Received: 1 August 2023 | Revised: 12 October 2023 | Accepted: 20 October 2023 | Published online: 30 October 2023

RESEARCH ARTICLE

Insights into Factors Influencing Academic Success: An Application of Classification Models in Higher Education

Artificial Intelligence and Applications
yyyy, Vol. XX(XX) 1–5
DOI: 10.47852/bonviewAIA32021465

BON VIEW PUBLISHING

Liliana T. Balderrama[1] and Danielli A. Lima[1,*]

1 Computational Intelligence and Robotics Laboratory (LICRO), Federal Institute of Education, Science and Technology of Triângulo Mineiro (IFTM) Campus Patrocínio MG, Brazil

Abstract: Understanding the factors that impact students' achievements and failures in higher education is crucial as it enables the development of targeted interventions and support mechanisms that can enhance academic performance and foster student success. Artificial intelligence can help understand the factors that influence students' academic achievements and failures in higher education by analyzing large volumes of data, identifying patterns and correlations, and providing valuable insights. This article presents the application of eight classification models on a "Dataset of Academic Performance Evaluation of Higher Education Students" consisting of 145 higher education students. The aim is to identify the best classification algorithm for predicting academic performance. The correlation filter was used for the discovery and selection of relevant attributes, resulting in the choice of four attributes for analysis. The best classification models were Random Forest, Support Vector Machine, and Decision Tree, with an average accuracy of 94.37% and a correlation filter of 0.1. These results demonstrate that the application of artificial intelligence and machine learning techniques is important for decision-making in higher education, allowing for a better understanding of the factors that influence academic success or failure. The study emphasizes the importance of careful attribute selection and the use of appropriate classification algorithms to ensure accuracy and reliability of the results. Additionally, the study was replicated and evaluated with nine Brazilian higher education students, achieving an accuracy of 88.89%. These results demonstrate the consistency and relevance of the proposed attribute filtering model.

Keywords: machine learning, higher education, students' performance, correlation filter, student attributes, classification algorithms

***Corresponding author:** Danielli A. Lima,Computational Intelligence and Robotics Laboratory (LICRO), Federal Institute of Education, Science and Technology of Triângulo Mineiro (IFTM) Campus Patrocínio MG, Brazil. Email: danielli@iftm.edu.br

1. Introduction

It is known that education is one of the most important means for the development of a society. It is exercised in a way that the individual develops their skills, adapting to society (Sendacz et al., 2022). Education is a social practice that aims at the development of the human being, their potentialities, skills and competences (Yılmaz & Sekeroglu, 2019). The education, a right for all and a duty of the State and the family, shall be promoted and encouraged with the collaboration of society (Akazaki et al., 2020), aiming at the full development of the individual, their preparation for the exercise of citizenship, and their

qualification for work. However, despite being supported by federal laws, most students still leave school with many deficiencies in knowledge (Matos et al., 2019; de Oliveira et al., 2008).

In this work, we will analyze a recent real-world database of higher education in foreign universities. The data was collected from the "Higher Education Students Performance Evaluation Dataset" (Yılmaz & Sekeroglu, 2019) from UCI Machine Learning Repository2, an open and online repository for data science (DS). It is known that DS, especially machine learning (ML) algorithms, can be used in education to personalize learning experiences for students based on their individual needs and performance (Wang, 2022; Yılmaz & Sekeroglu, 2019; Lima & Fagundes, 2020; Moonsamy et al., 2021). By analyzing large datasets of educational data, ML models can help identify patterns and relationships between variables, allowing for more effective decision-making in education (Dornelas & Lima, 2023; de Castro Soares et al., 2023; Aldowah et al., 2019; Romero & Ventura, 2020; Fernandes et al., 2019).

The goal of this paper is to predict the performance of higher education students based on some variables using data mining (DM) algorithms, an artificial intelligence (AI) field. In this sense, we aim to employ ML techniques for the purpose of modeling and predicting the academic success and failure of higher education students, based on a scientific analysis conducted by DS. Furthermore, our objective is to identify a concise and minimal set of attributes can accurately predict students' academic success and failure, using correlation filters (CF), aiming to provide valuable insights for educators and teachers, contributing to the enhancement of public education.

Predicting student behavior through some parameters is important for designing strategies to avoid dropouts and failures in graduation. That means that the results of this re- search have the potential to positively impact higher education institutions (HEI) by providing relevant information for decision-making and educational planning. It is hoped that these findings can contribute to the improvement of academic performance of university students, promoting the quality of education and the educational development of the country, including Brazil higher education. We collected data from students in a public institution through a questionnaire to verify if the proposed model is capable of correctly classifying the dataset.

This article is organized as follows: in section 1, we have the introduction and objectives. In section 2, we provide a brief literature review on the use of ML in education and its potential to predict academic success and failure. Section 3 presents the methodology used in this study, including data collection and preprocessing, feature selection, and model development. In section 4, we present the results of our experiments, including the accuracy of our models and the identification of the most important features for predicting academic success and failure. Section 5 discusses the implications of our findings for educators and higher education institutions, as well as some limitations and future research directions. Finally, in section 6, we conclude the paper by summarizing our main contributions and discussing some implications for future research and practice.

2. Background

In this section, we will present the theoretical foundation. First, we will present some detailed definitions of the database. Then, we will comprehensively present the definitions and tools for machine learning (ML). Finally, we will thoroughly present the extensive works related to this research.

2.1. Database description

In this work we will use a database (DB) with a feature vector of size 32 (31 parameters and 1 class), which were collected from 145 students from the Faculty of Engineering and Faculty of Education Sciences in 2019 (Yılmaz & Sekeroglu, 2019). During the determination of academic success, writ- ten exams, tests and oral exams are considered in the cognitive achievement of students and scales are generally used for components. The students were evaluated on concepts, which represent the classes abstracted from the problem, (0): Fail, (1): DD, (2): DC, (3): CC, (4): CB, (5): BB, (6): BA and (7): AA.

However, we have abstracted these 8 classes into two useful classes: R (0): reproved/fail and A (1): approved/pass. Many of the educational approaches that are not based on certain criteria of data and weights cause subjectivity during the evaluation process and, thus, false evaluations can be carried out (Yılmaz & Sekeroglu, 2019). Each of these attributes can directly interfere with the success and failure rate of students in higher education.

In this case, we have the following information on the 32 parameters of the dataset. The questions related to the data (Q1 - Q10) are the personal questions: (1) age, (2) gender, (3) type of high school attended, (4) type of scholarship, (5) additional employment, (6) sports and arts, (7) relationship, (8) salary, (9) transportation, (10) accommodation.

The questions related to the data (Q11 - Q16) are the family questions: (11) mother's level of education, (12) father's level of education, (13) number of siblings, (14) parents' relationship, (15) mother's occupation, (16) father's occupation.

The questions related to the data (Q17 - Q30) are the educational questions: (17) daily study hours, (18) non-scientific reading, (19) scientific reading, (20) participation in seminars and conferences, (21) effect of projects and activities, (22) participation in readings, (23) study type (group | individual), (24) study type (regular | last week), (25) grade- taking, (26) writing and reading, (27) effect of classroom discussion, (28) effect of flipped classroom, (29) GPA last semester, (30) expected CGPA at graduation, (31) course ID and (32) Output CLASS Grade: A approved or R failed.

2.2. Machine learning

Machine learning can be understood as machines with the ability to learn by themselves from volumes of data, recognizing patterns and creating relationships between them, this field of study is a subset of Artificial Intelligence algorithms (IA) (Ferreira et al., 2019; Wongvorachan et al., 2023). In this work, we will use ML with the objective of predicting the final performance of higher education students. One of the most common uses for ML techniques is to predict or classify new situations within the same context, bringing new information (Yılmaz & Sekeroglu, 2019, Witten & Frank, 2002).

In this work, we will focus on supervised learning, whose fundamental objective is to learn a function that maps an input to an output based on examples of input-output pairs (Wang, 2022). Supervised learning methods try to infer a function from labeled training data consisting of a set of training examples, systematizing and analyzing data that bring new findings (Ferreira et al., 2019). There are several supervised learning techniques, among the supervised learning techniques we have: Decision Tree Learner (DT), K- Nearest Neighbor (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Nayve Bayes Learner (NBL), Gradient Boosted Learner (GBL), Randon Forest Learner (RF) and Multilayer Network with Fast Backpropagation (RProp). These 8 algorithms are commonly used in classification problems in ML and each has its own advantages and disadvantages. Understanding how these algorithms work can be helpful in selecting the best algorithm for a given classification problem.

DT This is a ML algorithm that builds decision trees from a set of training data. The idea behind the algorithm is to divide the dataset (recursively splitting) into smaller and more homogeneous subsets based on a certain criterion, such as entropy or Gini impurity. It's commonly used for classification and regression problems.

KNN This algorithm classifies new data points based on their proximity to the k nearest data points in the training set. It's a simple and versatile algorithm used for classification, regression, and clustering tasks.

SVM This is a popular algorithm for binary classification that finds the hyperplane that best separates the two classes in the dataset. It's effective for handling complex, high-dimensional data.

MLP This is a type of neural network with multiple layers of interconnected neurons that can learn complex pat- terns in data. It's a powerful algorithm for classification and regression tasks but can require significant computational resources.

NBL This is a probabilistic algorithm based on Bayes' theorem that assumes the features in the dataset are independent of each other. It's a simple and efficient algorithm commonly used for text classification and spam filtering.

GBL This algorithm builds an ensemble of weak decision trees that are sequentially trained to correct the errors of the previous trees. It's a powerful algorithm for classification and regression tasks that can handle complex datasets.

RF This is another ensemble learning algorithm that builds a forest of decision trees, where each tree is trained on a random subset of the data and features. It's effective for handling noisy and high-dimensional data, and can prevent overfitting.

RProp This is another type of neural network that uses a faster and more efficient variant of the backpropagation algorithm to update the weights of the network during training. It's effective for handling high-dimensional data and can learn complex patterns in data.

Additionally, to run each of these algorithms efficiently, there are different tools for data mining: R, WEKA, Python, DataMelt, Trifacta Wrangler and in this work we will use KNIME® Analytics Platform, which is a tool that we will use in this work. KNIME® is open-source software for all your data needs, it is free to download and free to use. The tool contains all the main ML techniques based on visual programming and was chosen for being free and easy to use (Ferreira et al., 2019).

2.3. Related works

In work of Sekeroglu et al. (2021), authors conducted a systematic literature review to analyze student performance from 2010 to 2020. The results are presented as percentages and categorized as model, dataset, validation, evaluation, or aims. Commonalities and differences were identified, critical gaps were presented, and possible remedies were proposed. The study suggests the need for standardized evaluation and validation strategies, as well as a shift towards deep learning models in future research. Additionally, the advantages of a global education information consortium are high- lighted.

According to Yılmaz and Sekeroglu (2019), personal information, educational preferences, and family properties are some of the main indicators for improve students' performance. In work of Yılmaz and Sekeroglu (2019), authors applied artificial intelligence techniques to questionnaire results from three different courses across two faculties to classify students' final grade performances and determine the most efficient ML algorithm for this task. Through several experiments performed by the Yılmaz and Sekeroglu (2019), it was found that the Radial-Basis Function Neural Network can effectively classify student performance with an accuracy of 70% - 88%. In work by Wang (2022), the importance of subject development in higher education is discussed, and its relationship with student academic performance is evaluated using ML and AI techniques. A new AdaBoost Adaptive-Bidirectional Associative Memory (AA-BAM) network model is introduced, which uses Hebbian supervisedlearning to train and up- date the model parameters. The memory cell in the neural model stores processing information and recalls output pat- terns, identifying student academic performance and analyzing subject development quality in institutions. The system achieves 98.78% accuracy, considering all attributes, indicating a high correlation between subject development and student academic performance.

Silva Filho et al. (2023) introduce a frame- work for applying data mining and machine learning techniques to assess the importance of contextual features in predicting educational outcomes. This framework incorporates score-based feature contributions and informative metrics to enhance interpretability. Applied to a Brazilian Large-scale assessment spanning 2009-2019, the framework reveals new insights while validating existing hypotheses. The study underscores the influence of well-established factors on school performance and highlights the potential for exploring at- tributes related to school infrastructure and faculty for educational policy formulation.

Similarly, Cortez and Silva (2008) conducted a study aimed at addressing high student failure rates in core subjects, such as Mathematics and Portuguese, in Portugal. Despite improvements in the population's educational level, the country lagged behind in European educational statistics. Their study applied Business Intelligence (BI) techniques, assessed four ML models and three input selections for insights from raw educational data. It achieved high predictive accuracy, particularly with initial school period grades, and identified influential factors like absences, parental employment and education, and alcohol consumption in student performance. As a result, it suggests improving prediction tools for better education quality and resource management.

3.Methodology

In this proposal section, the workflow applied in KNIME will be presented to identify the most important attributes in detecting the characteristics of success and failure of students, through school, family and educational data. In this work, the collected data was previously pre-processed and tabulated in a .csv file, which is why we only used correlation filters to select the main features. A table with the six correlation filters ($C_F = 6$) used will also be displayed. The selection of these filters is crucial to determine the most relevant attributes for predicting students' academic performance. In this way, it is intended to provide a clear view of the methodological process of selection of the most relevant attributes and of the filters used for this choice.

3.1. Proposed workflow

Using the KNIME® Analytics Platform for data analysis and report building, 3 of the 8 workflows were used as example: DT, SVM and KNN, as shown in Figure 1. Initially, data of type (.csv) was read by CSV Reader. Later, some columns were filtered with the Column Filter, to filter the data of the 8 classes (Yılmaz & Sekeroglu, 2019) and we used only the two classes passed (1) and failed (0). Later we made some data visualizations using the Color Manager and Color Learner nodes for data coloring. The charts we per- form Pie Chart, Box Plot and Histogram for data visualizations. Another node used was Linear Correlation and Correlation Filter for filtering where most correlated columns survive while all correlated columns are filtered.

The X-Agregator node is the first in a cross-validation loop. At the end of the loop there should be an X-Aggregator to collect the results of each iteration. All nodes between these two nodes run as many times as iterations should run, in this case the value of $x = 20$ iterations. In this example, for reasons of page limit, we present only 3 nodes of the 8 algorithms for the ML (SVM, KNN and DT), in this case, the nodes SVM Learner and SVM Learner serve to make the prediction of the SVM algorithm,

while the nodes Decision Tree Learner and Decision Tree Learner are used to predict and classify data using the DT algorithm. The K Nearest Neighbor node does the KNN prediction, with k = 3 nearest neighbors. For this work, the other 5 remaining algorithms were also tested using KNIME nodes (MLP, NBL, GBL, RF and RProp), but for space reasons they were not included in the Figure 1.

Finally, the Scorer node evaluates each of the prediction algorithms that were used here by means of accuracy and a confusion matrix is presented. Accuracy (A) is the percent- age of hits (both true positives and true negatives) on all bets (TP = true positive; FP = false positive; TN = true negative; FN = false negative) of the algorithm, according to Equation 1.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The error is the cases in which the algorithm fails. That is, the error (E) is calculated considering the difference of the total minus the accuracy value, given by Equation 2.

$$E = 1.0 - A \quad (2)$$

3.2. Correlation filter

In this article, we will explore the selection of the most relevant attributes to detect the characteristics of students' academic success and failure based on the 32 attributes divided into: personal, family and educational data, as shown in Table 1. For this, we will use 6 correlation filters ranging from (1.0, 0.5, 0.4, 0.3, 0.2, 0.1) to select only the most relevant attributes that are still capable of generating good ac- curacy for the analysis. This filtering will allow us to identify which characteristics have the greatest impact on students' school performance, enabling a better understanding of the factors that influence academic success or failure.

4. Results

In this section, the results of the analysis of students' school, family and educational data will be presented through the selection of relevant attributes and the application of data mining algorithms. This analysis was divided into two subsections: the statistical analysis and the classification analysis. In the first subsection, we perform a descriptive analysis of the data, including summary statistics and graphs to visualize data distributions. In the second subsection, we apply 8 data mining algorithms to classify students into two categories: academic success and failure. This classification was based on a selection of the most relevant attributes, chosen through different correlation filters. The results obtained are presented and discussed in detail in these subsections.

4.1. Analysis by descriptive statistics

In this section the data collected from the data visualization will be presented. In this sense, in the pie charts of Figure 2 we have the results referring to: (i) the number of students who failed (5.52%) and the number of students who passed (94.48%), see Figure 2(a). Additionally, (ii) the gender of the students interviewed: (1) 40% female and 60% male, see Figure 2(b). Finally, in Figure 2(c), we have the 9 undergraduate courses.

In Figure 3 we have the range and median of some of the attributes listed in the boxplots, and in Figure 4 the histogram grouped by classes A (1) and R (0), and the attributes of type of high school, average of previous grades, transportation to the university are the ones that most interfere on average in the student performance. While the attributes of type of scholarship and presence in classes seem to be the attributes that least interfere, as they have similar averages in both cases, as shown in Figure 4.

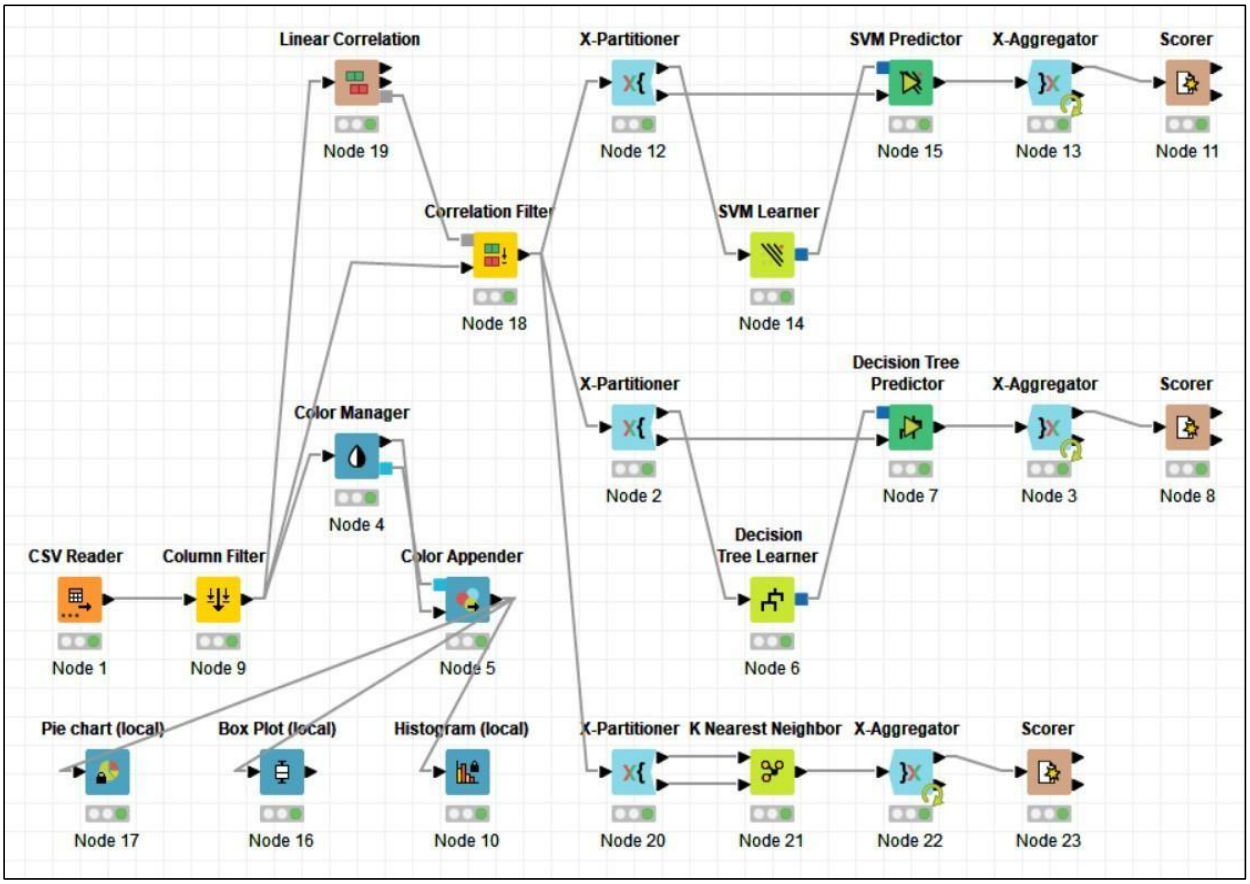
4.2. Analysis of classification algorithms

This article presents results of a study on the selection of the most relevant attributes for the classification of academic success and failure in higher education students. The results showed that the best classification algorithms were those with the

lowest number of attributes filtered by CF , which led to greater classification accuracy, as can be seen in Table 2. In particular, the DT, RF and SVM algorithms had an accuracy of 94.483% when the correlation filter was set to CF = 0.1. The parameters selected by the correlation filter indicated that Father Education, Parental Status, Discussion improves interest, and Course ID were the most relevant parameters for classifying success and failure in higher education. These parameters are related to the students' education and home environment, as well as their motivation and interest in the course.

On the other hand, the worst performance was obtained by the NBL algorithm, which had an accuracy of only 74.483% when no parameter was filtered by the correlation filter (1.0). This can be justified by the fact that NBL is a simple classification algorithm that is not able to deal well with many attributes or with data with high dimensionality. In general, the results obtained in this study suggest that the selection of relevant parameters is crucial to obtain good accuracy in classifying academic success and failure in higher education students. In addition, the results highlight the importance of students' education and home environment for their academic performance. The last line of Table 2 presents the average of the performance metrics of all models considered, and the algorithm with the best average was RF, with a value of 94, 368%, better than Yılmaz and Sekeroglu (2020) who achieved accuracies between ≈70% - 88%, and better in terms of processing than Wang (2022), who achieved ≈98.78% accuracy, however, considers many more database ' s attributes, unlike this work, in which only 4 attributes were needed. We can see that RF performed very well in all evaluated metrics, ranking first in several cases. Furthermore, its mean in the last row of the table is the highest among all the models, which indicates that RF is a good choice for the classification task of the given data set.

Figure 1
Workflow in KNIME used for data visualization and classification



4.3. Analysis of results collected through questionnaires

Given the above, we present the collected results from 9 students in higher education at a public institution. The collected data from these 9 students shows, for example, that the students' ages range from 2 to 4, with an average of approximately 3.33

(around 25 years) and a standard deviation of about 0.71. Most students are male, with an average of 1.56 and a standard deviation of 0.53. Addition- ally, it is observed that the mean and standard deviation for variables such as scholarship type, additional work, regular artistic or sports activities, presence of a partner, and type of accommodation have similar values. The variables related to parents' education, such as mother's and father's education, also have close means and standard deviations, indicating a relatively homogeneous distribution of data. The variable "Number of siblings" has an average of 2.44 and a standard deviation of 1.74, suggesting a larger variation in this characteristic. The variables related to academic performance, such as weekly study hours, reading frequency, and exam grades, are also analyzed based on their descriptive statistics.

The data does not contain missing values (NaN) or infinite values (+∞ or -∞), indicating that all information was recorded for every entry. Although the median is not provided in the statistics, we can infer that the data is close to the mean, as the standard deviation is relatively low compared to the means. The total number of students is 9 for all columns, suggesting a relatively small dataset.

In this regard, we will initially present the workflow for prediction in this task, as depicted in Figure 5. First, we separated the data into training and testing sets, with the training data being the data from the UCI Machine Learning Repository, and the testing data being the data from the 9 interviewed students. It is noteworthy that the training data was used to train the DT algorithm, which achieved the best result considering the minimum number of attributes. In this context, the data was created following the same pattern as the data from the studied dataset (collected in the UCI). Additionally, we conducted a score analysis, and the obtained result for these 9 interviewed students was 88.89%, suggesting that the proposed model has a high potential to predict student success or failure based on a few attributes.

Table 1
Resulting attributes after applying the correlation filters

Correlation Filter	Number of Attributes		Considered attributes
	Included	Excluded	
1.0	32	0	Age, Sex, Graduated high-school type, Scholarship type, Additional work, Regular artistic or sports activity, Partner, Salary, Transportation, Accommodation type, Mother Education, Father Education, Number of simblings, Parental status, Mother occupation, Father occupation, Weekly study hours, Reading frequency non-science, Reading frequency scientific, Attendance to the seminars, Impact of your projects, Attendance to classes, Preparation to midterm exams 1, Preparation to midterm exams 2, Taking notes in class,Listening in classes, Discussion improves my interest, Flip-classroom, Cumulative GPA, Expected GPA, Course ID, Output_CLASS
0.5	31	1	Age, Sex, Graduated high-school type, Scholarship type, Additional work, Regular artistic or sports activity, Partner, Salary, Transportation, Accommodation type, Mother Education, Father Education, Number of simblings, Parental status, Mother occupation, Father occupation, Weekly study hours, Reading frequency non-science, Reading frequency scientific, Attendance to the seminars, Impact of your projects, Attendance to classes, Preparation to midterm exams 1, Preparation to midterm exams 2, Taking notes in class,Listening in classes, Discussion improves my interest, Flip-classroom, Cumulative GPA, Course ID, Output_CLASS

0.4	28	4	Transportation, Accommodation type, Mother Education, Number of simblings, Parental status, Mother occupation, Father occupation, Weekly studyhours, Reading frequency non-science, Reading frequency scientific, Imp- act of your projects, Attendance to classes, Preparation to midterm exams 1,Preparation to midterm exams 2, Taking notes in classes, Listening in class, Discussion improves my interest, Flip-classroom, Cumulative GPA, Course ID, Output_CLASS
0.3	19	13	Graduated high-school type, Additional work, Partner, Total salary, Transportation, Mother Education, Parental status, Father occupation, Weekly study hours, Reading frequency non-science, Impact of your projects, Attendance to classes, Preparation to midterm exams 1, Taking notes in classes, Listening in classes, Discussion improves my interest, CumulativeGPA, Course ID, Output_CLASS
0.2	11	21	Age, Partner, Artistic or Sports, Partner, Salary, Number of simblings, Father occupation, Impacty of projects, Attendance to classes, Preparation to Midterm 2, Discussion Interest, Output_CLASS
0.1	5	27	Father Education, Parental Status, Discussion improves interest, Course ID, Output_CLASS

Table 2
Accuracy for each of the classification models with the respective filters

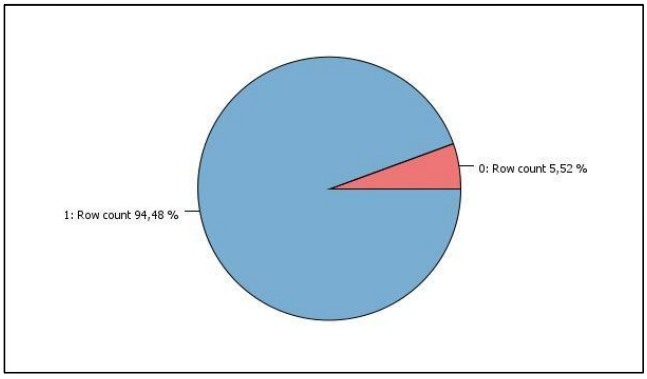
Correlation Filter	Number of Attributes		Classification Algorithms							
	Included	Excluded	DT	KNN	SVM	MLP	NBL	GBL	RF	RProp
1.0	32	0	90.345	93.103	86.207	94.483	74.483	91.724	94.483	91.724
0.5	31	1	90.345	93.103	86.897	90.345	75.862	93.103	94.483	93.103
0.4	28	4	93.793	93.103	91.034	91.034	90.345	91.724	94.483	92.414
0.3	19	13	92.414	94.483	94.483	88.276	86.207	94.483	94.483	90.345
0.2	11	21	91.724	94.483	94.483	91.724	90.345	92.414	93.793	91.034
0.1	5	27	94.483	93.103	94.483	93.103	90.345	91.724	94.483	91.034
Means	21	11	92.184	93.563	91.265	91.494	84.598	92.527	94.368	91.609

Figure 6 presents the results after data collection. In Figure 6(a), it is evident that only 1 student (11.11%) failed, while 88.89% of the students (8 students) did not fail. Furthermore, in Figure 6(b) it is possible to see the decision tree that was tested with the 9 students. The results from the collected data indicated 88.89% accuracy.

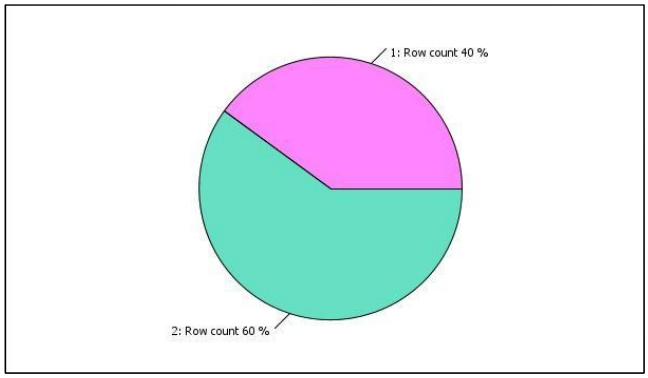
From these analyses, it is possible to see that the data collected reveals a series of challenges and gaps in student education. The average hours of study per week is relatively low, indicating a possible lack of dedication and commitment by students to

learning. The averages for reading frequency, participation in seminars and impact of projects are also lower than expected, suggesting a lack of student involvement and interest in these academic activities.

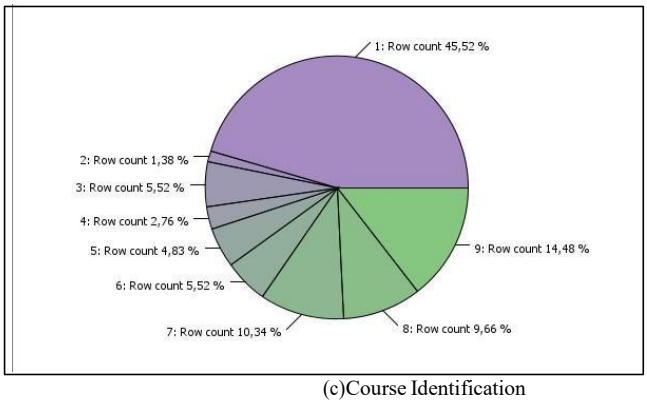
Figure 2
Pie charts made from the 145 student interviews



(a)Approved and failed classes



(b)Gender



(c)Course Identification

5. Discussion

Based on the data analysis using descriptive statistics and data mining, the parameters of Father's Education, Parental Status, and Discussion that increase interest and Course ID identified by the correlation filter as relevant for predicting academic success and failure in higher education are important for different reasons (Yılmaz & Sekeroglu, 2019).

The level of father's education can have a significant impact on the student's academic performance, as parents' education can influence the opportunities available to the student and their home environment, including parental motivation and expectations regarding the child's academic performance (Cortez & Silva, 2008). Students whose parents have a higher level of education generally have greater access to educational resources, such as books and teaching materials, as well as greater emotional and financial support for their studies (Yılmaz & Sekeroglu, 2019).

The attribute of Parental Status can also influence the student's academic success, as parents who work in jobs with flexible schedules or that allow them to be more present in their children' s lives, such as working from home, may have a positive impact on academic performance. Addition- ally, parents who are actively involved in their children's academic life can stimulate their interest and dedication to studies (Guerra et al., 2020).

The attribute of discussion improving interest and participation can be an indicator of students' engagement and interest in the subject, which can be a determining factor for their academic success. Students who are more engaged in class, asking questions, and actively participating in discussions may have a better understanding of the content and develop critical thinking skills (Wang, 2022).

The Course ID, or course identification, can influence academic performance, as different courses have different levels of difficulty, workload, and learning objectives (Cortez & Silva, 2008). Additionally, the course identification may also be related to other factors, such as the quality of the faculty, infrastructure, and resources available for the subject (Wang, 2022). The main limitations of this work lie in the fact that, when applying the questionnaire, only 9 students were interviewed for testing the model found. How- ever, when using the dataset, we achieved excellent performance, totaling 89.9% of accuracy. Thus, considering these parameters, it is possible to identify factors that may impact students' academic success and, consequently, allow schools and educational institutions to take measures to pro- mote a more effective and personalized learning environment for students. Subsequently, educational institutions can implement measures to foster a more efficient and customized learning environment for their students, empowering schools and educational institutions to adopt strategies aimed at creating a more effective and individually tailored learning atmosphere for their students.

Figure 3
Boxplots showing ranges and medians of the attributes

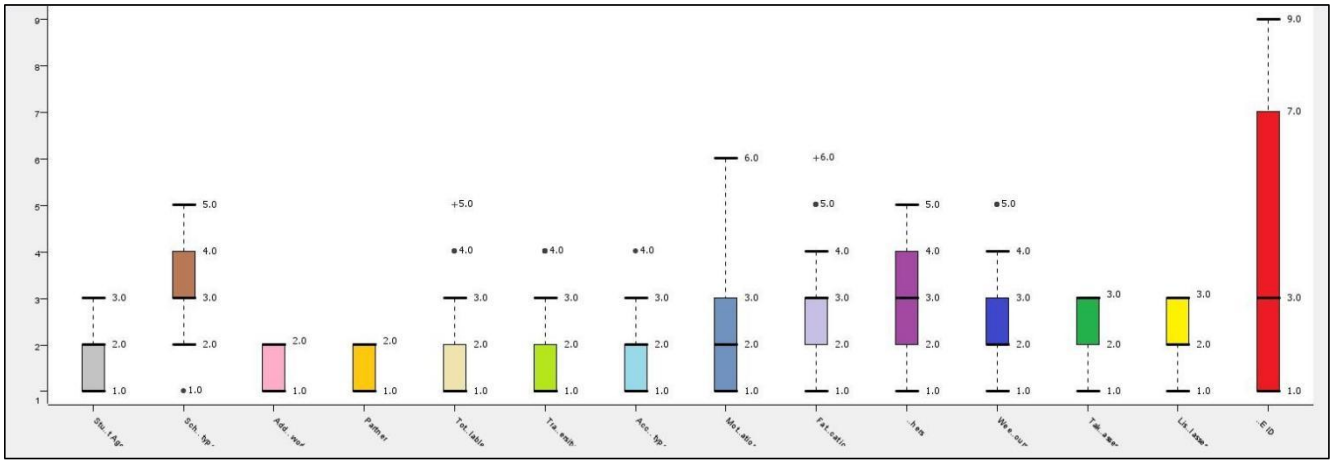


Figure 4
Histograms of the means of the attributes that lead to school success and failure

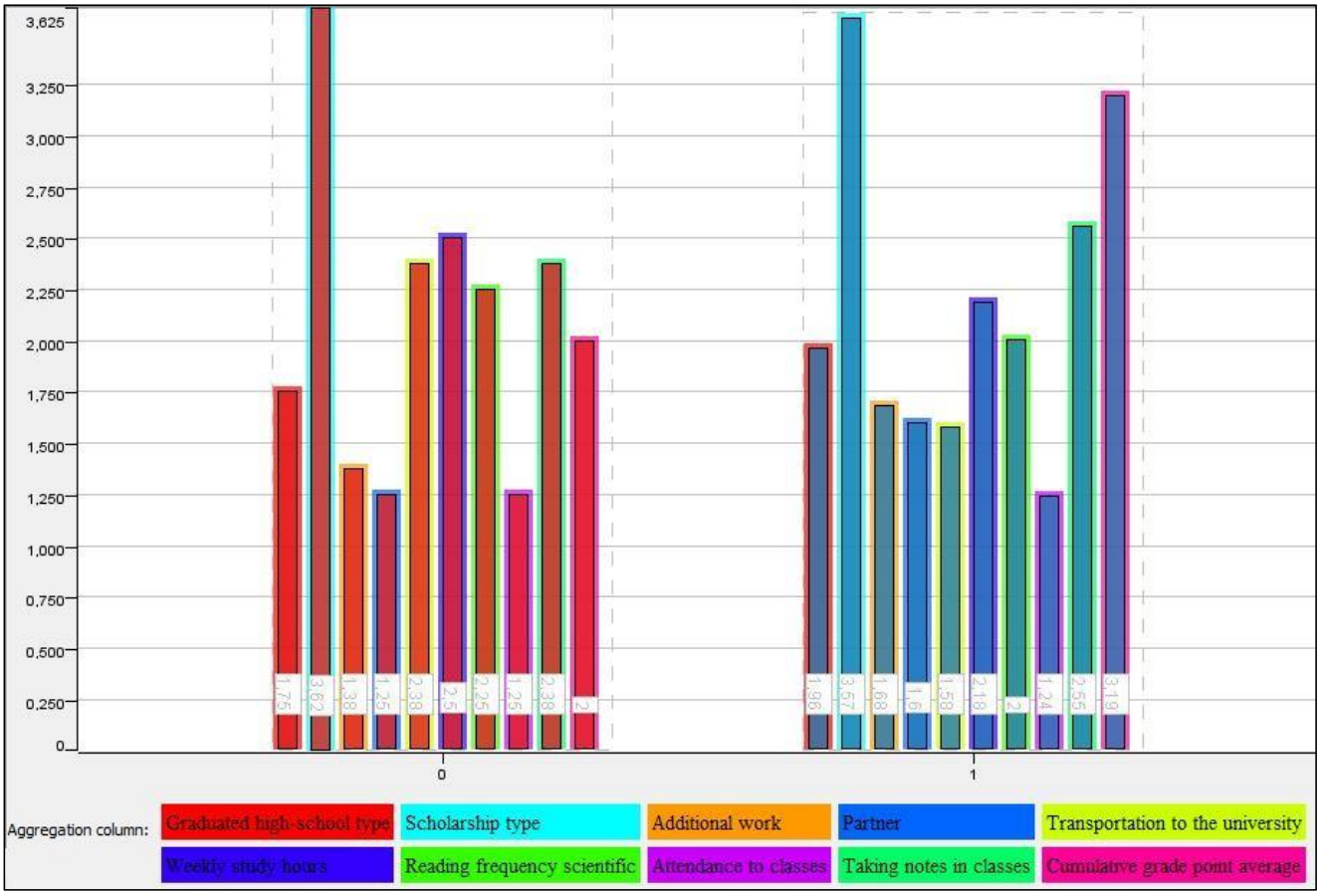


Figure 5
Workflow in KNIME to predict the success or failure of interviewed students

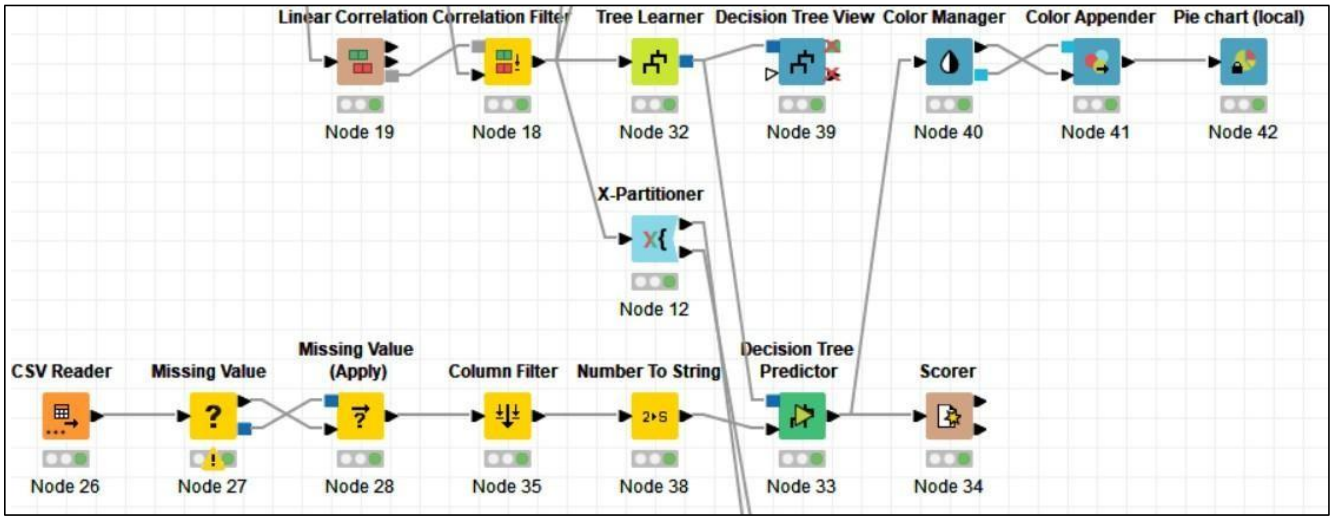
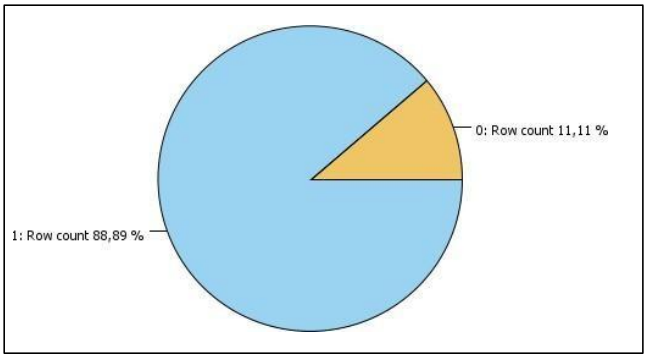
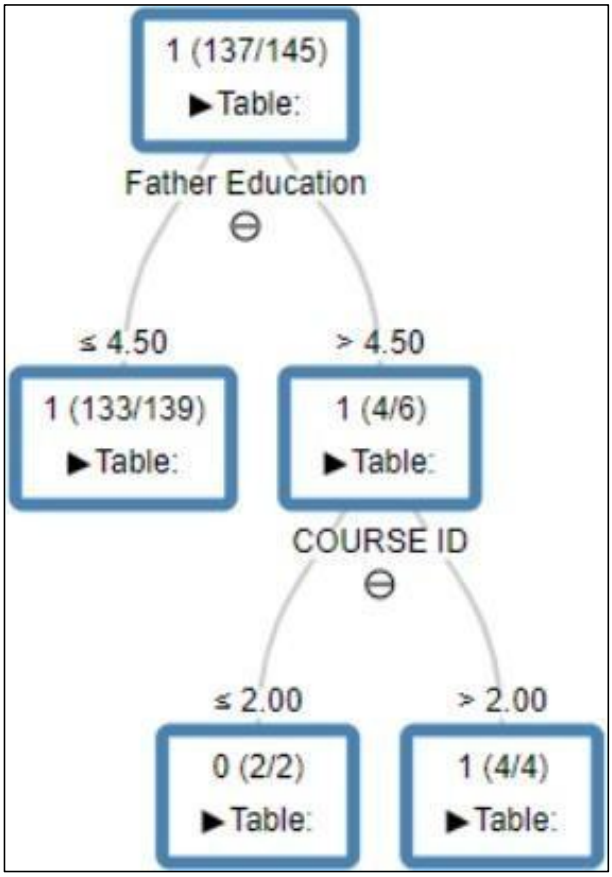


Figure 6
Results obtained from the KNIME Analytics Platform after collecting data from students



(a) Pie chart for class



(b) Decision tree

6. Conclusions

It is concluded that this study utilized the dataset "Student Performance Evaluation Dataset of Higher Education" to identify the most relevant attributes for predicting the academic performance of higher education students. The results showed that the correlation filter is an effective technique to reduce the data dimensionality and select the most important attributes for predicting student success and failure in school. Additionally, the DT, RF, and SVM algorithms presented the highest accuracy

(94.483%) when the correlation filter was set at 0.1, indicating that selected parameters, such as Fathers Education, Discussion improving interest, and Course ID, may have a significant impact on students' success or failure in higher education.

After data collection, it was evident that the 88.89% accuracy indicates good precision in classifying students based on the analyzed characteristics. This means that the model used to collect and process the data was able to correctly classify the vast majority of students based on their individual characteristics. This high accuracy rate shows that the model is reliable and can provide valuable insights into different student groups and their specific characteristics. Based on these results, those responsible for the analysis can make more informed and targeted decisions regarding the development and support of students, aiming to improve their academic outcomes and overall educational experience.

These results point to the need for greater attention to education and student development. It is essential to implement educational strategies and programs that stimulate student engagement, promote reading, encourage participation in extracurricular activities, and emphasize the importance of regular study. Moreover, effective support from parents, teachers, and society as a whole is essential for students to reach their full potential and become qualified citizens prepared for the challenges of the modern world.

Based on the data analysis conducted in this study, we rein- force the importance of continuous investments in education, both in terms of financial resources and appropriate educational policies. Through quality education, it is possible to promote equal opportunities, reduce social inequalities, and prepare students for a prosperous and productive future. It is fundamental that all stakeholders involved in the educational process, including the State, educational institutions, teachers, and society in general, work together to improve education and ensure a better future for future generations.

This study also highlighted the importance of using artificial intelligence and data mining techniques in the field of education, enabling teachers and educational institutions to make personalized decisions to improve students' academic performance. Future works may further explore the use of ML techniques in other areas of education, such as primary and secondary education, and in different cultural and geographical contexts. Another future work is the investigation of other ML algorithms, such as neural networks and deep learning, to further explore the relationship between the selected attributes and students' academic performance. With broader analyses, it will be possible to better understand the influence of various factors on higher education and thus sup- port more effective decisions for Brazilian educational policies in higher education.

Funding

The research project received funding from various sources, including PROPI, CNPq, CAPES, and FAPEMIG. These organizations played a vital role in supporting the research conducted by author DAL, providing essential financial resources to carry out the project successfully.

Authors' Contributions

RSD took charge of executing and simulating the results within the database. Additionally, RSD was responsible for collecting and analyzing the data, conducting simulations and tests, and generating the results showcased in this paper. On the other hand, DAL played a pivotal role in writing the paper, translating it, and providing guidance and mentorship to RSD throughout the research process.

Supportive Materials

We are not dealing with humans in our research, as the databases used in our research are open, public, scientific database available on the UCI Machine Learning Repository website: https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation. In this article, we used an open public database deposited in the repository.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Akazaki, J. M., Moro, F. F., Tarouco, L. M. R., Behar, P. A., Machado, L. R., & Cazella, S. C. (2020). Educational data mining for the profile of candidates from the Encceja in the state of Rio Grande do Sul–Brazil. *Redin - Práticas Educacionais e Inovação em Tempo de Isolamento Social, 9*(1).

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

Cortez, P., & Silva, A. M. G. (2008). *Using data mining to predict secondary school student performance*. Portugal: EUROSIS-ETI

de Castro Soares, R., Neto, N. W., Coutinho, L. R., dos Santos, D. V., da Silva, F. J., & Teles, A. S. (2023). Minerando dados para entender os fatores de influência da qualidade educacional do maranhão. *Revista Brasileira de Informática na Educação, 31*(1), 378-406.

de Oliveira, K. L., Boruchovitch, E., & dos Santos, A. A. A. (2008). Leitura e desempenho escolar em português e matemáticano ensino fundamental. *Paidéia (Ribeirão Preto), 18*(41), 531–540.

Dornelas, R. S., & Lima, D. A. (2023). Correlation filters in machine learning algorithms to select demographic and individual features for autism spectrum disorder diagnosis. *Journal of Data Science and Intelligent Systems, 1*(2), 105-127.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of business research, 94*, 335-343.

Ferreira, M. E., Lima, D. A., & Silva, A. (2019). Data analysis for robotics and programming project evaluation involving female students participation. In *2019 Latin American Robotics Symposium*, 417-422.

Guerra, G., Benedetti, E., Resce, G., Potente, R., Cutilli, A., & Molinaro, S. (2020). Status socioeconômico, escolar- idade dos pais, vínculo escolar e recursos socioculturais individuais em vulnerabilidade ao uso de drogas entre es- tudantes. Jornal internacional de pesquisa ambiental e saúde pública, *17*(4), 1306.

Lima, M. N., & Fagundes, R. A. D. A. (2020). Educational datamining: A study of the factors that cause school dropout in higher education institutions in Brazil. *Revista Novas Tecnologias Na Educação, 1*8(1).

Matos, J. D. V., Cruz, J. R., Ribeiro, A. F. S., Gomes, R. M. M., Ferreira, J. C., & Matos, F. B. (2019). Aprendizagem significativa por meio do Uso de TICs: Levantamento das produções da area de ensino de 2016 a 2018. *RENOTE, 17*(1), 466-475.

Moonsamy, D., Naicker, N., Adeliyi, T. T., & Ogunsakin, R. E. (2021). A meta-analysis of educational data mining for predicting students performance in programming. *International Journal of Advanced Computer Science and Applications, 12*(2), 97-104.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery, 10*(3), e1355.

Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. *Applied Sciences, 11*(22), 10907.

Sendacz, N., Isotani, S., & Lima, D. A. (2022). Literature review on technologies and games that motivated people to practice physical activity during the pandemic. *RENOTE, 20*(2), 280–289.

Silva Filho, R. L. C., Brito, K., & Adeodato, P. J. L. (2023). A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement. *Expert Systems with Applications, 221*, 119729.

Wang, Y. (2022). Analysis on the particularity of higher education subject development under the background of artificial intelligence. *International Transactions on Electrical Energy Systems, 2022*.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record, 31*(1), 76-77.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information, 14*(1), 54.

Yılmaz, N., & Sekeroglu, B. (2019). Student performance classification using artificial intelligence techniques. In *10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions*, 596 – 603.