills, the Random Forest, AdaBoost, and eXtreme Gradient Boosting algorithms were tested. The results indicated improved predictions, with eXtreme Gradient Boosting being highlighted as the most effective classifier. In (Siddique et al., 2021), the study aims to identify critical factors impacting student performance by proposing constructing classification models to predict academic performance. The approach combines multiple classifiers, showing promise in early identification of academic performance and improving the learning process. In (Singh & Pal, 2020), using student individual information data and academic information data, an approach was used to improve the prediction of students' performance using two ensemble methods, bagging classifiers and boosting classifiers. The results of the previous work were positive, where the boosting techniques showed an accuracy of 91.76% compared to 89.56% for bagging techniques. In (Ouatik et al., 2022), a system was made to predict students' academic success and failure using Big Data to minimize the execution time without losing the efficiency of the machine learning algorithms. This approach showed to be effective in this case, where they could arrive at a recognition rate of 87.32% by the SVM algorithm. In (Alalawi et al., 2024), predictive models were created based on continuous assessment data from undergraduate students, getting results of around 86.9% for binary classification and 80% for multiclass classification. In (Javed et al., 2024), it was created a framework capable of making predictions to identify students at risk of failure for higher education by using data collected from an online system where it showed to be highly effective in terms of evaluation metrics results getting 94.8% accuracy and 95% F1-score for binary classification using Random Forest. In (Rodrigues et al., 2024), proposed the use of Transformer architecture for predicting student performance in primary and secondary education using semi-structured data of assessment grades. The study compares Transformers, eXtreme Gradient Boosting, and a feedforward neural network, evaluating them with Area Under the Precision-Recall Curve (AUC-PR). Extreme Gradient Boosting initially outperformed the others with 94%, while Transformers started at 88% but improved to 97% as more data became available. These studies showed some limitations regarding the techniques used, high computational cost, and the learning environment in which they were applied. They focused solely on information from specific assessment activities, student-specific information, data from online interactions, and the use of synthetic datasets together with real-world data. Those factors may not capture the variability of the student interactions, limiting their applicability and generalizability of the model.

In this context, the main objective of this work was to develop an artificial intelligence tool to create a predictive system capable of predicting the probability of a student failing a subject following the classification of 'Approved' or 'Failed' using machine learning to learn historical patterns based on recorded assessment grades. The database used has records of more than three million assessment grades from more than three thousand students from the Social Service of Industry of Alagoas (SESI-AL) schools.

The proposed approach is aimed at student performance regarding grade assessment from the academic year. Furthermore, by using students' grades and not including personal student information, there is a lower risk of bias regarding specific details such as gender, age, or ethnicity of students. Also, the student's data was obtained from real-world classrooms, which provides improved model generalizability for predicting concrete outcomes. Based on the results presented by the predictive system, it was possible to have indications of students who are at risk of failing.

This work was organized into sections detailing the process of the development of a predictive system, presenting the machine learning methods applied, explaining the phases employed, the use of data in supervised classification algorithms, and finally, the attainment of performance results of the developed predictive models, as well as prediction results. Section 1 presents this work introduction, defining the project's scope and setting the context for the work. Section 2 reviews the existing literature for predicting students at risk of failure in education. Section 3 focuses on describing the processes conducted to apply the machine learning methods to develop the predictive models. Section 4 presents the performance evaluation results for selecting the classification algorithm to build the classification model and justifies the decision behind the chosen classifier for the final version of the system. It also demonstrates how the system's output results are displayed and what actions can be taken based on the obtained information. Section 5 summarizes the proposed system and the key results of the developed system.

## 2 Related works

In the literature, a few studies have aimed to develop systems to predict students at risk of failure or academic performance using machine learning techniques to create models capable of identifying those students. Most of those studies used students' individual demographic data found in the schools' system to build the predictive models like (Singh & Pal, 2020), while (Alalawi et al., 2024) and (Rodrigues et al., 2024) used data provided by academics from the school system for assessments grades to make predictions for students performance. Moreover, some of them were developed for higher education students((Singh & Pal, 2020), (Alalawi et al., 2024)) or for specific environments that were used to collect the data to be used in the predictive models while in (Rodrigues et al., 2024), the authors emphasize the importance of predicting student performance in primary and secondary education, arguing that early detection of academic challenges in this stage is essential for ensuring quality higher education. In these studies, there was a variety of classification algorithms used, predominantly ensemble algorithms such as Random Forest, eXtreme Gradient Boosting, and MultiBoost, along with individual classifiers like Logistic Regression, Support Vector Machines, Decision Trees, Naïve Bayes, and K-Nearest Neighbors and deep learning strategies as the Transformer Architecture for predicting academic performance.

In (Siddique et al., 2021), the authors highlight the scarcity of studies using databases focused on secondary education and propose a model to predict academic performance at this level. It suggests that predicting performance in secondary education is important for student performance at higher educational levels, as it mostly says that the early detection of problems and improvement of the performance of secondary education level in science subjects is crucial for higher education quality and so, identifying and addressing weaknesses in academics and personal traits can enhance academic performance. The model was developed using data mining techniques and academic and personal data to predict the educational progress of secondary school students. The approach utilized was the use of ensemble methods such as bagging and boosting to predict students' academic performance, with the best-performing model being a MultiBoost with Multilayer Perceptron, achieving an accuracy of 98.7% and precision, recall, and F-Score of 98.6%. Moreover, it was implied that the proposed model can be used to identify academic performance in the early stages to improve learning. However, the study has limitations, including a focus on physical learning environments, an inability to recommend suitable learning paths based on individual performance, and a lack of a comprehensive framework across all educational levels. The model also has potential biases and constraints when generalized due to the specification of individual student traits.

In (Amal Asselman & Aammou, 2021), the objective was to identify whether ensemble methods may improve student performance prediction. The authors have used three different real-world databases: student data, activities carried out, and skills associated with the activities carried out. The study presented a Knowledge Tracing (KT) exploratory approach to enhance the Performance Factors Analysis (PFA) algorithm using Ensemble Learning techniques. Results showed that the approach used was effective, with an improvement in prediction performance. The test showed that Random Forest achieved better results when data was in smaller quantities. Regarding large datasets, eXtreme Gradient Boosting achieved the highest prediction performance due to its scalability. However, this study could have been more extensive in its applicability to the specific conditions analyzed. The study aimed to enhance the PFA algorithm using ensemble methods but faced challenges in generalizing the variables used for model creation.

The work of (Singh & Pal, 2020) proposed an approach for predicting academic performance by using ensemble methods and comparing classification algorithms based on bagging, such as Random Forest, and boosting, such as eXtreme Gradient Boosting, for predicting students' performance in universities. The single classifiers initially used were Decision Tree, Naïve Bayesian, K-Nearest Neighbors, and Extra Tree, applied to a database consisting of individual student information, socioeconomic information, and academic performance. Ensemble techniques like bagging (Random Forest) and boosting(eXtreme Gradient Boosting) were employed to enhance the accuracy of the single classification. The results indicate that the boosting technique achieved the highest accuracy of 91.76% in predicting academic performance, surpassing both bagging and individual classifiers. Specifically, the best accuracy among the particular machine learning classifiers was 86.83% using Naïve Bayesian. It demonstrated the effectiveness of ensemble methods in improving classification accuracy compared to single classifiers. However, the limitations of this study include potential biases since the information used was regarding a lot of individual details of the students, limiting generalization considering not all institutions could access such data from the students. The study was also applied to a specific scenario.

In (Ouatik et al., 2022), the focus was on the recognition of the academic success of students using Big Data and data mining methods such as K-Nearest Neighbors, C4.5, and the Support Vector Machine algorithm based on metrics such as engagement, course completion and learning. The data used encompasses the students' personal information, the academic evaluation, the activities of the students, and the psychological and the environment of the students applied to university students. The results of this study, by comparison of performance measures and execution time between the methods used, were that the Support Vector Machine algorithm had the highest classification rate at 87.32%. Regarding data selection for the chosen machine learning methods, it was possible to see that some factors were crucial to the prediction made, and one of those factors was academic assessment. However, even though some other factors that involved specific personal information such as economic status and parent educational level appeared to be important in this case for the prediction, those could lead to biases, and it also could limit the generalization of the models generated for institutions that don't have access to such data.

In (Alalawi et al., 2024), the authors focus on presenting a learning analytics intervention framework for educators to use on their courses to get results of predictions of at-risk failures students using historical continuous assessment data. The dataset used in the study was provided by two academics for their specific courses for undergraduate students. The academics used 497 student assessment records from 2017–2020 to develop predictive models, relying solely on assessment data, which was easily accessible data for those academics. The predictive models offer educators early insights into students at risk of underperformance, failure, or dropout, enabling timely and targeted intervention. In this case, the algorithms used for the prediction task were Logistic Regression, Support Vector Machine, Decision Tree, k-nearest Neighbours, and Naïve Bayes. Two types of classification were applied in this case for binary classification; at early assessments, Logistic Regression performed the best with an accuracy of 86.9%, F-measure of 92.9%, and recall of 98.6%, while K-Nearest Neighbors performed best after more data was inserted with an accuracy of 93%, F-measure of 95.9%, and recall of 96.2%. For multiclass classification, logistic regression outperformed the other models in all assessments with around 80% accuracy. This study shows an effective use of assessment data for students for specific courses based on the data provided by the academics that used the system, getting good metrics results for the predictive models; however, its implementation was limited to the use of a specific undergraduate course, the need for course-specific predictive models, challenges in ensuring intervention effectiveness, the requirement for educator training, and limited generalizability.

In (Javed et al., 2024), is proposed that predicting students' performance is an important subject for learning environments because it aids in the development of effective strategies to prevent dropouts and proposes a framework to make predictions about students' failures and introduces the factors that influenced the prediction via feature engineering in an online environment for higher education. The data used in this study encompass students' behavior and performance, the behavior being from the interactions with the system as well as the engagement and satisfaction measured by surveys in the system and performance based on the student grades. This information was recorded as daily summaries for each of the 32593 students. The machine learning approach of this work was aimed at using The Synthetic Minority Oversampling Technique to correct imbalanced data, and then the preprocessed input data was inserted in the Cat Boost, Random Forest, Logistic Regression, Linear Discriminant Analysis, Decision Tree, Support Vector Machine, Naive Bayes and Neural Network to generate prediction models. The results from this study showed that the Random Forest classifier outperformed all the others with 94.8% accuracy and 95% F1-score for binary classification for pass or fail students. This study showed the efficacy of the use of machine learning to predict outcomes for students and identify students at risk; however, it was applied to an online scenario where some of the variables used were only capable of being obtained in the online system, lacking generalization for on-site classroom environments.

In (Rodrigues et al., 2024), explores the use of deep learning models to predict the academic performance of students in Primary and Secondary Schools. The data used in this student was semi-structured data from periodical assessment grades belonging to different institutions. The models were created from three different architectures, the Transformer, eXtreme Gradient Boosting and a feedforward neural network, these were evaluated using AUC-PR as the evaluation metric. eXtreme Gradient Boosting performed best initially, achieving an AUC-PR of 94% with early evaluation data, while Transformers scored 88%. However, as more assessments were included, the Transformer model improved significantly, reaching 97%, surpassing the other models. This suggests that while eXtreme Gradient Boosting is effective with limited data, Transformers excel when more information is available, making them a strong choice for long-term academic predictions. For this study, it focused on the application of deep learning models in semi-structured data, the use of the Transformer architecture can be computationally intensive for real-time data predictions because it requires substantial resources to train and deploy so an less intensive approach can be used in the available data to obtain effective predictive models. Besides, the Transformer model showed less efficacy in the early stages of prediction, but these early stages are important so that educators can take measures as early as possible for students at risk.

# 3 Materials and Methods

The SESI-AL schools have records of approximately three million assessment scores. These schools, linked to the Social Service of Industry (SESI) and the National Confederation of Industry (CNI), provide private education from preschool to Basic Secondary School, emphasizing academic advancement and job readiness. They prioritize technical skills, offer extracurricular activities, and partner with companies for comprehensive student training. SESI, funded by industry, ensures quality education, often prioritizing enrollment for children of industrial workers, aligning with its mission to serve industrial communities. In these schools, basic education refers to middle school students aged 11 to 14 years old and basic secondary school students aged 15 to 17 years old in Brazil. In the SESI-AL schools, 16.78% of Middle School students are considered as failed regarding their assessment scores performance in the years 2017 to 2022 and around 5% for Basic Secondary School in the year 2022, and that indicates a moderately high number of students are failing in the subjects according to their performance in assessments.

The assessment system in the SESI-AL schools is divided into three quarters, each lasting three months. Each quarter includes three assessments, which are conducted throughout the quarter. At the end of each quarter, there is a general assessment known as the end-of-quarter assessment. The grades from all assessments in the quarter are combined to determine whether a student passes or fails. If a student fails, they can take a re-evaluation assessment. If insufficient grades persist by the year's end, a class council will review the student's overall performance, including grades, behavior, and participation, to possibly assign a passing grade. For the predictive system models, only the regular assessments and final re-evaluations were considered; class council grades were not included.

## 3.1 System architecture

Historical data from Basic Education student assessment grades was used to develop a predictive system, employing modeling techniques to anticipate approval or failure outcomes as 'Approved' or 'Failed'. The system's general process began with acquiring student data from the available database of the SESI-AL schools located throughout the State of Alagoas, Brazil. The database consisted of individual grades for each student, for each subject, for their respective school class, and for each school grade. Then, the system was designed to separate the data from Middle and Basic Secondary Education so that it could be treated distinctively.

The overall system process began with acquiring student data from the available database. Educational data was obtained from the internal SESI-AL database, which collected students' grades from educational management software and stored the data in a specific database after the data was transformed and standardized. The process allowed the system to access the individual grades of each student for each subject and their respective school class in a consolidated manner within a single database.

Next, the system had to differentiate the data handling for Middle School and Basic Secondary School. This step of the process was justified by the fact that the two levels may be at different assessment stages in the year when the predictive system is used. Additionally, historical Basic Secondary School data was recorded in the database differently from how Middle School data was registered. As a result, two separate code sections were created to handle this data. Model creation and data mining started with Middle School data, and the Basic Secondary Education process began when the first process ended.

To make predictions based on the new data that was stored throughout the year, milestones were defined according to the grades available in the system at the execution time. Once these milestones were defined, it was determined which assessment grades could be used to begin training the predictive model. Predictive model training occurred for each subject in the school grade, so each subject model was trained, referring only to historical data from that subject. Throughout the system's execution, the system created predictive models for each school grade across seven grades. For each model created, each one of them made predictions using new data from enrolled students.

At the end of the process, the tool predicted whether each student would be 'Approved' or 'Failed' based on their accumulated grades in each subject. Besides classification, the tool provided the probability of the student's risk of failure. The results were recorded in a dataset alongside information about students enrolled in the current year and saved in the database. Figure 1 presents an overview of the system execution process, which consists of data acquisition and processing, training predictive models, and using the developed models to predict outcomes for enrolled students. Part 1 indicates where the Middle School data preprocessing began, followed by all the other processes regarding only that data. Part 2 indicates that the beginning of the Basic Secondary School data was processed after all the processes and the registry of results in the database for Middle School data had ended.

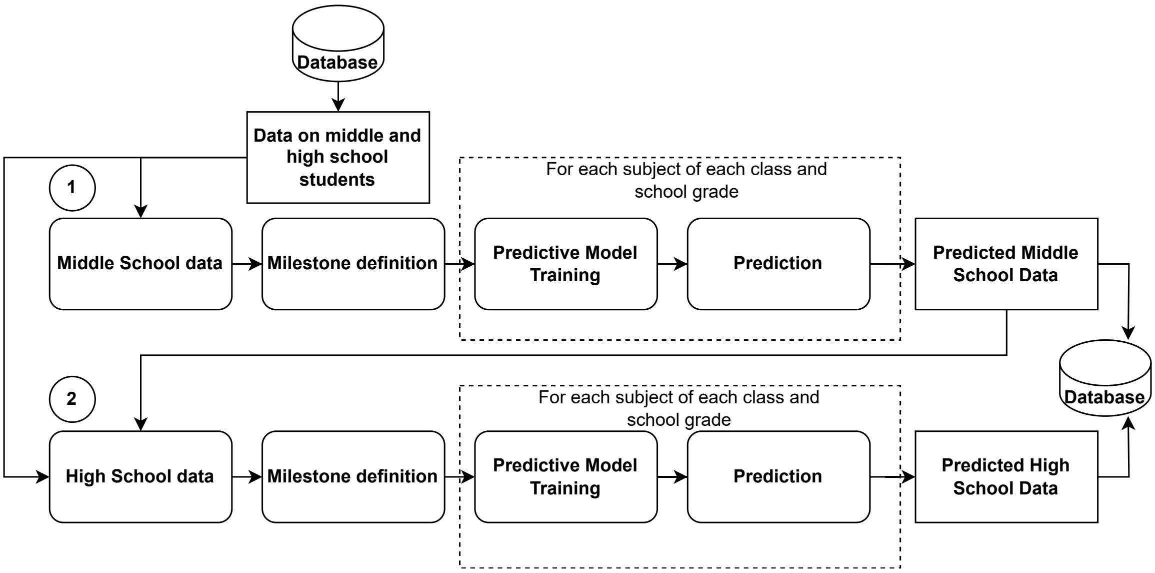

Database
Data on middle and high school students
1
For each subject of each class and school grade
Middle School data
Milestone definition
Predictive Model Training
Prediction
Predicted Middle School Data
Database
2
For each subject of each class and school grade
High School data
Milestone definition
Predictive Model Training
Prediction
Predicted High School Data

Figure 1: System architecture flowchart.

The predictive system was created using Python 3.8 programming language and its available data manipulation and machine learning libraries. Those libraries were Pandas(2.2.0), Scikit-learn(1.4.0), Numpy(1.26.3), SQLAlchemy(1.4.51), XGBoost(2.0.3) and Lazy Prediction(0.2.12). Furthermore, Microsoft SQL Server and DBeaver obtained the students' data from the school database. The development was carried out in collaboration with the Federation of Industry of the State of Alagoas (FIEA). The project's source code is private due to privacy issues, and the people involved signed a confidentiality agreement.

Furthermore, the data obtained from the database were related to existing individuals, so it was necessary to take measures such as using only the numeric representation to identify a student instead of using the actual student names and assuring that the data obtained were only the needed assessment grades as well as the identification of school grade, school class and subject associated to each assessment grades. All the access was treated with restrictions aimed at safe handling, so given the sensitive nature of the data, measures have been taken to prevent undue disclosure. In conjunction with this, the project was conducted in accordance with Brazil's General Data Protection Law (LGPD).

The system was created following the Knowledge Discovery in Database (KDD) phases, which essentially consisted of five phases: data selection, data preprocessing, data transformation, data mining, analysis, and interpretation of results (Fayyad et al., 1996). Figure 2 shows the project workflow according to these stages. Initially, the data was selected based on its relevance to the problem and the quality and quantity of data available in the database. Based on the selected data, the information extracted from the database was from the 6th, 7th, 8th, and 9th years of Middle School and the 1st, 2nd, and 3rd years of the new Basic Secondary School curriculum. Once the data was obtained, preprocessing was performed to prepare the input data for the classification models. Subsequently, data mining was conducted, using the processed data as input for the classification algorithms.

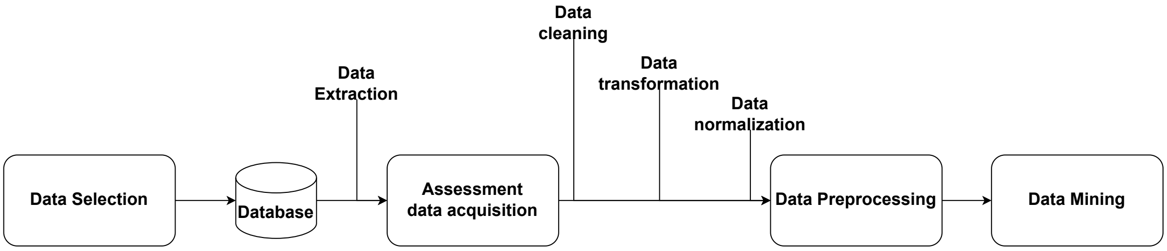


Figure 2: Project execution flowchart.

### 3.2 Data Selection

The data selected were the grades from the quarterly assessments recorded in the database. These grades are the criteria teachers use to determine whether a student was 'Approved' or 'Failed'. Assessments were scored on a scale of 0 to 10, with scores equal to or greater than 6.0 indicating a positive evaluation result. In contrast, scores below 6.0 indicated a negative result, which is an indication of failure. The database contained more than three million student assessment grades in various subjects from 2017 to 2023. Data were extracted from the 6th, 7th, 8th, and 9th years of Middle School from 2017 to 2023 and the 1st, 2nd, and 3rd years of Basic Secondary Education for 2022 and 2023. To differentiate the data from Middle School to Basic Secondary schools, the names of the school classes were considered, the data corresponding to the school classes from Middle School were considered as the data selected for the Middle School dataset, and then the data corresponding to the Basic Secondary school classes were considered to the Basic Secondary School dataset.

The assessment grade was chosen as the only measurement of students' performance due to its standardized nature. It was seen as the better factor to measure students' performance because it is the most crucial factor that directly links to academic outcomes since the teachers use the assessment grades to determine if the student is approved or failed the subject. Other factors, such as classroom presence, were not included as they may not consistently reflect students' grasp of the material or their actual academic achievement. Consequently, the assessment grade was considered a more reliable indicator for evaluating performance.

**Evaluative milestones:** Since assessment grade score data was registered throughout the year, assessment milestones consist of lists indicating the assessment grades used for training prediction models. Assessments were conducted over three quarters of the year, each composed of three assessments, an end-of-quarter assessment, and a re-evaluation assessment. Figure 3 shows how evaluation milestones were defined throughout the year.

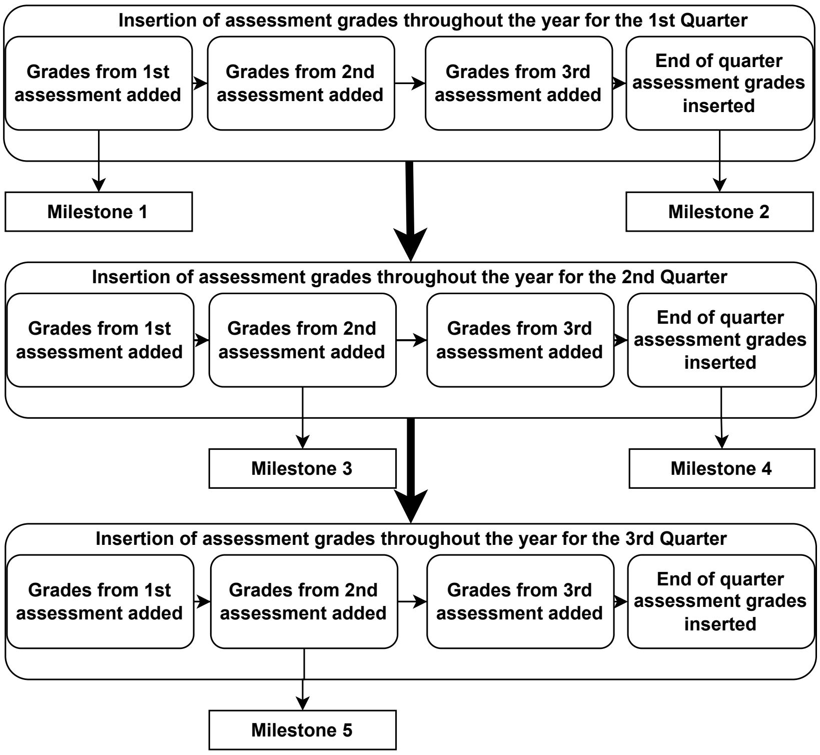

Insertion of assessment grades throughout the year for the 1st Quarter
Grades from 1st assessment added
Grades from 2nd assessment added
Grades from 3rd assessment added
End of quarter assessment grades inserted
Milestone 1
Milestone 2
Insertion of assessment grades throughout the year for the 2nd Quarter
Grades from 1st assessment added
Grades from 2nd assessment added
Grades from 3rd assessment added
End of quarter assessment grades inserted
Milestone 3
Milestone 4
Insertion of assessment grades throughout the year for the 3rd Quarter
Grades from 1st assessment added
Grades from 2nd assessment added
Grades from 3rd assessment added
End of quarter assessment grades inserted
Milestone 5

Figure 3: Evaluative milestones.

The prediction phases were divided into five evaluative milestones, created so that the result of the algorithm indicated the current milestone. Thus, it was possible to determine the specific assessment grades used in the prediction presented in the results, the classification results of the prediction as ‘Approved’ or ‘Failed’, and the probability of the corresponding prediction.

These milestones were divided and defined together with the pedagogical team of SESI-AL based on the time of the year of the assessments and the registry of grades in the system and what was thought to be the best to determine the student’s performance in the said quarter together with the previous ones. The SESI-AL schools work on a system of three-quarters of assessments where each quarter has four assessments, three regular assessments, and the end-of-quarter assessment. Regarding the time of the year, the quarters, the first quarter goes from February to May, the second quarter goes from June to September, and the third quarter goes from September to December. Hence, the milestones had to make sense regarding the quarters of the year the students have assessments. Moreover, the last milestone was defined until the closest date the prediction could make sense, as in the teachers could do something to help the student, and that was around the second assessment of the third quarter in October, not getting too close to the end of the year.

Therefore, all assessment grades were used, except re-evaluations, for the entire first and second quarters. Regarding the third quarter, the assessment grades referring only to the first two assessments were used because the other evaluations for that quarter were very close to the end of the academic year.

In particular, the predictions regarding Milestone 1 were made with the assessment grades from the first assessment of the first quarter of the year, together with the existing grades from the fifth assessment milestone of the previous year. In other words, for students who were not new and were in the 7th, 8th, and 9th grades of Elementary School and the 2nd and 3rd grades of the new Basic Secondary School, the assessment grades data for the subject to be assessed from the previous year were obtained for this student, this process was done since this way it is possible to have historical data directly related to the student. For the other students, the grades from the current milestone were repeated to fill in the data. These conditions were initially established because the first milestone only included one grade from the student, the first assessment, so it was seen that it was necessary to add more data to the training related to the student. Furthermore, in initial tests, Milestone 1 performed worse than the other milestones regarding prediction performance. Therefore, for new students, those who do not present information from previous years in the system, the grade of the first current milestone is repeated, and only the student's current performance is reflected. Students without registration, students entering the 6th grade of Elementary School, and students entering the 1st year of Basic Secondary School were considered new students since, even if the student had attended the 9th grade at a SESI-AL school, the performance related to Middle and Basic Secondary School must be differentiated since the subjects were different. The other students were considered veterans, and the grades in the system were used. If there was a missing grade, it was filled in with the average of the existing grades.

Furthermore, students' personal data were not used to avoid biases, so the only information presented to identify a student was a unique numerical identifier, and the other attributes besides assessment grades were linked to school grade identification, classes, and subjects.

## 3.3 Data Preprocessing

In both cases, for Middle School and Basic Secondary School the school grades were split into subjects, and a predictive model was created for each subject. Thus, historical data from the years before the current were used for each subject, where the label that defines whether the student has been considered as 'Approved' or 'Failed' the subject was determined at this stage based on the student's final grade in the subject.

Figure 4 shows the flowchart for defining the student's final grade in a subject. These considerations reflected the student's academic performance regarding assessment grades, disregarding student approvals based on class council grades. Therefore, if the class council grade is not registered in the database, the student's final subject grade was based on their last assessment, which may have been the subject's final re-evaluation or the subject's annual partial average grade.

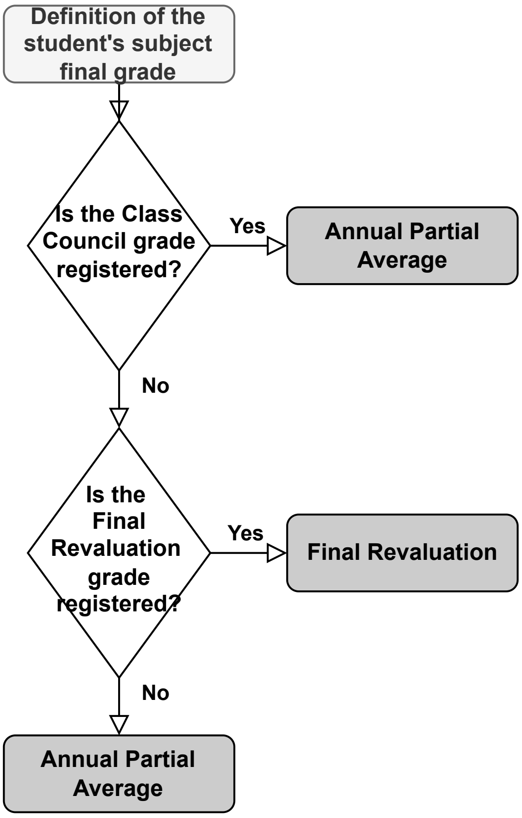

Definition of the student's subject final grade
Is the Class Council grade registered?
Yes
Annual Partial Average
No
Is the Final Revaluation grade registered?
Yes
Final Revaluation
No
Annual Partial Average

Figure 4: Final Subject Grade definition flow.

Therefore, the following considerations were made to determine the student's final grade and define the student as 'Approved' or 'Failed':

- If the Class Council grade is filled in:
  - Final grade = Annual partial average of the subject
- If the Class Council grade is not filled in:
  - If the final re-evaluation grade is filled in:
    - Final grade = Final Re-evaluation
  - If the final re-evaluation grade is not filled in:
    - Final grade = Annual partial average of the subject

Transformations were applied to the data to maximize its inclusion in the training set, in addition to the need to normalize it. In this scenario, data from previous years had to be combined into a common format, given the evolving methods of recording assessment grades over time. Regarding data normalization, the values of the scores obtained were normalized so that all grades were within a scale from zero to ten.

Another important step in Data Preprocessing was cleaning all acquired data to eliminate missing and duplicate data. Missing data marked as null or Not a Number (NaN) was removed for each subject. Furthermore, duplicated data, such as duplicate records of a student's grades in the same assessment for the same class and school grade, were removed.

### 3.4 Data mining

In the data mining phase, the data obtained after preprocessing went through the oversampling process to balance the number of 'Approved' and 'Failed', avoiding biases in learning patterns learned by the classifiers. In this phase, an adjustment was made in which failing students were oversampled to equal the number of failing students to the number of approved students.

Comparative tests were conducted to create the predictive models using a definitive classifier algorithm according to the accuracy, recall, precision, and F1-score metrics. The algorithms tested were the K-Nearest Neighbors (KNN) Classifier, Decision Tree Classifier, Random Forest Classifier, and eXtreme Gradient Boosting(XGBoost) Classifier. These were chosen in the initial tests using the Lazy Prediction library in Python, which provides the results of tests with base parameters of several algorithms.

Grid Search, Random Search and Bayesian Discrete Search (Liashchynskyi & Liashchynskyi, 2019)(Bergstra & Bengio, 2012)(Bischl et al., 2023) identified optimal hyperparameters for predictive models across chosen algorithms. Table 1 shows the parameters and the range of the tests made for XGBoost.

Table 1: XGBoost Booster-Specific Parameters range and description..

| Parameter | Range | Description |
|---|---|---|
| n_estimators | 50 to 200 | Number of boosting rounds (trees). |
| max_depth | 3 to 7 | Maximum depth of a tree. |

Table 2 shows the parameters and the range of the tests made for Random Forest.

Table 2: Random Forest Parameters range and description..

| Parameter | Range | Description |
|---|---|---|
| n_estimators | 50 to 500 | Number of trees in the forest. |
| criterion | gini or entropy | Function to measure the quality of a split. |
| max_depth | 3 to 10 | Maximum depth of a tree. |

Table 3 shows the parameters and the range of the tests made for KNN.

Table 3: KNN parameter range and description..

| Parameter | Range | Description |
|---|---|---|
| n_neighbors | 1 to 31 | Number of neighbors to consider when classifying a new point. |

Table 4 shows the parameters and the range of the tests made for Decision Tree.

The parameters in Table 1, 2, 3 and 4 were chosen aimed at balancing between performance, overfitting prevention and efficiency regarding computational time (Yang & Shami, 2020).

Table 4: Decision Tree parameter range and description..

| Parameter | Range | Description |
|---|---|---|
| max_depth | 1 to 20 | Maximum depth of the tree. Limits tree growth to prevent overfitting. |

The models' input data were divided into training and testing bases, with a data division of 80% for the training base and 20% for the test base. The percentage of the splits was selected firstly based on common practices (Joseph, 2022) and then after performance tests that were made on the models created. Furthermore, the training and testing samples were stratified, and each set contained approximately the same percentage of samples from each target class in the input database. That made a total of about 231 thousand training samples, 57 thousand test samples for Middle Schools, about 29 thousand training samples, and 7 thousand test samples for Basic Secondary Schools across all the years registered.

The goal of evaluating the performance of each classifier is to discern which one can best identify and classify a student as 'Failed'. This evaluation was conducted at the four evaluation milestones throughout the year since, at the time of testing, the assessment grades for the fifth milestone had yet to be recorded.

Predictive models were created for each class subject for each milestone, as there were seven school grades composed of approximately ten subjects, with results obtained for four evaluation milestones. So, around two hundred predictive models were created to evaluate each classifier algorithm. As the results were extensive, the arithmetic mean was used to synthesize the metrics' values, depending on the specific metric to be evaluated. In order to calculate the arithmetic mean for the assessment milestones, the metric values for each subject per school grade were obtained. Then, based on these values, the average for each milestone for each Middle and Basic Secondary Education grade was defined. Therefore, the arithmetic mean values of the milestones were obtained for each school grade. Then, a new average calculation was made according to the values of the metrics for each school grade, thus getting the average evaluation values for each milestone encompassing all school grades.

To evaluate the predictive models, classification models evaluation metrics were used. The metrics used were accuracy, precision, recall, and F1-Score. Accuracy indicates the measure of how close the predictions are to the true values of the test, and it is calculated as the sum of the correct predictions divided by the total data. Precision indicates how many of the data classified as positive are actually positive, and it is calculated as the number of true positives divided by the number of total positive predictions. Recall indicates the percentage of data classified as positive compared to the amount of real positives, and it is calculated as the number of true positives divided by the sum of the true positives and false negatives. Lastly, the F1-Score is used as an indicator of precision and recall since this value is the harmonic mean of these metrics(Vujovic, 2021).

These metrics have the following corresponding equations:

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative} \tag{1}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

The predictive results obtained were saved in the database and entered into a panel where the SESI-AL pedagogical team could view this information on students' failure risk so that interventions defined by the team could be applied together with the educators at the SESI-AL schools.

## 4 Results and Discussion

To train the models for the classification algorithms, the observed optimal parameters obtained from the tests made with Grid Search, Random Search and Bayesian Discrete Search for XGBoost focused primarily on learning rate values frequently settled around 0.2–0.3, indicating that moderate learning rates provided better performance. Extremely low learning rates (e.g., 0.08) appeared less frequently, suggesting they may lead to slower convergence without significant gains. Max depth commonly ranged between 5 and 7, showing that deeper trees generally improved model capacity but were not always necessary. The number of estimators varied widely, with many models favoring 200 estimators. For KNN, the parameter search consistently indicated that the models performed best with a relatively small number of neighbors. For Random Forest, the gini criterion was more common, higher tree depths with maximum depths of 9 or 10, and larger ensembles with number of estimators going form 300 to 500 were generally preferred, suggesting that the dataset could benefit from complex decision boundaries. For Decision Tree, it generally favored higher maximum tree depth values, particularly around 17-20, across all search methods, suggesting that deeper trees performed well on the dataset.

The models metrics results were obtained through evaluation metrics such as accuracy, precision, recall, and F1-Score. Those results were obtained for every model created regarding school grade, class, subject, and milestone, which led to using the arithmetic mean for each classifier.

Table 5 presents results of average accuracy (AC) and average F1-Score (F1) obtained from the results per class generated by tests carried out with the KNN, Decision Tree, Random Forest, and XGBoost classifiers. The tables also show the standard deviation of accuracies (STDA) and the standard deviation of F1-Score (STDF1). The values obtained indicate that throughout the evaluation milestones, the metrics were improving, which is consistent with the inclusion of new variables.

The Mann-Whitney test was conducted to analyze the two independent samples. The test checks whether a variable presents values superior to the other (Hart, 2001). The Mann-Whitney test was performed to evaluate the values obtained for each metric and compare sets of classifier performance results two by two. In the case of KNN and the Decision tree, the test was inconclusive ($p = 0.1963 > 0.05$), indicating that there may be no difference between the distributions of the two samples for the metrics results. However, for KNN and Random Forest ($p = 0.0001224 < 0.05$) and KNN and XGBoost ($p = 0.0007437 < 0.05$), and so the test indicated statistical significance. Besides, for the Decision tree and Random Forest ($p = 0.01134 < 0.05$) and the Decision

Table 5: Evaluation of classifiers by milestone.

| Milestone | Classifier | AC Mean | STDA | F1 Mean | STDF1 |
|---|---|---|---|---|---|
| Milestone 1 | KNN | 79.65% | **0.0793** | 79.65% | 0.0793 |
| | Decision Tree | 78.60% | 0.1097 | 78.60% | 0.1097 |
| | Random Forest | **79.81%** | 0.0983 | 79.81% | 0.0983 |
| | XGBoost | 79.48% | 0.1032 | 79.48% | 0.1032 |
| Milestone 2 | KNN | 82.00% | 0.0672 | 82.00% | 0.0672 |
| | Decision Tree | 83.61% | 0.0628 | 83.61% | 0.0628 |
| | Random Forest | **84.65%** | **0.0519** | 84.65% | 0.0519 |
| | XGBoost | 84.52% | 0.0528 | 84.52% | 0.0528 |
| Milestone 3 | KNN | 86.29% | 0.0728 | 86.29% | 0.0728 |
| | Decision Tree | 88.21% | 0.0732 | 88.21% | 0.0732 |
| | Random Forest | 89.46% | **0.0504** | 89.46% | 0.0504 |
| | XGBoost | **89.61%** | 0.0557 | 89.61% | 0.0557 |
| Milestone 4 | KNN | 88.84% | 0.0580 | 88.84% | 0.0580 |
| | Decision Tree | 90.24% | 0.0492 | 90.24% | 0.0492 |
| | Random Forest | **92.28%** | **0.0317** | 92.28% | 0.0317 |
| | XGBoost | 91.95% | 0.0416 | 91.95% | 0.0416 |

tree and XGBoost (p = 0.04241 < 0.05), the test indicated statistical significance. However, the test was inconclusive for Random Forest and XGBoost (p = 0.588 > 0.05).

The tests show statistically significant differences between some classifiers, with a tendency for higher Random Forest and XGBoost values than the Decision tree and KNN. Random Forest and XGBoost classifiers stood out from the others, exhibiting a smaller standard deviation, indicating less dispersion concerning the arithmetic mean and, therefore, a more robust consistency in the results. XGBoost showed a lower standard deviation of approximately 6.24% concerning the arithmetic mean compared to 8.14% for Random Forest, indicating that XGBoost had a lower dispersion regarding the arithmetic means performance obtained.

Furthermore, based on the averages per milestone for all school grades, these values were used to calculate the overall average for each classifier. Then, the general averages shown in Table 6 were obtained by calculating the average of the milestones per school grade.

Table 6: Classifiers Evaluation.

| Classifier | AC Mean | STDA | F1 Mean | STDF1 |
|---|---|---|---|---|
| KNN | 85.15% | 0.0762 | 85.15% | 0.0762 |
| Decision Tree | 85.83% | 0.0689 | 85.83% | 0.0689 |
| Random Forest | **87.60%** | 0.0814 | 87.60% | 0.0814 |
| XGBoost | 86.39% | **0.0624** | 86.39% | 0.0624 |

The F1-Score values were very close to the accuracy values. This closeness is due to the balancing of the training base and the use of F1-Score Micro, calculating the score globally. As previously mentioned, the Mann-Whitney test was used to assess the significance of these differences, further confirming the consistency of the model's performance.

The results indicate that the best models were generated by the Random Forest and XGBoost classifiers, with similar results. By milestone, as shown in the table 5, Random Forest and XGBoost were the algorithms with the best performance.

The next step after obtaining the performance results was the definitive choice of the classification algorithm, which ended up being XGBoost. The performance results justify this, as well as the efficiency of its use and superiority to Random Forest in cases of overfitting, which can happen with this data since the database may not present many assessment grades for some school subjects. Furthermore, XGBoost showed a lower standard deviation of approximately 0.0624 concerning the arithmetic mean compared to 0.0814 for Random Forest, indicating that XGBoost had a lower dispersion regarding the arithmetic means performance obtained. Furthermore, XGBoost, compared to Random Forest, presents an iterative improvement of the loss function, thus making adjustments according to the creation of trees instead of maintaining fixed parameters throughout training. XGBoost performs tree pruning when gain settings stop trees from being built to avoid overfitting. In addition, XGBoost emphasizes reducing the cost of the model and is scalable and effective in preventing overfitting and creating biased patterns, which are crucial aspects to avoid.

The models created were focused only on recorded assessment grades and did not use any personal information of the students; this factor, compared to works, was differential and can make it easier to generalize the system to other schools' databases, considering that it's more accessible to have information such as school subject assessment grades than have specific information that could be restricted or create biases. Regarding the methods used to make the models, XGBoost showed good results for larger datasets. In this context, it was used in a database with more than three million grades. We constantly evaluated the models generated to limit overfitting and biases. Besides that, the database used was from basic education students, in contrast with other works that focus on higher education data, such as university-level data. The approach used involved a large dataset using only assessment records rather than integrating individual-specific information, yet the system maintained robustness despite that distinctiveness.

Table 7: Result example.

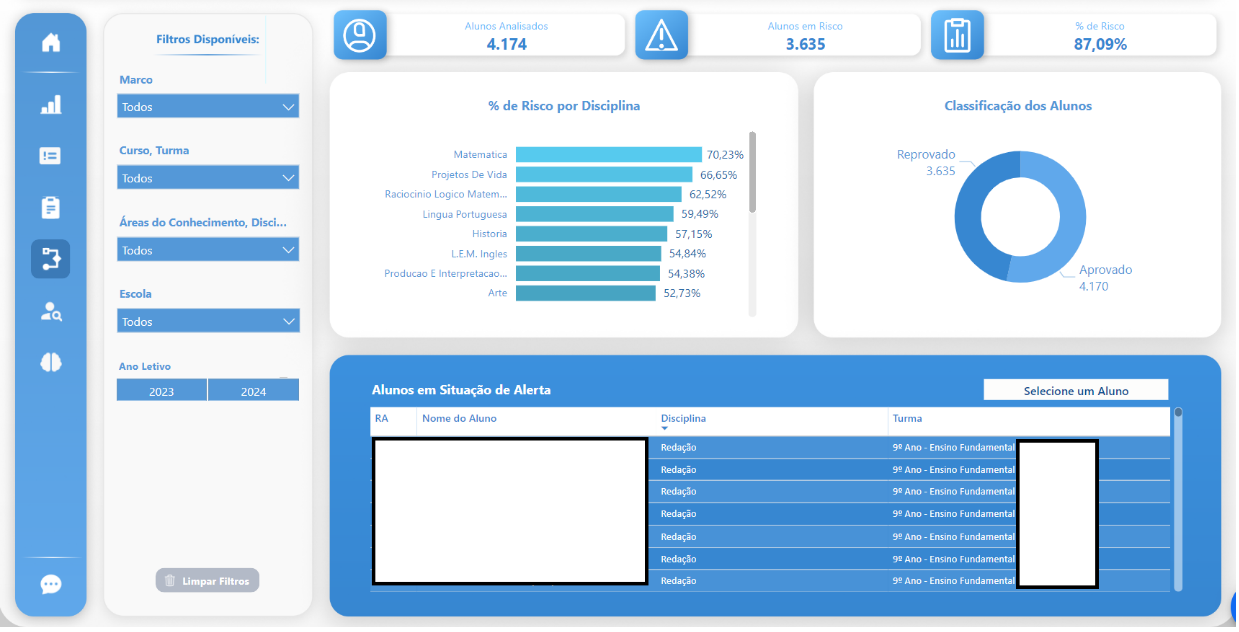
| Student ID | Grade/Class | Subject | Prediction | Prediction Probability | Milestone | Year |
|---|---|---|---|---|---|---|
| 001 | 6º Grade/D | Mathematics | Approved | 0.885 | Milestone 1 | 2023 |
| 002 | 6º Grade/C | Mathematics | Failed | 0.630 | Milestone 2 | 2023 |
| 003 | 2º Grade/A | Mathematics | Approved | 0.886 | Milestone 2 | 2023 |

An example of the results obtained by the system is shown in Table 7, presenting information related to the identification of students with a unique identification number, as well as the school class to which they belong, the school grade, the subject in which the prediction was made, the prediction obtained, the probability values associated with prediction, the evaluative Milestone associated with that prediction, and the year to which the prediction was made. The distinction of Milestones is important to recognize in what phase of the year the prediction was made based on the historical data that were used according to the Milestone to know the assessment used to make that prediction.

All the results obtained were stored in a database and inserted in a Business Intelligence Panel shown in Figure 5 where the educational team of the SESI-AL schools can access the results. In the panel, they can use filters for data filtering to analyze the results.

After the panel was implemented, the pedagogical team began using the pilot version of the system, actively engaging in its evaluation and providing ongoing feedback to refine both the panel and the implemented algorithm. Maintenance has been made regarding the automation of updates in the database, and any changes in the teaching methods in the SESI-AL schools have been tracked. This process allowed for continuous improvements, ensuring that the system evolved to meet the team's needs using the system.

The feedback given by the pedagogical team was positive. The predictions made by the system were seen by the team through the panel provided as shown in Figure 5, and they were able to filter the results by school class and subject to see the students indicated as 'Failed' and were able to analyze the environment variables that could influence in this prediction for the students. Based on that, it was said that the system helped them and the teachers in class to understand that some school classes were lacking in terms of learning and that the subjects given in class could be improved. This information proved to be crucial for timely intervention, enabling both the pedagogical team and teachers to implement corrective actions when necessary.



Filtros Disponíveis:

Marco: Todos

Curso, Turma: Todos

Áreas do Conhecimento, Disci...: Todos

Escola: Todos

Ano Letivo: 2023 2024

Alunos Analisados 4.174

Alunos em Risco 3.635

% de Risco 87,09%

Alunos em Situação de Alerta

Selecione um Aluno

| RA | Nome do Aluno | Disciplina | Turma |
|---|---|---|---|
| [redacted] | [redacted] | Redação | 9º Ano - Ensino Fundamental [redacted] |
| [redacted] | [redacted] | Redação | 9º Ano - Ensino Fundamental [redacted] |
| [redacted] | [redacted] | Redação | 9º Ano - Ensino Fundamental [redacted] |
| [redacted] | [redacted] | Redação | 9º Ano - Ensino Fundamental [redacted] |
| [redacted] | [redacted] | Redação | 9º Ano - Ensino Fundamental [redacted] |
| [redacted] | [redacted] | Redação | 9º Ano - Ensino Fundamental [redacted] |
| [redacted] | [redacted] | Redação | 9º Ano - Ensino Fundamental [redacted] |

Figure 5: Analytics Dashboard of the predictive system for failure.

## 5 Conclusion

In this work, we developed an artificial intelligence tool to predict the probability of a student failing in a school subject of Basic Education students based on assessment grades scores recorded in subjects for Middle School and Basic Secondary Education with the classification of 'Approved' or 'Failed' using machine learning to learn historical patterns. For this, more than three million assessment grades of students from the years 2017 to 2023 from SESI-AL schools were obtained through a provided database.

Predictive models were created from clean and balanced data for each subject, according to the historical patterns in the recorded grades, and learned through machine learning using supervised learning classifier algorithms to predict student performance regarding assessment grades.

The results of the chosen algorithms were evaluated using performance metrics, where arithmetic averages of performance values per subject were calculated. The Random Forest and XGBoost algorithm obtained the better results, with Random Forest with overall average accuracy and F1-Score of 87.60% and XGBoost with base parameters presented the best results, with an overall average accuracy and F1-Score of 86.3%. Furthermore, the XGBoost algorithm demonstrated less dispersion concerning the arithmetic mean of performances per subject, indicating greater consistency in the results. XGBoost was chosen as the classification algorithm for the final version of the predictive system due to its simplicity, scalability, and performance in preventing overfitting and avoiding the creation of biased patterns. Thus, a predictive system was developed using the XGBoost classification algorithm with supervised learning, which can build predictive models for each school subject in the system. This system can make 'Approved' or 'Failed' predictions for each student enrolled in the subjects and provide the probability of the prediction. The system has been deployed in the SESI-AL schools and is being used by teachers and administrators to identify students at risk of failing. The feedback regarding the system has been positive, with it being used to improve the learning curriculum in class after it was identified as lacking. With this information, educators can intervene effectively to support students and prevent failure.

## References

Alalawi, K., Athauda, R., Chiong, R., & Renner, I. (2024). Evaluating the student performance prediction and action framework through a learning analytics intervention study. *Educ. Inf. Technol.* https://doi.org/10.1007/s10639-024-12923-5 [GS Search].

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, *11*(9). https://doi.org/10.3390/educsci11090552 [GS Search].

Al-Tameemi, R. A. N., Johnson, C., Gitay, R., Abdel-Salam, A.-S. G., Hazaa, K. A., BenSaid, A., & Romanowski, M. H. (2023). Determinants of poor academic performance among undergraduate students—a systematic literature review. *International Journal of Educational Research Open*, *4*, 100232. https://doi.org/10.1016/j.ijedro.2023.100232 [GS Search].

Amal Asselman, M. K., & Aammou, S. (2021). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235 [GS Search].

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, *13*, 281–305. https://dl.acm.org/doi/abs/10.5555/2188385.2188395 [GS Search].

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2), 1–43. https://doi.org/10.1002/widm.1484 [GS Search].

Byrd, R. S. (2005). School failure: Assessment, intervention, and prevention in primary pediatric care. *Pediatr. Rev.*, *26*(7), 233–243. https://doi.org/10.1542/pir.26-7-233 [GS Search].

Chanlekha, H., & Niramitranon, J. (2018). Student performance prediction model for early-identification of at-risk students in traditional classroom settings. *MEDES '18: Proceedings of the 10th International Conference on Management of Digital EcoSystems*, 239–245. https://doi.org/10.1145/3281375.3281403 [GS Search].

Crosnoe, R. (2002). High school curriculum track and adolescent association with delinquent friends. *Journal of Adolescent Research - J ADOLESCENT RES*, *17*, 143–167. https://doi.org/10.1177/0743558402172003 [GS Search].

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230 [GS Search].

Giavrimis, P., & Papanis, E. (2008). Sociological dimensions of school failure: The views of educators and students of educational schools. *The Journal of International Social Research*, *1*(5), 326–354. https://www.sosyalarastirmalar.com/articles/sociological-dimensions-of-school-failure-the-views-of-educators-and-students-of-educational-schools.pdf [GS Search].

Hart, A. (2001). Mann-Whitney test is not just a test of medians: Differences in spread can be important. *BMJ*, *323*(7309), 391–393. https://doi.org/10.1136/bmj.323.7309.391 [GS Search].

Javed, D., Jhanjhi, N. Z., Ashfaq, F., Khan, N. A., Das, S. R., & Singh, S. (2024). Student performance analysis to identify the students at risk of failure. *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, 1–6. https://doi.org/10.1109/ETNCC63262.2024.10767511 [GS Search].

Joseph, V. R. (2022). Optimal ratio for data splitting. *Stat. Anal. Data Min.*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583 [GS Search].

Kamal, M., & Bener, A. (2009). Factors contributing to school failure among school children in very fast developing arabian society. *Oman Med. J.*, *24*(3), 212–217. https://doi.org/10.5001/omj.2009.42 [GS Search].

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: A big comparison for NAS. https://doi.org/10.48550/arXiv.1912.06059 [GS Search].

Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting student success using big data and machine learning algorithms. *Int. J. Emerg. Technol. Learn.*, *17*(12), 236–251. https://doi.org/10.3991/ijet.v17i12.30259 [GS Search].

Rodrigues, L. S., Santos, M., Gomes, C. F. S., Choren, R., Goldschmidt, R., & Barbará, S. (2024). Transformers para previsão de desempenho acadêmico no ensino fundamental e médio. *Revista Brasileira de Informática na Educação*, *32*, 213–241. https://doi.org/10.5753/rbie.2024.3661 [GS Search].

Siddique, A., Jan, A., Majeed, F., Qahmash, A., Quadri, N. N., & Wahab, M. (2021). Predicting academic performance using an efficient model based on fusion of classifiers. *Applied Sciences*, *11*, 11845. https://doi.org/10.3390/app112411845 [GS Search].

Silva, P., Souza, F., & Fagundes, R. (2020). Approaches to predicting educational problems: A systematic mapping. *Proceedings of the XVI Brazilian Symposium on Information Systems*. https://doi.org/10.1145/3411564.3411657 [GS Search].

Simanca, F., Gonzalez Crespo, R., Rodríguez-Baena, L., & Burgos, D. (2019). Identifying students at risk of failing a subject by using learning analytics for subsequent customised tutoring. *Appl. Sci. (Basel)*, *9*(3), 448. https://doi.org/10.3390/app9030448 [GS Search].

Singh, R., & Pal, S. (2020). Machine learning algorithms and ensemble technique to improve prediction of students performance. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*, 3970–3976. https://doi.org/10.30534/ijatcse/2020/221932020 [GS Search].

Sunny, B. S., Elze, M., Chihana, M., Gondwe, L., Crampin, A. C., Munkhondya, M., Kondowe, S., & Glynn, J. R. (2017). Failing to progress or progressing to fail? Age-for-grade heterogeneity and grade repetition in primary schools in Karonga district, northern Malawi. *International Journal of Educational Development*, *52*, 68–80. https://doi.org/10.1016/j.ijedudev.2016.10.004 [GS Search].

UNICEF Brasil, Instituto Claro & Cenpec. (2021, January). Enfrentamento da cultura do fracasso escolar: Reprovação, abandono e distorção idade-série [UNICEF Brasil]. https://www.unicef.org/brazil/relatorios/enfrentamento-da-cultura-do-fracasso-escolar

Vujovic, Z. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670 [GS Search].

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061 [GS Search].