Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6, 370-395
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i6.7814

Enhancing academic performance prediction through machine learning in cloud environments

Mirko Talajić[1], Robert Matijević[1], Zlatan Morić[1*]
[1]Algebra University, 10000 Zagreb, Croatia; mirko.talajic@algebra.hr (MT) robert.matijevic@algebra.hr (RM) zlatan.moric@algebra.hr (ZM)

Abstract: This study aims to enhance the accuracy and applicability of academic performance prediction by integrating machine learning techniques within cloud-based environments. It seeks to address critical gaps in leveraging predictive analytics to support at-risk students and optimize educational outcomes through scalable solutions. The research utilizes a dataset from Portuguese secondary schools, applying advanced machine learning models, including ensemble techniques and cloud-based frameworks such as Azure Machine Learning. Exploratory data analysis, preprocessing techniques like SMOTE for class balancing, and automated machine learning pipelines are employed to develop and evaluate predictive models. The Voting Ensemble model emerged as the most effective, achieving an F1 score of 0.836 and an AUC of 0.973. Historical academic performance, attendance, and parental education were identified as the most influential predictors of student success. The study emphasizes the potential of cloud-integrated machine learning to deliver scalable and interpretable predictive analytics, enabling proactive interventions and promoting equal access to educational opportunities. This study contributes to the field by integrating automated machine learning pipelines with cloud-based solutions, offering a replicable framework for educational institutions. By addressing class imbalance and enhancing model interpretability through feature importance analysis, the research bridges critical gaps in the practical deployment of predictive analytics for academic performance. The findings provide a foundation for future advancements in adaptive, data-driven education systems.

Keywords: *Academic performance prediction, Class imbalance handling, Machine learning in education, Smote technique. Voting ensemble model.*

1. Introduction

The rapid advancement of technology has been recognized as a catalyst for transformative opportunities within educational institutions, allowing for the enhancement of learning methodologies and the provision of personalized educational experiences. The prediction of academic performance has been identified as a critical area among these opportunities, allowing for timely interventions and tailored support for students. The challenges educators face in identifying and assisting at-risk students are addressed by this approach, thereby promoting equitable access to quality education. Accurate predictions of academic success are considered vital for fostering proactive interventions; however, it has been noted that traditional methods for monitoring performance rely on reactive measures, which are seen to limit their effectiveness. In recent years, the potential of machine learning algorithms to revolutionize the analysis and utilization of student data has been demonstrated in educational data mining and learning analytics. Historical academic performance, attendance, and socioeconomic factors have been highlighted as key predictors of success, with the promise of data-driven educational strategies being underscored. However, significant gaps are noted in the practical deployment of these predictive models within real-world educational environments, particularly in their integration into existing systems for scalable and efficient use.

History: Received: 11 March 2025; Revised: 16 May 2025; Accepted: 21 May 2025; Published: 4 June 2025
* Correspondence: zlatan.moric@algebra.hr

This study aims to bridge the gap by exploring the integration of cloud computing, data analytics, and machine learning to predict academic performance in secondary education. A cloud-based predictive model is developed using data collected from Portuguese secondary schools, which is designed to identify students at risk of underperforming and provide actionable insights for educators. Cloud computing offers scaling, real-time data processing, and accessibility, making it a compelling choice for deploying such predictive models.

Two critical questions are grounded in the research. The most significant predictors of student academic success are being sought after. In what ways might the implementation of a cloud-based predictive model be expected to enhance student outcomes? The growing body of evidence is contributed to by the study by addressing these questions, indicating that cloud-integrated machine learning models are considered effective and practical solutions for tackling complex educational challenges.

The importance of leveraging historical performance data, attendance records, and socioeconomic indicators for the accurate prediction of academic outcomes is highlighted by the findings of this research. Moreover, the study underscores the need for scalable, user-friendly technological tools that can be seamlessly integrated into educators' workflows. Both theoretical understanding and practical applications are advanced to foster more adaptive and inclusive educational environments.

2. Literature Review

Traditional machine-learning approaches have been foundational in predicting academic performance. Decision trees, k-nearest neighbors (KNN), and support vector machines (SVM) are widely recognized for their simplicity and effectiveness. Decision trees, for instance, excel in handling categorical data and providing visual interpretability. They have been successfully applied to classify and predict student outcomes based on demographic, academic, and social data, revealing important patterns in educational settings [1]. Random forests, an extension of decision trees, improve prediction accuracy by combining multiple trees into an ensemble model. This method has shown robust performance in identifying the key predictors of academic success, such as prior grades and attendance, while reducing overfitting tendencies [2]. KNN, a non-parametric algorithm, has also been extensively utilized due to its ability to capture similarity among data points. By analyzing the closest neighbors, KNN predicts academic outcomes effectively in cases where data distribution does not follow a clear pattern [3]. Similarly, SVM has demonstrated its strength in dealing with high-dimensional data by constructing hyperplanes that separate academic success categories with maximal margins. This approach efficiently involves complex interrelationships among input features, such as socioeconomic factors and parental education levels [4]. Naïve Bayes and logistic regression are also prominent methods in this domain. Naïve Bayes, relying on probabilistic models, performs well in scenarios with independent feature assumptions, while logistic regression provides insights into the likelihood of academic outcomes based on input features like study habits and classroom engagement [5]. However, these traditional algorithms often require careful preprocessing and feature engineering, which can be time-intensive but essential for achieving optimal performance [6].

Ensemble methods like boosting, bagging, and random subspace classifiers have been extensively explored to improve predictive accuracy and robustness. Boosting combines weak learners iteratively to build a strong classifier, enhancing its accuracy with each iteration. For instance, AdaBoost has been applied in academic prediction contexts, achieving superior accuracy by focusing on misclassified samples during training [7]. Bagging methods, including random forests, build multiple independent models and aggregate their predictions to reduce variance and improve stability. This approach is practical for datasets with noisy data or complex feature interdependencies, as demonstrated in various academic performance prediction tasks [8]. Random subspace classifiers explore diverse feature subsets, improving generalization by reducing overfitting. This method is advantageous when dealing with high-dimensional data, as it prevents the model from overly focusing on dominant features [9]. Optimization techniques play a crucial role in improving ensemble model performance. Inspired by natural selection, genetic algorithms have optimized feature selection, ensuring that only the most relevant features contribute to the prediction model. This approach enhances accuracy and reduces computational complexity [10]. The double particle swarm optimization-based categorical boosting model further exemplifies how optimization techniques can fine-tune ensemble models. This approach uses particle swarm optimization to adjust hyperparameters, achieving a remarkable accuracy of 96.62% in predicting student grades [11]. Fine-tuning SVM parameters has also proven effective, showcasing how careful calibration can significantly improve model precision and reliability [12].

Deep learning approaches have transformed the landscape of academic performance prediction, leveraging their ability to model complex relationships and high-dimensional data. Convolutional Neural Networks (CNNs) have effectively captured spatial hierarchies within input data. For instance, CNNs have been employed to analyze patterns in student engagement metrics, achieving high levels of prediction accuracy for academic outcomes [13]. Recurrent Neural Networks (RNNs), designed to handle sequential data, have been applied to capture temporal dependencies in student learning behaviors. These models help predict long-term academic performance by analyzing time-series data, such as attendance records and progressive assessment scores [14]. Deep Neural Networks (DNNs) that incorporate advanced techniques, such as dropout regularization, have been shown to reduce overfitting in academic datasets. By leveraging dropout, DNN models have achieved up to 89.36% accuracy in predicting student success in mathematics, emphasizing the effectiveness of advanced deep learning techniques in handling structured educational data [15]. Hybrid models combining CNNs and RNNs have further improved predictive capabilities by integrating spatial and temporal patterns. Such models have been deployed to analyze complex educational datasets, demonstrating their robustness and scalability in predicting student performance [16]. Despite these advancements, deep learning methods face challenges, including the need for large datasets and extensive computational resources. Overfitting remains a concern, mainly when dataset size is limited. Recent efforts have focused on augmenting data through synthetic generation and transfer learning, addressing these limitations effectively [17]. Studies have also highlighted the adaptability of linear regression combined with feature optimization techniques, showing its relevance as a complement to deep learning in less complex datasets [18].

Cloud computing has revolutionized educational data analysis, providing scalable and efficient solutions for processing large datasets. Streamlit applications have been employed to deploy machine learning models for real-time academic performance prediction, achieving notable improvements such as 73.20% test accuracy using Random Forest models [17]. Hadoop Distributed File Systems (HDFS) and MapReduce frameworks have demonstrated significant advantages in handling distributed educational datasets. These technologies enhance computational efficiency and have improved recognition rates, such as 87.32% using Support Vector Machines for academic prediction tasks [19]. Virtual learning environments benefit significantly from cloud-integrated solutions. For instance, deploying gradient boosting algorithms in these settings has resulted in remarkable performance, with a maximum accuracy of 97.5% for predicting student success [20]. Emerging trends in cloud-integrated machine learning include serverless architectures and cloud-native frameworks. These technologies allow educators to leverage computational resources dynamically, eliminating the need for dedicated infrastructure. By automating resource management, they reduce costs while maintaining high processing capabilities [21]. Platforms like Microsoft Azure and Google Cloud provide pre-built tools to develop and deploy machine learning models in educational settings. These platforms enhance scalability and accessibility, enabling institutions to perform real-time predictive analytics supporting data-driven decision-making [22]. Data security and privacy remain critical concerns in cloud-based educational systems. Research emphasizes the need for robust encryption and access control mechanisms to protect sensitive student data while ensuring compliance with regulatory standards such as GDPR [23]. Hybrid cloud systems offer a promising solution for balancing data accessibility with security. These systems allow institutions to store sensitive data in private clouds while using public clouds for computational tasks, optimizing performance and privacy [24].

Additionally, integrating cloud computing with automated machine learning (AutoML) enhances the efficiency of predictive modeling. AutoML frameworks deployed on cloud platforms automate complex tasks like hyperparameter tuning and feature selection, achieving higher accuracy with reduced manual intervention [25]. Edge computing complements cloud-based systems by processing data locally on edge devices. This approach reduces latency, enabling real-time decision-making in classroom environments. The synergy between edge computing and cloud technologies further optimizes resource utilization and enhances the responsiveness of educational systems [26]. As the adoption of cloud technologies in education expands, their potential to transform academic performance prediction grows. By combining cloud-based solutions with advanced machine learning models, educational institutions can deliver personalized and compelling learning experiences. One study demonstrated the power of AutoML by achieving over 90% accuracy in predicting student academic success. The approach involved ensemble models that combined multiple machine learning algorithms to provide robust and reliable predictions. AutoML frameworks, such as Google AutoML and H2O.ai, enable users to focus on problem-solving rather than the intricate technicalities of machine learning pipelines [20].

Moreover, AutoML has been shown to outperform traditional approaches in virtual learning settings. By leveraging large-scale data, these frameworks can optimize model configurations dynamically, ensuring that the best-performing model is selected for specific educational datasets. For example, experiments using AutoML in real-time feedback systems for virtual classrooms revealed improved predictive capabilities compared to manually tuned models [21]. In addition to improving accuracy, AutoML enhances scalability. Cloud-based implementations of AutoML allow educational institutions to process vast datasets with minimal computational overhead. These systems can dynamically allocate resources based on the complexity of the task, ensuring cost-effective operations without sacrificing performance [22]. Despite its advantages, AutoML also faces challenges. The reliance on large datasets can lead to biases if the input data lacks diversity. Furthermore, the "black-box" nature of AutoML-generated models often makes it difficult for educators to interpret the results, hindering trust and adoption. Addressing these limitations through explainable AI techniques and ensuring data diversity are key steps toward maximizing the potential of AutoML in education [23].

Numerous variables significantly influence academic performance, with diverse studies highlighting intrinsic and extrinsic factors. Academic achievement is often shaped by socioeconomic, psychological, and institutional elements that interact in complex ways to impact student outcomes. Institutional factors, including the quality of teaching and access to educational resources, are equally influential. Studies have found that effective teaching methodologies and supportive teacher-student relationships enhance learning experiences and academic success. Additionally, access to modern learning tools, such as digital resources and technology-enhanced learning environments, can significantly improve outcomes [24]. Peer influences and social networks are increasingly recognized as important determinants of academic performance. Collaborative learning environments and supportive peer groups encourage knowledge sharing and provide emotional support, fostering an environment conducive to academic excellence [25]. Recent research has expanded on these findings, highlighting multidimensional determinants of academic performance. A study analyzing data from over 5,000 undergraduate students in Morocco emphasized the role of psychological support, balanced lifestyle practices, and societal engagement in shaping academic success [27]. Similarly, research focusing on student engagement identified strong correlations between academic performance and behavioral engagement, demonstrating the value of fostering active participation in learning [28]. Cultural and community factors also play a significant role in academic performance. For instance, cultural attitudes toward education, community engagement, and extracurricular participation have been linked to improved student outcomes. Promoting inclusive educational practices that respect cultural diversity is critical for creating equitable learning environments [29]. Moreover, studies exploring environmental and lifestyle factors such as sleep quality, physical activity, and internet use have revealed their impact on academic outcomes, underscoring the need for holistic approaches to student well-being and success [30].

The existing literature reveals significant gaps in the comprehensive understanding of factors influencing academic performance and the methodologies employed for predictive analytics. While prior studies often discuss demographic, socioeconomic, and institutional factors, they examine these elements in isolation rather than exploring their complex interplay. This fragmented approach limits the ability to capture the nuanced relationships between diverse factors that shape student outcomes. Furthermore, although machine learning techniques are frequently applied, the focus remains predominantly on traditional methods, with limited integration of advanced technologies such as automated machine learning (AutoML), ensemble methods, and cloud computing. This lack of technological advancement constrains the scalability and robustness of predictive models, leaving untapped potential for addressing real-world educational challenges. A critical shortcoming is the insufficient emphasis on the interpretability of machine learning models and their practical applicability. Existing studies prioritize predictive accuracy without adequately addressing how these models can inform actionable strategies for educators or policymakers. Moreover, the datasets used are often small, localized, and context-specific, hindering findings' generalizability across diverse educational environments. Ethical considerations, particularly concerning the privacy and security of sensitive student data, are also underrepresented, leaving significant gaps in ensuring the responsible use of predictive analytics in education. In response to these limitations, our approach seeks to bridge these gaps by adopting a multidimensional analytical framework that integrates demographic, behavioral, and institutional factors. By leveraging AutoML frameworks and ensemble techniques within a cloud computing infrastructure, our methodology enhances the scalability and robustness of predictive models. This integration improves accuracy and facilitates the dynamic allocation of computational resources, making real-time interventions feasible.

Additionally, our study prioritizes the interpretability of models through explainable AI techniques, ensuring that predictions are actionable and transparent for educational stakeholders. The utilization of large and diverse datasets, enabled by multi-cloud platforms, further enhances the generalizability of our findings. By embedding rigorous data security protocols and aligning with regulatory standards such as GDPR, we address ethical concerns and reinforce the trustworthiness of our approach. This comprehensive and technologically advanced methodology represents a significant step toward addressing the limitations identified in existing research, offering a more inclusive and practical framework for academic performance prediction.

3. Methodological Framework for Academic Performance Prediction

A comprehensive methodological framework has been used in this study to analyze and predict student academic performance. The framework integrates exploratory data analysis, advanced preprocessing techniques, machine learning algorithms, and cloud-based computational resources to address the challenges associated with complex educational datasets, ensuring scalability, efficiency, and accuracy.

An exploratory data analysis (EDA) phase is initiated, during which a detailed examination of the dataset is conducted to uncover patterns, relationships, and trends. Visualization tools such as histograms, scatter plots, and box plots are employed in EDA to provide a clear picture of the data distributions, detect outliers, and assess relationships between academic performance and other features. The integral part of this phase is correlation analysis, in which Pearson's correlation coefficient measures linear relationships among continuous variables, and nonlinear or monotonic relationships are employed using Spearman's rank correlation. Additionally, demographic and school-related factors are examined to identify potential disparities, with insights into which attributes might be considered strong predictors of academic success.

Extensive data preprocessing is conducted to ensure the dataset is suitable for machine learning. Missing values are addressed by applying imputation techniques or removal, determined by their relevance and distribution within the dataset. Numerical representations are obtained from categorical variables by applying dummy encoding, allowing for compatibility with machine learning algorithms.

The issue of class imbalance, commonly encountered in educational datasets, is mitigated by applying the Synthetic Minority Over-sampling Technique (SMOTE), whereby synthetic samples are generated for underrepresented classes. The model training process is ensured to fairly represent all categories, thereby improving the robustness of predictions.

The critical step of feature engineering is recognized within this framework. Redundant or irrelevant features are eliminated to minimize noise and enhance the efficiency of the predictive models. The relative contribution of each variable to the model's predictions is determined through feature importance analysis, which provides interpretable insights into the key factors influencing academic performance. The transparency of the model is enhanced, and the understanding of the underlying relationships between input features and the target variable is aided.

Cloud computing platforms are leveraged in the study to address the computational demands of data preprocessing, analysis, and model development. The Google Cloud Platform is utilized extensively for data cleaning and transformation, with the automation of inconsistency detection and suggestion of appropriate transformations performed by Dataprep by Trifacta. This visual interface facilitates data preparation in a format optimized for analysis. Sophisticated visualization tools, including dynamic charts and graphs, are provided by Oracle Cloud Analytics, enhancing the interpretability of the data and supporting meaningful insights into feature distributions and relationships. The pivotal role of Microsoft Azure Machine Learning in this framework is highlighted, mainly through the automated machine learning (AutoML) capabilities, which are utilized to streamline the model-building process by automatically selecting the best-performing algorithm and optimizing its parameters. Seamless code execution and analysis are enabled by the integration of Jupyter Notebooks within Azure.

Various machine learning algorithms are employed in the predictive modeling phase, tailored to multiclass classification tasks, where multiple categories comprise the target variable. The algorithms tested include logistic regression, decision trees, random forests, and neural networks. Training and validation of these models are conducted using cross-validation techniques to ensure generalizability to unseen data. Evaluation metrics, including accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC), are utilized to assess model performance comprehensively. It is ensured that balanced learning across all classes is achieved using SMOTE during model training, which is critical for reducing prediction bias.

Statistical methods are also included in the analysis to deepen the understanding of relationships within the data. The Mann-Whitney U test is employed to identify statistically significant differences between two independent groups, particularly when the data do not meet the assumptions of normality required by parametric tests. This non-parametric test provides insights into group-level differences based on ordinal or continuous data. The Phi coefficient is utilized to analyze binary variables, with the strength of associations being quantified in a manner analogous to Pearson's correlation yet specific to binary data. The integration of ensemble learning techniques, such as Voting Ensembles, is a key feature of this methodological framework, whereby the predictions of multiple algorithms are combined to enhance overall performance. These ensemble methods capitalize on the diverse strengths of individual models, resulting in improved prediction stability and accuracy. The framework addresses Challenges in imbalanced datasets, which also evaluates multiclass classification models using weighted metrics, ensuring that performance assessments account for class proportions.

The primary programming language for data manipulation, analysis, and machine learning implementation is served by Python. Efficient data handling is supported by libraries such as Pandas and NumPy, while Scikit-learn provides a robust suite of machine learning algorithms and evaluation tools. Visualization libraries such as Matplotlib and Seaborn are utilized to produce insightful graphs and charts that aid in interpreting findings.

This methodological framework presents a holistic approach to academic performance prediction. The limitations of traditional educational data mining practices are addressed by integrating exploratory analysis, rigorous preprocessing, advanced machine learning techniques, and scalable cloud-based tools. The framework ensures accurate and interpretable predictions, while a replicable and adaptable model is established to enhance decision-making and support student success in diverse educational contexts.

4. Exploratory Insights

The dataset utilized in this research is regarded as a critical foundation for understanding and predicting academic performance in secondary education, specifically in Mathematics. In 2008, a unique blend of real-world academic, behavioral, and socio-demographic information was provided by this dataset, which was collected from 395 students enrolled in two Portuguese secondary schools, Gabriel Pereira and Mouzinho da Silveira, located in Porto. Students aged 15 to 18 are catered to by these schools, which are aligned with the high school level in Portugal. The dataset was initially introduced in the seminal research titled Using Data Mining to Predict Secondary School Student Performance [1] and has since been established as a benchmark resource for educational data research.

The significance of this dataset is found in its comprehensiveness, with a broad spectrum of factors that influence student performance being captured. The factors include demographic details, socioeconomic conditions, personal habits, and academic records. The exploration of variables that affect student achievement is facilitated by such diversity, which is considered an ideal basis for constructing predictive models. A detailed examination of the dataset's structure, variable categories, and their relevance to the study objectives is provided in this section.

A total of 33 features are comprised in the dataset, which are categorized into demographic, socioeconomic, school-related, personal, and academic performance variables. These variables capture objective and subjective aspects of a student's educational experience. Table 1 is presented below, containing all variables along with their descriptions, types, and values.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 370-395, 2025
DOI: 10.55214/25768484.v9i6.7814

Table 1.
Variables in the dataset.

Variable	Description	Type and Values
school	Student's school	Binary: 'GP' or 'MS'
sex	Student's gender	Binary: 'F' (female) or 'M' (male)
age	Student's age	Numeric: 15 to 22
address	Home address type	Binary: 'U' (urban) or 'R' (rural)
famsize	Family size	Binary: 'LE3' (≤3) or 'GT3' (>3)
Pstatus	Parent's cohabitation status	Binary: 'T' (together) or 'A' (apart)
Medu	Mother's education	Numeric: 0 to 4 (higher education)
Fedu	Father's education	Numeric: 0 to 4 (higher education)
Mjob	Mother's occupation	Nominal: 'teacher', 'health', 'services', 'at_home', or 'other'
Fjob	Father's occupation	Nominal: 'teacher', 'health', 'services', 'at_home', or 'other'
reason	Reason for choosing school	Nominal: 'home', 'reputation', 'course', or 'other'
guardian	Student's guardian	Nominal: 'mother', 'father', 'other'
traveltime	Travel time from home to school	Numeric: 1 (<15 m) to 4 (>1 h)
studytime	Weekly study time	Numeric: 1 (<2 h) to 4 (>10 h)
failures	Number of past class failures	Numeric: 0 to 4 (≥4 failures)
schoolsup	Extra educational support	Binary: Yes or No
famsup	Family educational support	Binary: Yes or No
paid	Extra paid classes	Binary: Yes or No
activities	Particip. in extracurricular activities	Binary: Yes or No
nursery	Attended nursery school	Binary: Yes or No
higher	Plans for higher education	Binary: Yes or No
internet	Internet access at home	Binary: Yes or No
romantic	Romantic relationship status	Binary: Yes or No
famrel	Quality of family relationships	Numeric: 1 (v. bad) to 5 (excellent)
freetime	Free time after school	Numeric: 1 (v. low) to 5 (v. high)
goout	Frequency of going out with friends	Numeric: 1 (v. low) to 5 (v. high)
Dalc	Workday alcohol consumption	Numeric: 1 (v. low) to 5 (v. high)
Walc	Weekend alcohol consumption	Numeric: 1 (v. low) to 5 (v. high)
health	Current health status	Numeric: 1 (v. bad) to 5 (v. good)
absences	Number of school absences	Numeric: 0 to 93
G1	First-period grade	Numeric: 0 to 20
G2	Second-period grade	Numeric: 0 to 20
G3	Final grade	Numeric: 0 to 20 (output target)

The relationships between academic performance indicators—such as the first-period grade (G1), second-period grade (G2), and final grade (G3)—and other features are understood through correlation analysis, which is regarded as a foundational step. Statistical measures, including Pearson and Spearman correlation coefficients, quantify these relationships' strength and direction. It has been revealed that a high correlation exists between G1 and G2 grades with G3, indicating a strong predictive value for prior performance. It has been observed that students excelling in the first two periods are more likely to achieve higher final grades, indicating that academic consistency plays a critical role in outcomes. Socioeconomic factors, including parental education and employment, exhibit moderate to strong correlations with performance. It has been observed that better academic achievements are associated with higher parental education levels, likely reflected in greater access to resources, support, and an emphasis on the value of education within these families. It has been observed that a positive correlation exists between weekly study time and grades, highlighting the significance of dedicated effort outside the classroom. Conversely, it has been shown that variables such as frequent alcohol consumption and excessive social outings are negatively correlated, indicating the adverse effects of lifestyle choices on academic success.

The performance of various student subgroups is compared to uncover disparities that might require targeted interventions. Gender-based comparisons display nuanced patterns, with female students being found to outperform males in specific contexts, while male students are observed to excel in others, reflecting subject-specific trends and behavioral differences. It has been observed that students from urban areas are often found to perform better than those from rural areas, which may be attributed to variations in access to educational resources and support systems.

School-related factors also influence outcomes, such as the availability of extra educational support and the reasons for choosing a particular school. For instance, it has been observed that additional educational support is received by students who generally perform better, thereby emphasizing the importance of tailored assistance. On the other hand, it has been observed that higher engagement and motivation are often exhibited by students who choose a school based on reputation, resulting in superior academic performance.

The distribution of final grades across the two schools is analyzed, and percentages of grades are presented in Table 2 for better understanding. This table provides detailed insights into the grade distribution for Gabriel Pereira (GP) and Mouzinho da Silveira (MS), along with the overall dataset. The results reveal notable differences in performance between the two schools.

Table 2.
Percentages of grades received across the two schools.

School	Excellent (%)	Very Good (%)	Good (%)	Sufficient (%)	Fail (%)	Total (%)
GP	4.3	12.66	20.25	22.53	28.61	88.35
MS	0.25	1.27	2.28	3.54	4.3	11.65
Overall	4.56	13.92	22.53	26.08	32.91	100.0

The data indicate that an "Excellent" grade was achieved by 4.30% of students from GP, in comparison to only 0.25% from MS. It was found that "Very Good" grades were earned by 12.66% of GP students, whereas only 1.27% of MS students achieved the same level. Higher percentages of students classified as "Good" (20.25%) and "Sufficient" (22.53%) were observed in GP compared to MS, where the corresponding values were 2.28% and 3.54%, respectively. However, significant failure rates were observed at both schools, with 28.61% of GP students and 4.30% of MS students receiving failing grades.

It has been revealed by the overall dataset that "Excellent" grades were achieved by 4.56% of students, while "Very Good" was earned by 13.92%, "Good" was scored by 22.53%, "Sufficient" classification was given to 26.08%, and failure was recorded for 32.91%. The imbalanced nature of academic outcomes is underscored by these percentages, with challenges in achieving passing grades faced by a substantial proportion of students.

The relative contribution of each grade category to the overall grade distribution for each school is focused on in a second analysis. Table 3 provides a summary that compares grade distributions within the schools with insights into their respective performance trends.

Table 3.
Percentage of total grades across the two schools.

School	Excellent (%)	Fail (%)	Good (%)	Sufficient (%)	Very Good (%)	Grand Total (%)
GP	4.87	32.38	22.92	25.5	14.33	100.0
MS	2.17	36.96	19.57	30.43	10.87	100.0

It has been highlighted that a stronger performance profile is exhibited by GP students across all categories, with a total of 88.35% of the dataset being attributed to GP. In comparison, only 11.65% is attributed to MS. While GP students dominate higher grade categories, MS students exhibit lower performance across the board. For example, a significantly higher proportion of students earning "Good" (22.92%) and "Very Good" (14.33%) grades are observed in GP compared to MS, where the corresponding percentages are 19.57% and 10.87%, respectively. Additionally, it has been observed that GP students achieve a higher percentage of "Excellent" grades (4.87%), while a higher proportion of students receiving "Fail" grades, at 36.96%, is noted in MS, compared to 32.38% in GP.

The presence of a significant difference in performance between the two groups of students (Gabriel Pereira and Mouzinho da Silveira schools) is evaluated through the conduct of statistical analysis using the Mann-Whitney U test, also referred to as the Wilcoxon rank-sum test. The appropriateness of this non-parametric test for comparing two independent groups is established when the assumption of normality is not met.

The test statistic, referred to as *U*, is calculated by comparing the ranks of observations in both groups. The smaller value between U_1 and U_2 is utilized as the test statistic. The *p-value* is derived from the *U-statistic*, with the probability of observing the given test statistic (or one more extreme) under the null hypothesis being provided.

The hypotheses for the Mann-Whitney U test are established as follows:

- The null hypothesis (H0) states that no difference is present between the distributions of the two groups. The probability of an observation from one group being less than or equal to an observation from the other group is considered the same as the probability of it being greater. $P(X<Y)$ is equal to $P(X>Y)$.
- The alternative hypothesis (H1) is stated as a difference being present between the distributions of the two groups. The probability of an observation from one group being less than or equal to an observation from the other group is differentiated from the probability of it being greater. $P(X<Y)$ is not equal to $P(X>Y)$.

The results of the Mann-Whitney U Test were obtained.

The test yielded a *U-statistic* of 8835.0 and a *p-value* of 0.25. The null hypothesis is not rejected since the *p-value* exceeds the significance threshold of 0.05. It is implied that insufficient evidence exists to conclude that a statistically significant difference exists between the two schools' performance distributions at the 0.05 significance level.

It is suggested that the "school" variable does not contribute meaningful differentiation between student performance outcomes. Therefore, its relevance as a predictive feature is considered limited and excluded from the set of predictors in subsequent modeling efforts.

The relationship between study time and academic performance is explored in a second analysis and is presented in Table 4. Overall percentages across all study time groups are examined in this analysis rather than focusing on school-based groupings.

Table 4.
The overall percentage of the final grade received by study time.

Study Time Group	Excellent (%)	Very Good (%)	Good (%)	Sufficient (%)	Fail (%)	Total (%)
1 (<2 h)	1.77	2.78	6.58	6.08	9.37	26.58
2 (2-5 h)	1.01	6.84	10.13	14.43	17.72	50.13
3 (5-10 h)	1.01	3.29	4.05	4.05	4.05	16.46
4 (>10 h)	0.76	1.01	1.77	1.52	1.77	6.84
Total	4.56	13.92	22.53	26.08	32.91	100.0

It is observed that the majority of students in the dataset are categorized into Study Group 2, which is defined as studying 2-5 hours per week, accounting for 50.13% of the dataset. Notably, 17.72% of students in this group fail in Math, highlighting the challenges faced even among moderately studious individuals. A high failure rate of 9.37% is exhibited by Group 1, characterized by studying less than 2 hours per week, despite this group accounting for only 26.58% of the dataset.

In terms of academic excellence, it is noted that only 0.74% of students in the entire dataset are found to achieve an "Excellent" grade within Study Group 4, which is comprised of students studying more than 10 hours per week. The relatively small proportion of high-performing students across all study time groups is underscored.

It is revealed by the overall dataset that most students fail in math (32.91%), while "Excellent" grades are achieved by only 4.56%, further emphasizing the imbalanced distribution of performance outcomes.

An analysis of study time and academic performance focused on the average grades received on the first and second Math tests (G1 and G2). The spread of G1 grades across different study time groups is illustrated by a box-plot visualization (Figure 1). This visualization highlights the variability in performance within each study group, and insights into the correlation between consistent study habits and academic outcomes are provided.

A narrower spread of grades tended to be observed among students in Study Group 4 (studying more than 10 hours per week), indicating greater consistency and potentially better preparation. However, it has been determined that their median grades are not significantly higher than those of students in Study Group 2 (studying 2-5 hours per week). It is suggested that while stability is contributed to by increased study time, a significant grade improvement is not necessarily led beyond a certain threshold. Conversely, a wider spread of grades is exhibited by students in Study Group 1 (studying less than 2 hours per week), reflecting more variability and a higher likelihood of lower performance.

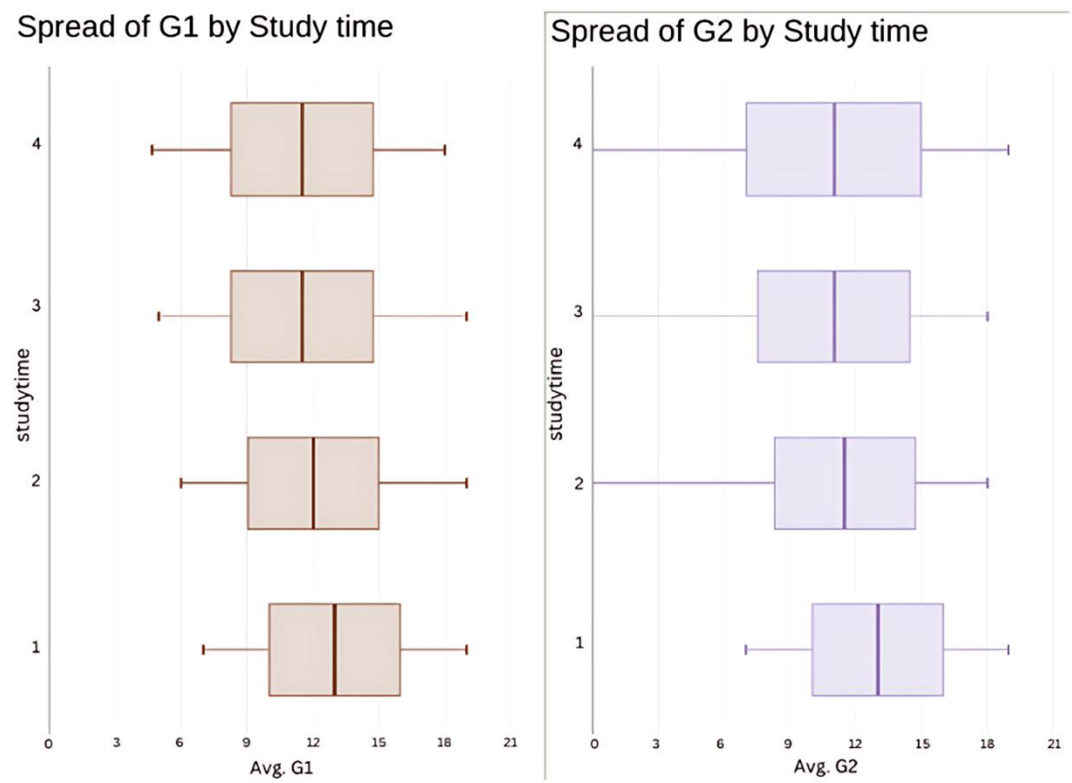

Figure 1.
Box-Plots of the spread of G1 by study time.

Additional analysis is conducted to examine the "romantic" variable, which is indicated by whether a student is in a relationship (encoded as 1) or not (encoded as 0), and its relationship to final grades is assessed. It has been revealed that students not in a romantic relationship are significantly outnumbered by those with higher academic performance, demonstrated by the former group across all grade categories. Among students not in a romantic relationship, "Excellent" grades were achieved by 17, "Very Good" was earned by 38, "Good" was scored by 55, "Sufficient" was classified for 75, and Math was failed by 78. In contrast, it was found that among students in a relationship, only one student was awarded an "Excellent" grade, 17 students were given "Very Good," 34 students scored "Good," 28

ISSN: 2576-8484
Vol. 9, No. 6: 370-395, 2025
DOI: 10.55214/25768484.v9i6.7814

students were classified as "Sufficient," and 52 students failed Math. These results suggest a correlation between relationship status and academic performance, with a higher likelihood of excelling observed among students not in a relationship. The distribution of final grades for both groups is visualized in Figure 2, where bar plots illustrate the grade categories for students in a relationship (left) and those not in a relationship (right). The plots emphasize the disparity in performance, particularly in the higher-grade categories.

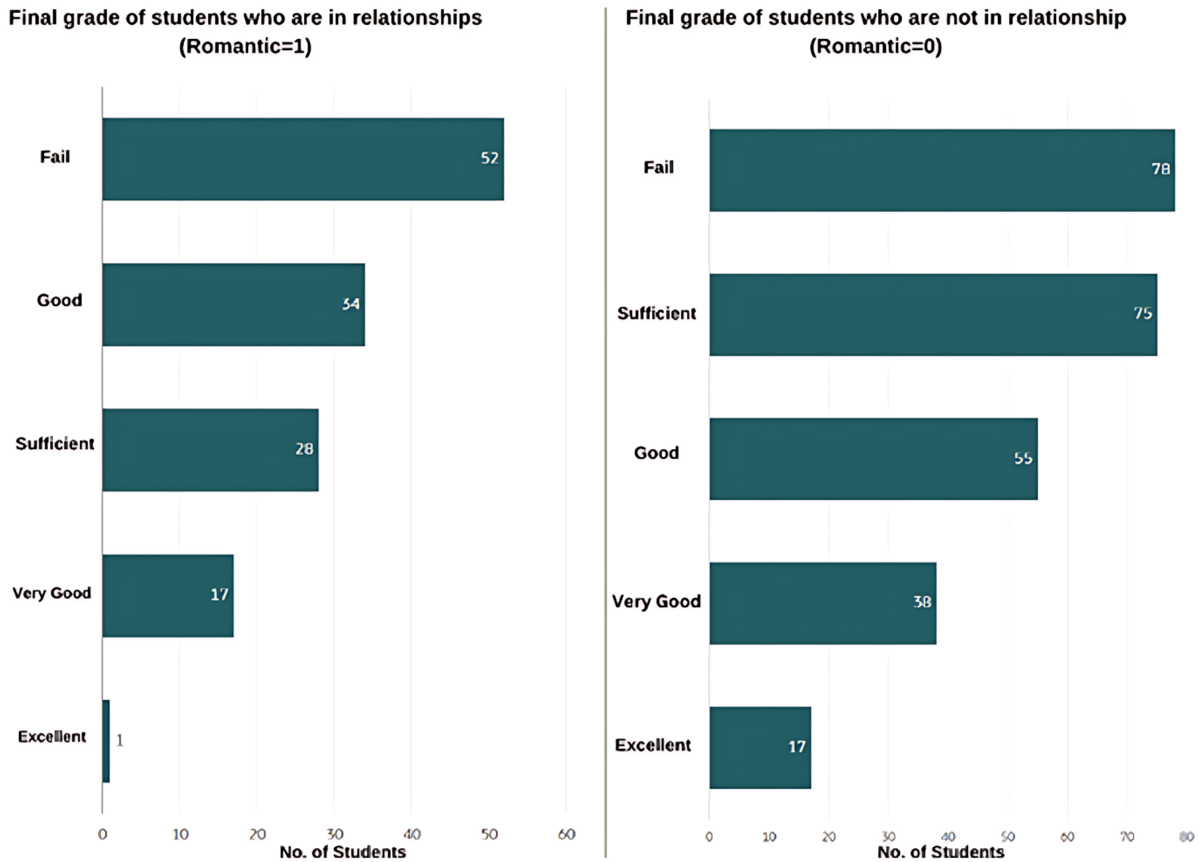

Figure 2.
Bar plots of students in a relationship and their final grade (left) against students who are not in a relationship and their final grade (right).

The strength and direction of linear relationships between continuous variables were first measured using the Pearson correlation coefficient. A visual representation of the correlations is provided by a heatmap (Figure 3), which facilitates the identification of key relationships.

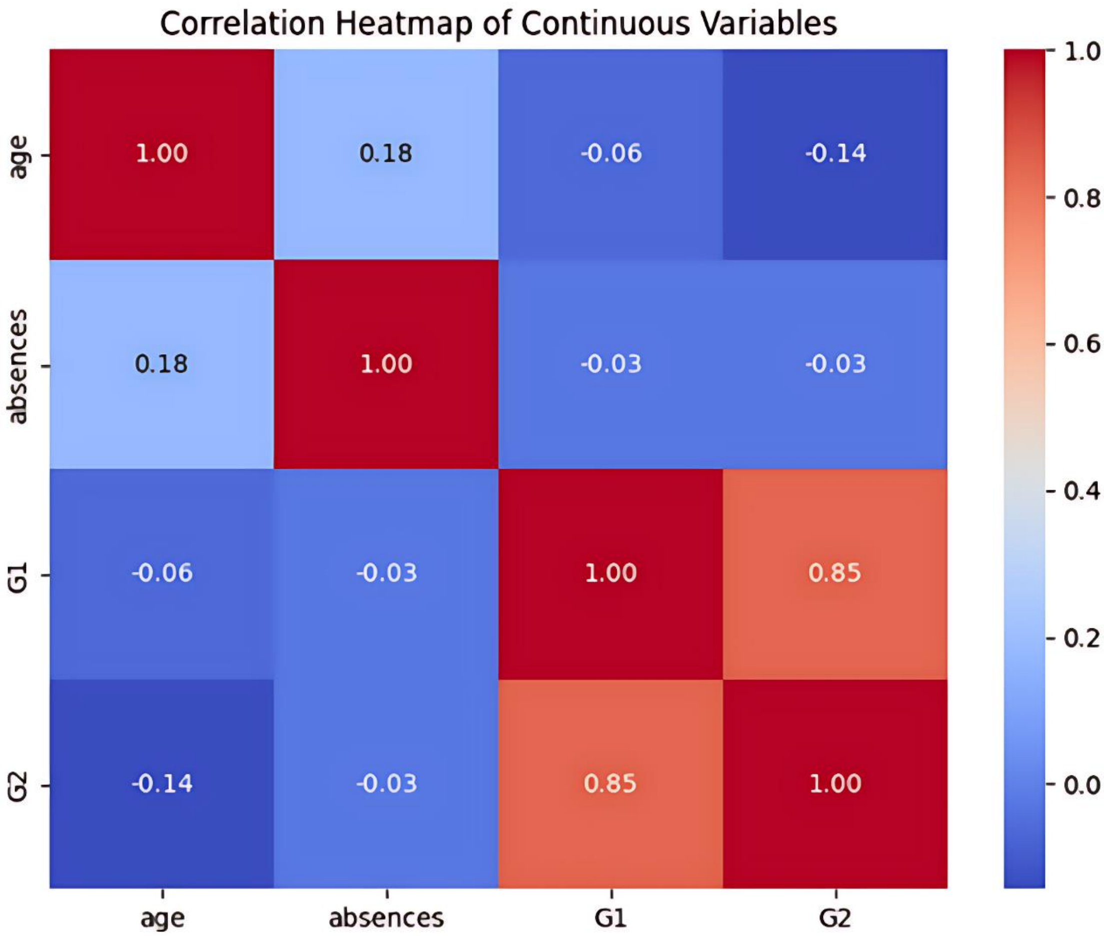

Figure 3.
Correlation heatmap of continuous variables.

The heatmap indicates a strong positive correlation of 0.85 between the variables G1 (first Math grade) and G2 (second Math grade). This finding has been observed to align with prior observations, indicating that strong performance in the first math test is predicted to influence subsequent performance strongly in the second test. The lowest correlation was observed between the variables "absences" and G1/G2 (-0.03), indicating that almost no relationship exists between the number of absences and performance on Math tests. These results underscore the importance of prior performance as a predictor, while it is suggested that absences have a negligible direct impact on grades in this dataset.

Given that the Pearson correlation coefficient is deemed unsuitable for ordinal variables, the Spearman correlation coefficient was utilized to analyze the monotonic relationships between these variables. The non-parametric measure, Spearman's rank correlation, captures linear and nonlinear relationships.

DOI: 10.55214/25768484.v9i6.7814

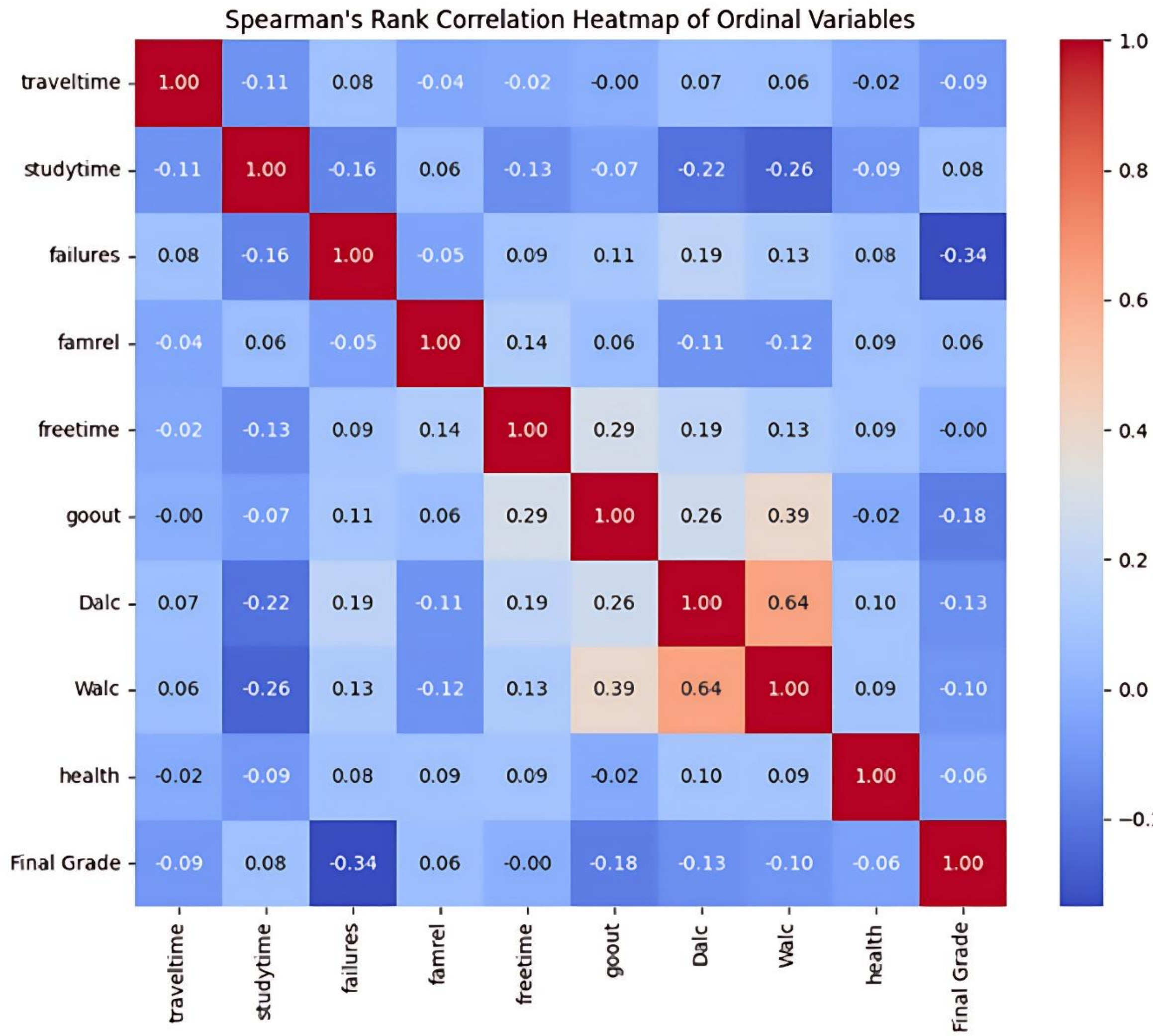

Figure 4.
Correlation heatmap of ordinal variables.

A heatmap (Figure 4) visualizes the results, highlighting significant relationships among ordinal variables. A strong positive correlation of 0.64 was identified between "daily alcohol consumption" (Dalc) and "weekly alcohol consumption" (Walc), indicating that students with higher daily alcohol intake are also likely to consume more alcohol weekly. A moderate positive correlation (0.39) was observed between "weekly alcohol consumption" (Walc) and "going out" (goout), indicating that it is suggested that students who go out frequently tend to drink more alcohol weekly.

The lowest correlation was identified between "travel time" and "going out" (0.00), indicating that no relationship exists between the time taken by students to travel to school and their social activities. Additionally, it has been determined that "free time" is uncorrelated with the "final grade" (0.00), indicating that leisure time does not exert a direct impact on academic performance in this dataset.

The highest correlation with the target variable "final grade" is indicated to be with "failures" (-0.34), which is interpreted as a weak negative correlation. It is suggested that a higher number of past class failures is associated with lower final grades among students. Other weak negative correlations with the final grade were identified, including "going out" (-0.18) and "daily alcohol consumption" (-

ISSN: 2576-8484
Vol. 9, No. 6: 370-395, 2025
DOI: 10.55214/25768484.v9i6.7814

0.13), indicating the detrimental impact of excessive social activities and alcohol consumption on academic outcomes.

A correlation analysis is conducted on binary variables using the Phi coefficient to analyze the dataset further. The Phi coefficient is specifically designed to measure the strength and direction of association between two binary variables, in contrast to Pearson and Spearman correlations, which are utilized for continuous and ordinal variables. The closeness of the relationship between two binary variables, such as "yes/no" (0/1) responses, is quantified.

A heatmap (Figure 5) is utilized to visualize the results of the Phi coefficient analysis, with the relationships between binary variables being highlighted.

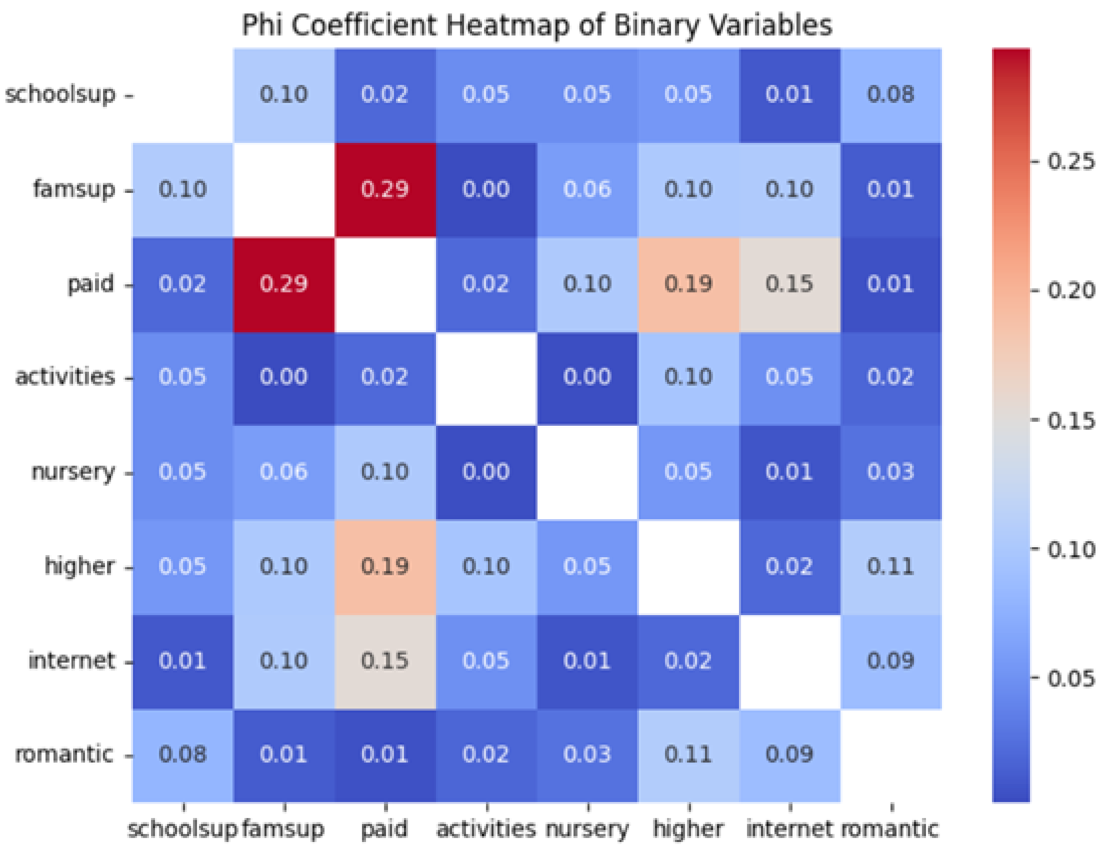

Figure 5.
Correlation heatmap of binary variables.

Noteworthy patterns are revealed by analyzing binary variables using the Phi coefficient. A strong correlation is observed between the variables "extra paid classes" (paid) and "family support" (famsup), characterized by a weak positive correlation of 0.29. It is suggested that students provided with family support are slightly more likely to attend extra-paid classes. A weak positive correlation (0.19) is observed between students aspiring to pursue higher education and those attending extra-paid classes, indicating a modest association between academic ambitions and seeking additional educational support.

It is noted that specific pairs of variables exhibit no correlation. For instance, no relationship (0.00) was observed between participation in "extracurricular activities" (activities) and having attended "nursery school" (nursery), implying that these aspects of a student's experience are unrelated. Furthermore, the heatmap revealed an absence of negative correlations among the binary variables. It

ISSN: 2576-8484
Vol. 9, No. 6: 370-395, 2025
DOI: 10.55214/25768484.v9i6.7814

has been observed that all relationships are either weakly positive or non-existent, indicating that no inverse relationships exist among the binary variables in this dataset.

To ensure that more reliable and accurate research results are obtained, it is essential that the class imbalance in the target variable "Final Grade" is addressed. Class imbalance is characterized by the underrepresentation of certain categories of the target variable, particularly minority classes, compared to others. It has been observed that biased machine-learning models can be produced due to this imbalance, leading to difficulties in accurately predicting minority classes. The Synthetic Minority Over-sampling Technique (SMOTE) handles this challenge. The class distribution is balanced by this widely used technique, leading to improvements in the fairness and effectiveness of machine learning models.

SMOTE generates synthetic samples for the minority class. The algorithm first identified instances of the minority class in the dataset. For each of these instances, one or more of their nearest neighbors in the feature space is selected. New synthetic samples are generated along the line segments joined by the original instance and its neighbors. Additional instances of the minority class are generated by this process without data duplication, resulting in an effective balance of the dataset.

The "Final Grade" variable was addressed for the observed class imbalance by applying SMOTE in this research. The distribution of the target variable is visualized in Figure 6 before and after the application of SMOTE. The original frequency of each class is shown in the left graph, with significant underrepresentation observed in the minority classes. The right graph illustrates the balanced class distribution achieved through SMOTE. It is ensured that machine learning models are trained on a more equitable dataset, which enables better performance across all classes.

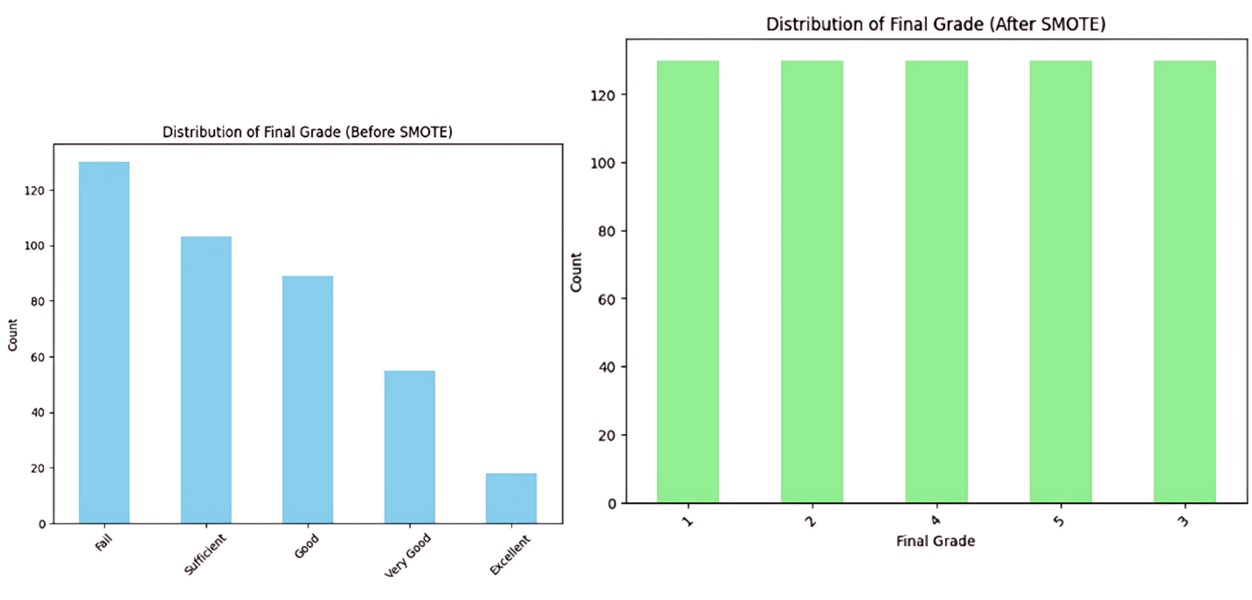

Figure 6.
Distribution of values in the final grade before (left) and after (right) the SMOTE technique.

Following addressing the class imbalance, the dataset is transformed into an appropriate format for machine learning. The purpose is served by Microsoft Azure Machine Learning Studio, which is an integrated development environment (IDE) based in the cloud. Azure Machine Learning Studio offers a user-friendly interface with drag-and-drop capabilities, enabling data preprocessing, model building, training, and deployment to be conducted without the necessity of extensive coding knowledge.

The platform automates many aspects of data preprocessing, including feature engineering, imputation of missing values, scaling, and encoding of categorical variables. Additionally, Automated

Machine Learning (AutoML) is employed by Azure ML Studio to streamline the model development process. AutoML performs an extensive and iterative exploration of machine learning algorithms, with training and optimization conducted to produce high-quality models tailored to the dataset and research objectives.

During the model development phase, various machine learning algorithms suitable for classification tasks are evaluated by Azure ML Studio. The best-performing models are identified through iterative hyperparameter tuning, which is employed to search for optimal combinations to maximize model performance based on predefined evaluation metrics. Approximately 45 minutes were taken for this process, specific to this task, due to the extensive search space explored during hyperparameter optimization.

Ensemble learning techniques, such as bagging, boosting, and stacking, are also leveraged by Azure ML Studio to improve model performance. Predictions from multiple models are combined by ensemble methods, with the diversity of these models being utilized to achieve improved accuracy and generalizability. After training a set of models, evaluations are conducted using the test set, and rankings are assigned based on performance metrics, including accuracy, precision, recall, and F1-score.

5. Modeling Outcomes, Feature Insights, and Evaluation

In this section, the results obtained will be analyzed, and potential improvements for future research will be suggested.

5.1. Model Performance and Evaluation Metrics

The importance of robust performance metrics in assessing the efficacy of machine learning algorithms is highlighted by the results obtained from the modeling process. The effectiveness of various classification models was determined by comprehensively examining metrics such as accuracy, precision, recall, and F1 score. The most effective algorithms for predicting the "Final Grade" target variable were identified by comparing these metrics.

The Voting Ensemble's best-performing model achieved an F1 score of 0.836 and an AUC (Area Under the Curve) of 0.973, indicative of excellent model performance in distinguishing between positive and negative classes. The Precision-Recall curve balances precision (the accuracy of optimistic predictions) and recall (the proportion of true positive cases identified). The confusion matrix further confirms the model's reliability, with 109 out of 130 test instances being correctly classified, resulting in an accuracy of 83.85%.

5.2. Feature Importance and Interpretation

The importance of features is understood to be critical for interpreting the model's predictions and refining future modeling approaches. The multiclass classification models were utilized to address the multi-category nature of the "Final Grade" target variable. An independent feature importance vector was generated by each class, allowing for a detailed understanding of which features were most significantly influenced by each grade category.

The aggregate importance of features was determined by averaging the absolute values of their importance across all classes, as illustrated in Figure 7. The grade achieved in the second Math test (G2) was identified as the most influential feature, with an importance value of 2.15445. The grade from the first Math test (G1) was followed by an importance value of 0.73537, and the number of school absences was noted as 0.35636. Other significant features were identified, including family size (famsize_GT3: 0.18699), student age (age: 0.1852), and health status (health: 0.16891). Behavioral and environmental factors were identified as influencing elements, including weekend alcohol consumption (Walc: 0.16239), mother's education (Medu: 0.16233), frequency of going out with friends (goout: 0.15822), and father's job category (Fjob_other: 0.14869).

ISSN: 2576-8484
Vol. 9, No. 6: 370-395, 2025
DOI: 10.55214/25768484.v9i6.7814

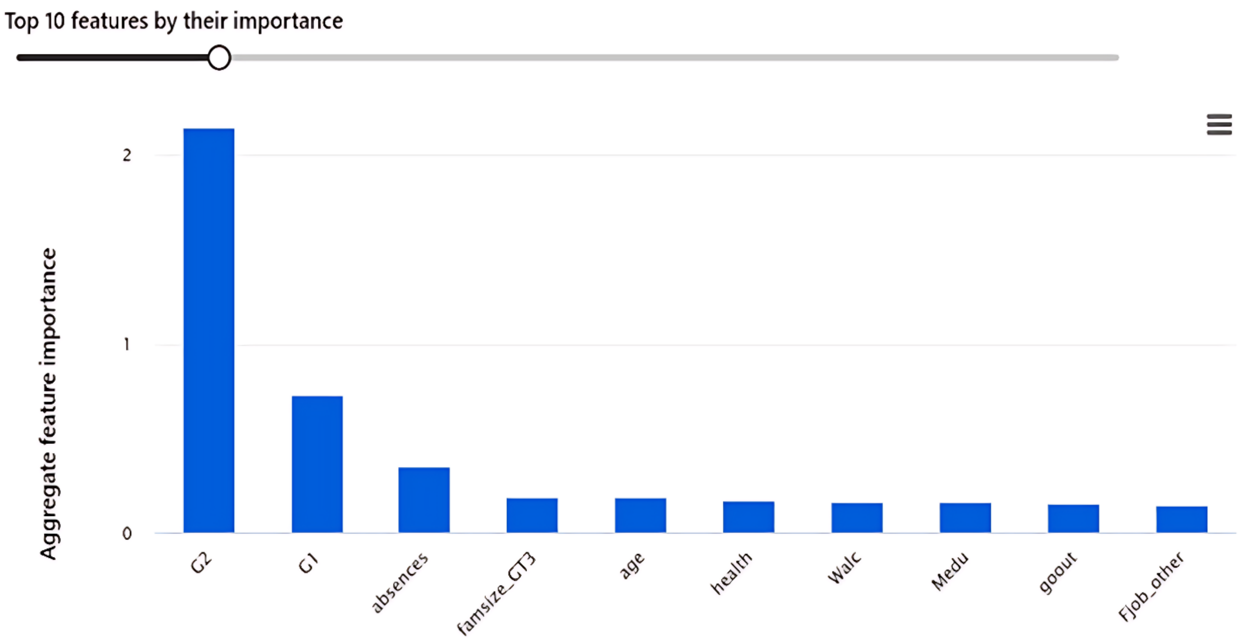

Figure 7.
Top 10 features ranked by their importance on average among all classes.

Further analysis was conducted on the importance of features for specific classes, such as grade 1, and additional insights were revealed. For instance, it is shown in Figure 8 that the results of the first Math test (G1: 0.97) and the number of school absences (0.6) were identified as the most critical predictors for students in this grade category.

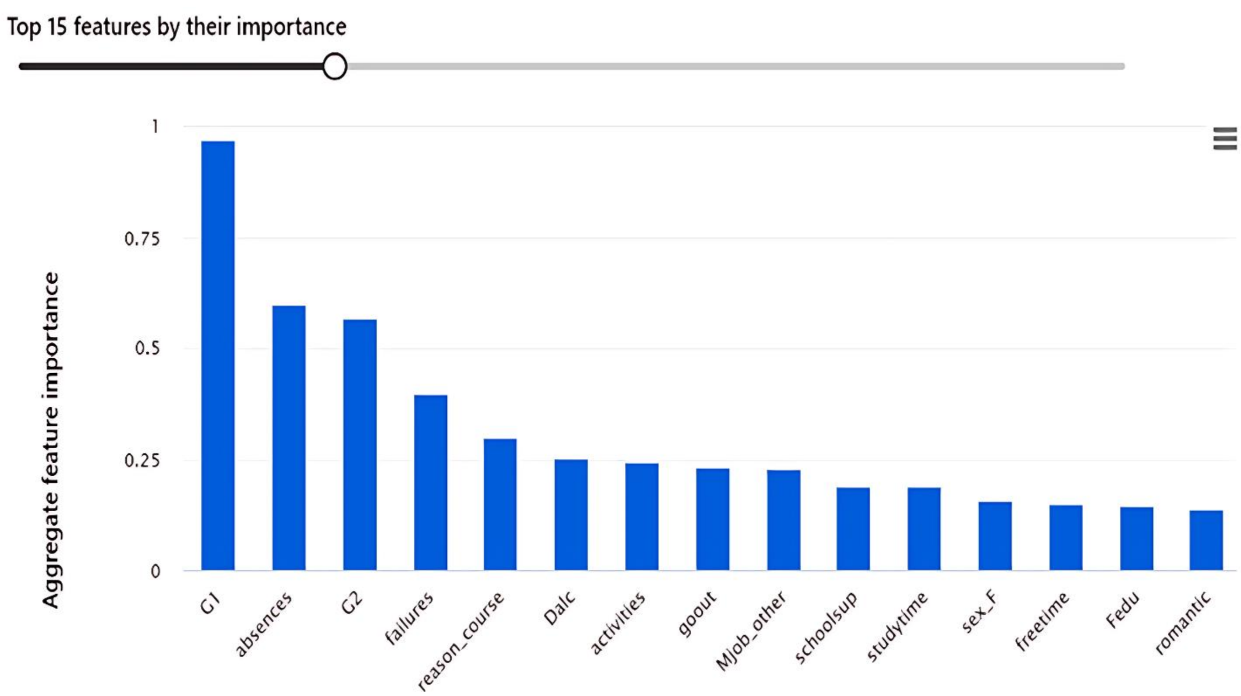

Figure 8.
Top 15 features ranked by their importance on the final grade 1.

ISSN: 2576-8484
Vol. 9, No. 6: 370-395, 2025
DOI: 10.55214/25768484.v9i6.7814

The grade from the first Math test (G1) is the most significant predictor, followed by the number of school absences and the grade from the second Math test (G2). These three features collectively strongly impact the model's predictive capabilities. Factors considered moderately important include prior class failures, the reason for course selection, daily alcohol consumption (Dalc), and participation in extracurricular activities, which are reflected in the influence of behavioral and environmental variables. It was found that features such as study time, gender (sex_F), and parental education (Fedu) have a comparatively weaker influence, indicating a lesser direct relationship with academic outcomes. Insights into the factors most strongly associated with student performance are provided by this analysis, guiding future feature selection and modeling efforts.

5.3. Voting Ensemble Model

The performance of the Voting Ensemble model is evaluated through a range of metrics, which include the Precision-Recall (PR) curve, the Receiver Operating Characteristic (ROC) curve, and a confusion matrix. These tools comprehensively understand the model's accuracy, generalization ability, and predictive reliability.

5.3.1. Precision-Recall and ROC Curves

The trade-off between precision and recall across different threshold values is illustrated by the Precision-Recall curve (Figure 9). The accuracy of positive predictions is measured by precision, while the ability to identify all positive instances is quantified by recall. The model achieves an F1 score of 0.836, indicating a good balance between precision and recall. It is suggested that the model is accurate in its predictions and effective at capturing a significant proportion of positive instances, with very good overall performance.

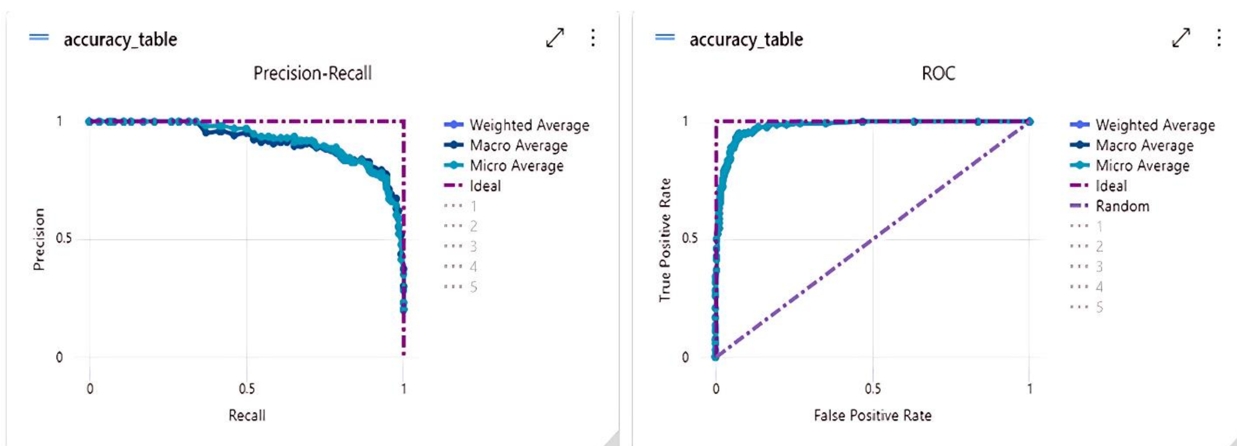

Figure 9.
Voting Ensemble PR curve (left) and ROC curve (right).

The ROC curve provides another perspective on model performance by measuring the ability to distinguish between positive and negative instances. The ability is quantified by the Area under the ROC curve (AUC), with an AUC of 1 being achieved by a perfect classifier. An AUC of 0.973 was achieved by the Voting Ensemble model, reflecting excellent discriminatory power and highlighting its robustness in classification tasks.

5.3.2. Confusion Matrix

A confusion matrix (Figure 10) was applied to the test set containing 130 values to analyze the model's predictions further. The confusion matrix reveals that 109 out of 130 test instances were correctly predicted, resulting in an overall accuracy of 83.85%.

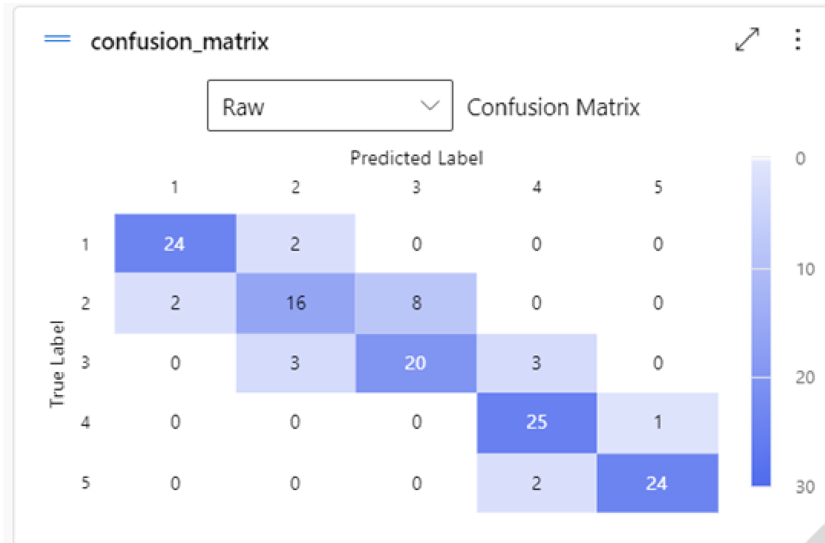

Figure 10.
Confusion matrix of classes prediction of the Voting Ensemble model.

This high accuracy supports the model's effectiveness in correctly classifying data across multiple classes.

5.3.3. Class-Level Precision, Recall, and F1 Scores

The model's performance was assessed for each class of the target variable, as summarized in Table 5.

Table 5.
Class-Level Precision, Recall, and F1 Scores.

	Precision	Recall	F1 Score
Class 1	0.923	0.923	0.923
Class 2	0.615	0.762	0.68
Class 3	0.77	0.714	0.741
Class 4	0.961	0.833	0.892
Class 5	0.923	0.96	0.941
Average	0.838	0.838	0.835

The highest F1 scores were achieved by Class 4 and Class 5, indicating that the model's ability to accurately and consistently predict these classes is reflected in the results. Relatively lower precision and recall were demonstrated by Class 2, suggesting that potential areas for improvement in identifying this category exist.

The Voting Ensemble algorithm was identified as the best-performing model based on the weighted AUC score. The predictions of multiple algorithms are combined in this ensemble to achieve superior performance by leveraging the strengths of each component. The algorithms that are included in the Voting Ensemble are (StandardScalerWrapper and XGBoostClassifier, TruncatedSVDWrapper and LogisticRegression, SparseNormalizer and ExtremeRandomTrees, MaxAbsScaler and LogisticRegression, StandardScaler and ExtremeRandomTrees, StandardScalerWrapper and

RandomForest, MaxAbsScaler and LightGBM). Table 6 summarizes the performance metrics (weighted AUC and weighted F1 score) for individual algorithms used in the Voting Ensemble:

Table 6.
Performance of Individual Algorithms in the Ensemble.

Algorithm	AUC Weighted	F1 Score Weighted
Extreme Random Trees	0.968	0.83
Random Forest	0.965	0.82
XGBoostClassifier	0.965	0.81
LightGBM	0.963	0.82

The Extreme Random Trees algorithm achieved the highest AUC weighted score (0.968), with Random Forest and XGBoostClassifier closely following at 0.965 each. The weighted F1 scores are also demonstrated to show consistent performance, with slightly better results achieved by Extreme Random Trees across these metrics.

The predictive accuracy is improved by combining these algorithms in the Voting Ensemble. However, the risk of overfitting is introduced by training multiple models within an ensemble, mainly when smaller datasets are utilized. The model's performance was rigorously tested on a separate test set to address this. The ensemble's ability to generalize was assessed using unseen data in this evaluation. The results were confirmed to indicate that overfitting was effectively managed, as strong performance was demonstrated by the Voting Ensemble across multiple classes while its ability to generalize to new data was maintained.

5.3.4. Automated Preprocessing in Azure Machine Learning

The integration of this model within the Azure Machine Learning environment is considered a key aspect of its development. The model-building process is automated using Azure AutoML, with preprocessing techniques, including scaling, normalization, and dimensionality reduction, selected and applied according to the dataset's characteristics. For instance, StandardScaler, SparseNormalizer, and MaxAbsScaler were applied automatically as part of the pipeline.

While this automated approach optimizes preprocessing for enhanced model performance, manual intervention is limited. Researchers do not possess direct control over the selection of scalers or hyperparameters for individual algorithms. Preprocessing techniques and hyperparameters that are considered most suitable for the dataset are selected by Azure AutoML. The process is simplified, and efficiency is ensured; however, fine-tuning possibilities that could further enhance performance are restricted.

The Voting Ensemble model effectively combines multiple algorithms' strengths to deliver robust and accurate predictions. The efficiency is enhanced, and complexity is reduced by the automated preprocessing and model-building capabilities of Azure Machine Learning, albeit with limited manual control. Despite these constraints, strong generalization was demonstrated by the ensemble, and overfitting was avoided, making it a reliable choice for multiclass classification tasks. The manual tuning of preprocessing steps and hyperparameters could be explored in future work to refine the ensemble's performance further.

6. Discussion

The multifaceted factors influencing academic achievement are investigated in this study, encompassing demographic characteristics, educational background, health and well-being, school environment, and academic performance indicators. The findings are aligned with existing research, such as that conducted by Cortez and Silva [1] in which the role of these factors in shaping students' academic outcomes is emphasized.

Students' learning environments and support systems were shaped by demographic variables such as gender, age, family size, and parental status rather than being directly determined by them in terms of academic success. Larger family sizes may limit the individual attention received at home, while the emotional and academic support available can be influenced by parental status. These findings are consistent with other studies that highlight the indirect influence of family dynamics on academic performance [20].

It has been observed that higher levels of parental education are correlated with better academic outcomes, which is likely attributed to greater academic support and access to resources [13]. Challenges for students from less privileged backgrounds may be exacerbated by socioeconomic disparities, which include limited access to educational materials and technology. These factors highlight the need for targeted interventions to address inequities and enhance academic opportunities.

Academic performance is significantly affected by health and well-being. The ability to focus, retain information, and perform academically is hindered by poor physical health, chronic illnesses, and mental health challenges among students [13]. It was found that lifestyle choices, such as excessive alcohol consumption and frequent socializing, negatively impacted grades. For instance, it was observed that higher grades were more likely to be achieved by students with lower weekend alcohol consumption and minimal social activities, indicating that balanced lifestyle choices are considered critical for academic success.

The school environment, characterized by teacher quality, class size, extracurricular activities, and available resources, plays a vital role in academic success. A positive school culture is associated with enhanced student engagement, motivation, and improved outcomes, as indicated by earlier findings [1, 14]. It has been observed that students from schools characterized by supportive environments exhibit better academic performance.

The grades achieved in the first and second academic periods were identified as the most significant predictors of the final grade, with the highest correlation exhibited by the second-period grade. It has been established that early academic success is a strong foundation for later achievements, consistent with previous findings [15]. Attendance was identified as a crucial factor, with frequent absences negatively correlated with performance [17]. The influential variables were family size, age, parental education, and health status.

A pattern of consistent historical achievements, regular attendance, and strong family support is typically exhibited by students who excel (achieving a final grade of 5). Excessive social activities and romantic relationships are less likely to be engaged, and studying for more than 10 hours per week tends to occur more frequently. Conversely, poor historical achievements, frequent absences, and higher levels of daily alcohol consumption are often exhibited by students with failing grades. Fewer hours are typically spent studying, while greater engagement in extracurricular activities and socializing is observed [17].

The discussion of several limitations of this study is warranted. The dataset comprises 395 students from only two schools in Porto, Portugal, which limits the generalizability of the findings to other regions or educational systems. Additionally, the dataset is from 2008, and the evolving educational landscape, including technological advancements and shifts in teaching methods, may impact the relevance of these findings. The use of machine learning models, such as the Voting Ensemble, was demonstrated to achieve high accuracy; however, the small sample size was found to increase the risk of overfitting, despite efforts being made to mitigate this through SMOTE.

Future research should address these limitations by including larger, more diverse samples from multiple schools and regions. Longitudinal studies tracking students over several years would provide insights into long-term trends and the impact of early interventions. Additional variables related to socioeconomic status, school infrastructure, teaching quality, and psychological factors could be incorporated to provide a more comprehensive understanding of academic success. Advanced machine learning techniques like deep learning are expected to improve predictive accuracy and provide deeper insights into variable interactions. Finally, ethical considerations, such as student data privacy, should be prioritized in future studies.

This study underscores the complex interplay of demographic, behavioral, and environmental factors influencing academic performance. By addressing these factors holistically and integrating advanced modeling techniques, future research is anticipated to enhance the understanding of student success further and inform the development of targeted interventions.

7. Future Work

This study's findings emphasize the potential of machine learning and cloud-based solutions in enhancing academic performance prediction. However, several opportunities are present for the expansion and refinement of the research that has been presented. Future efforts are suggested to focus on addressing the limitations that have been identified and exploring innovative approaches that may be utilized to advance this field further.

The reliance on a single dataset from two schools is a significant limitation of the current research. It is suggested that more extensive and more diverse datasets spanning multiple regions, educational levels, and socioeconomic contexts be incorporated into future research to enhance the generalizability of findings. Broader patterns and trends will be identified through this approach, with regional and cultural differences being accounted for.

Cross-sectional data was utilized in this study, allowing for a snapshot of student performance to be provided at a specific point in time. Incorporating longitudinal data would allow for exploring the temporal dynamics of academic success and examining how student performance evolves. Insights into the long-term impact of interventions and the predictive value of early academic indicators could be provided.

While traditional machine learning and ensemble methods were employed in this study, the application of advanced techniques such as deep learning, transformer-based models, and graph neural networks could be explored in future research. These approaches can capture complex, nonlinear relationships and dependencies within the data, which may lead to higher prediction accuracy and deeper insights into influencing factors.

Their black-box nature can hinder the practical application of many machine-learning models in educational settings. The integration of explainable AI (XAI) techniques for enhancing the transparency and actionability of model predictions should be prioritized in future research. Clear explanations of how specific factors contribute to academic performance can benefit educators and policymakers.

Cloud platforms such as Microsoft Azure enable real-time data processing and analytics. It is suggested that future studies be conducted to leverage these capabilities to develop systems that provide real-time feedback to students and educators. For example, models could predict at-risk students in real-time, and immediate interventions could be suggested, resulting in a proactive approach to academic support.

As machine learning is integrated into educational systems, ethical concerns regarding data privacy and security are recognized as increasingly important. It is suggested that future research be directed towards developing robust data governance frameworks compliant with global standards, such as GDPR. Transparent policies for data usage and informed consent are required to build trust among stakeholders.

Personalized learning pathways can be designed by extending the predictive models developed in this study. The predictions are integrated with adaptive learning platforms, allowing curricula and teaching strategies to be tailored to individual student needs, resulting in improved engagement and outcomes.

Integrating diverse data sources, such as video analysis of classroom interactions, speech patterns during oral assessments, and physiological data (e.g., stress levels), could be explored in future research to gain a more comprehensive understanding of student performance. The fusion of these modalities with traditional academic and demographic data is expected to reveal novel predictors of success.

While predictors of academic performance are highlighted in this study, the effectiveness of specific intervention strategies based on these predictions should be evaluated in future work. The impact of

ISSN: 2576-8484
Vol. 9, No. 6: 370-395, 2025
DOI: 10.55214/25768484.v9i6.7814

targeted support programs, such as mentorship, tutoring, or wellness initiatives, on student outcomes could be assessed through controlled trials.

Collaboration between educators, policymakers, data scientists, and psychologists is crucial for developing holistic solutions. Diverse expertise is brought together, allowing for addressing complex challenges and creating practical, scalable tools for enhancing education systems in future research.

By addressing these areas, future research can build a foundation to improve the prediction of academic performance and its application in real-world educational settings. That will ultimately contribute to more equitable and effective learning environments for students worldwide.

8. Conclusions

Integrating machine learning techniques within cloud-based environments presents a comprehensive exploration of academic performance prediction. A dataset from Portuguese secondary schools has been leveraged to highlight the multifaceted nature of academic success, which is influenced by demographic, socioeconomic, behavioural, and educational factors. The research underlines the transformative potential of advanced machine learning algorithms, with robust classification performance achieved by ensemble methods such as the Voting Ensemble, which recorded an F1 score of 0.836 and an AUC of 0.973.

The methodological advancements are found to constitute the scientific contribution of this work. Exploratory data analysis is systematically combined with rigorous preprocessing techniques, such as SMOTE, for handling class imbalance, and the deployment of machine learning models is facilitated within a scalable and automated cloud infrastructure. This approach enhances predictive accuracy, and the generalizability and reliability of the models are ensured. Furthermore, the interpretability of predictions is emphasized through feature importance analysis, with actionable insights offered for educators and policymakers. Critical factors influencing student outcomes, including historical academic performance, attendance, and parental education, were identified.

The practical application of machine learning in education is advanced by this research, which demonstrates the feasibility of integrating automated machine learning pipelines within cloud platforms such as Microsoft Azure, in addition to methodological contributions. These systems provide real-time, scalable solutions that educational institutions can readily adopt to identify at-risk students and implement targeted interventions proactively.

While the findings are considered promising, the research acknowledges limitations, including the geographic and temporal constraints of the dataset and the challenges posed by a small sample size. It is suggested that future studies be conducted to address these gaps by incorporating diverse datasets, exploring advanced machine-learning techniques, and emphasizing the ethical use of student data.

The potential of data-driven approaches to enhance educational outcomes is underscored by this study, paving the way for more adaptive, inclusive, and effective learning environments. A foundation for further exploration in leveraging machine learning and cloud computing to address complex educational challenges is provided.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Author Contributions:

Conceptualization, M.T. and R.M.; methodology, M.T.; validation, M.T., and Z.M.; formal analysis, R.M.; investigation, R.M.; writing—original draft preparation, R.M.; writing—review and editing, Z.M. and M.T.; supervision, Z.M.

References

[1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *Proceedings of the 2008 IEEE International Conference on Computer-Based Medical Systems (CBMS)*, 2008.

[2] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," *International Journal of Modern Education and Computer Science*, vol. 8, no. 11, pp. 36-42, 2016. https://doi.org/10.5815/ijmecs.2016.11.05

[3] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques," *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415-6426, 2015. https://doi.org/10.12988/ams.2015.53289

[4] Z. Fan, J. Gou, and C. Wang, "Predicting secondary school student performance using a double particle swarm optimization-based categorical boosting model," *Engineering Applications of Artificial Intelligence*, vol. 124, p. 106649, 2023. https://doi.org/10.1016/j.engappai.2022.106494

[5] J. Maurício and G. Marques, "Predicting the performance of mathematics students through data mining techniques for enhanced education systems," presented at the Conference on Online Teaching and Mobile Education, 2020.

[6] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," *Computers & Electrical Engineering*, vol. 89, p. 106903, 2021. https://doi.org/10.1016/j.compeleceng.2020.106903

[7] S. Begum and S. S. Padmannavar, "Prediction of student performance using genetically optimized feature selection with multiclass classification," *International Journal of Engineering Trends and Technology*, vol. 70, no. 4, pp. 223-235, 2022. https://doi.org/10.14445/22315381/IJETT-V70I7P226

[8] Y. Rong, T. He, Z. Li, and G. Liu, "Achievement performance prediction model based on deep neural network of student similarity," in *Proceedings of the 4th International Conference on Modern Education and Information Management*, 2023.

[9] M. Shoaib, N. Sayed, J. Singh, J. Shafi, S. Khan, and F. Ali, "AI student success predictor: Enhancing personalized learning in campus management systems," *Computers in Human Behavior*, vol. 158, p. 108301, 2024. https://doi.org/10.1016/j.chb.2023.108301

[10] O. Opeyemi, I. Yasser, W. Simonne, and L. Albertha, "Cloud-based STEM student academic success prediction web application," in *Proceedings of the Society for Information Technology & Teacher Education International Conference*, 2023.

[11] H. Al-Shehri *et al.*, "Student performance prediction using support vector machine and k-nearest neighbors," in *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2017.

[12] S. Ashtari and A. Eydgahi, "Student perceptions of cloud computing effectiveness in higher education," presented at the 2015 IEEE 18th International Conference on Computational Science and Engineering, 2015.

[13] Y. Li and Y. Chen, *Improving students' achievement prediction in blended learning environments with integrated machine learning methods. In Machine Learning in Education Science*. Springer. https://doi.org/10.1007/978-3-030-73426-8_19, 2024.

[14] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 4067721, 2024. https://doi.org/10.1155/2024/4067721

[15] J. C. Muñoz-Carpio, Z. Jan, and A. Saavedra, "Machine learning for learning personalization to enhance student academic performance," in *Proceedings of the Conference on Online Teaching and Mobile Education*, 2021.

[16] M. M. E. Khoudier *et al.*, "Prediction of student performance using machine learning techniques," presented at the 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2023.

[17] A. Merchant, N. Shenoy, A. Bharali, and M. A. Kumar, "Predicting students' academic performance in virtual learning environment using machine learning," presented at the 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 2022.

[18] E. Ismanto, H. A. Ghani, and N. I. B. M. Saleh, "A comparative study of machine learning algorithms for virtual learning environment performance prediction," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 4, p. 1677~1686, 2023.

[19] N. Arora and N. Singh, "Factors affecting the academic performance of college students," *Journal of Educational Technology*, vol. 14, no. 1, pp. 47-53, 2017.

[20] V. Kassarnig, P. Sapiezynski, and S. Lehmann, "Social patterns in academic performance predictions," *EPJ Data Science*, vol. 5, pp. 101–112, 2026.

[21] M. Shafiq and G. Berhanu, "Socioeconomic influences on STEM success," *Education Review*, vol. 82, pp. 312–328, 2024.

[22] C. Sandoval, D. Tabla, and H. Florez, "Barriers in educational resources and academic outcomes," *Education Innovation*, vol. 11, pp. 142–157, 2023.

[23] I. A. J. Al-Muslimawi and A. Hamid, "Linking economic dimensions to academic success: Insights for educators," *Economic Education Review*, vol. 40, pp. 128–140, 2022.
[24] P. Kukreti, *Google cloud platform all-in-one guide: Get familiar with a portfolio of cloud-based services in GCP*. New Delhi, India: BPB Publications, 2023.
[25] P. Sarkar and G. Ruiz, *Oracle cloud infrastructure for solutions architects*. Birmingham, UK: Packt Publishing, 2021.
[26] A. Merchant, N. Shenoy, A. Bharali, and M. A. Kumar, "Edge computing and its role in enhancing classroom technologies," *Educational Technology*, vol. 19, pp. 72–85, 2025.
[27] Z. Alj and A. Bouayad, "Multidimensional determinants of academic performance: Insights from undergraduate students in Moroccan universities," *Journal of Technology and Science Education*, vol. 14, no. 2, pp. 607-621, 2024.
[28] T. J. Chan and M. Dai, "Factors influencing academic achievement of university students," *Journal of Communication, Language and Culture*, vol. 3, no. 2, pp. 14-26, 2023.
[29] S. Chapagai, "Factors affecting the academic achievement of the students in community campuses," *KMC Journal*, vol. 6, pp. 1–15, 2024.
[30] L. Redondo-Flórez, D. J. Ramos-Campo, and V. J. Clemente-Suárez, "Relationship between physical fitness and academic performance in university students," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 14750, 2022.

DOI: 10.55214/25768484.v9i6.7814