Lab 2: Predicting Academic Success from Study Habits: A Machine Learning Approach to Student Performance Classification

Priyal Rawat

DeVos Graduate School, Northwood University

MGT 665: Solving Bus Problems W/ Machine Learning

Dr. Itauma Itauma

06/15/2025

Abstract

This research has evaluated the effectiveness of machine learning models for predicting student academic success using behavioral, demographic, and academic performance features. Given a dataset of 10,000 students from Kaggle, five different classification models such as Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, and Naive Bayes were built and evaluated. The current findings were expected, as the new models were developed using a variety of features, including exam scores, assignment completion rate, study hours, sleeping habits, and discussion participation. The dataset was pre-processed, and a binary outcome of success was created as the target variable. The individual model and collective evaluation results exhibited outstanding classification performance. All models were perfect with the evaluation metrics, specifically Decision Tree, Random Forest, and Logistic Regression models. Feature importance analysis indicated that exam scores and assignment completion rate were the most significant predictors, and the impact of discussion participation was of significance. The overall findings are consistent with predictive analytics applications in education context while raising more questions related to dataset limitations and the need for more complicated and diverse educational data in future research.

Introduction

Education is one of the most important part of any individual. There is common question which comes to our mind that why do some students thrive while others fall behind, even when they follow the same curriculum and have access to similar resources. The question has been a major concern for educators, majorly when it is linked to broader social and economic outcomes. Student's success not only affects an individual future but also shows the quality of education by an institute and determine their ability to contribute to economic development of the nation in the longer run (Guanin-Fajardo et al., 2024). However, one of the challenges of success is not able to identify the underperformance in the beginning itself, which can be a game changer for educators and institutions and for student's success as well. While traditional measures like scores, attendance offer some insight, they often fail to focus on behavioral study patterns like study habits of a student, that plays an important role in learning outcomes (Ahmed, 2024).

Study habits like number of hours studied, completion of required assignments, engagement in group or class discussion, and frequency of working independently have long been recognized as an important part of academic success. In fact, at times the manner in which students approach with their learning process is a better indicator of potential performance than a one-time examination or assessment. (Domínguez et al., 2024, Orji & Vassileva, 2022). Still many academic support systems do not use behavioral data to predict. Using machine learning models to identify patterns in student behavior and predicting their outcomes, helps in creating proactive and personalize strategies to increase student's success (Ahmed 2024, Yildiz & Borekçi, 2020).

The objective of this study is to develop and evaluate predictive models that classify students into success and failure categories based on their study habits. The study will use machine-learning algorithms including Logistic Regression, k-Nearest Neighbors (kNN), Decision Tree, Random Forest and Naïve Bayes to identify the most effective model for identifying students at risk of failing. This study provides institutions with tools to improve academic assistance, which can ultimately lead to improved student outcomes.

Literature Review

Predicting students learning performance is not easy, and this is something the faculty and educational institutions have been caring about for some time. Grades are not just determined by intelligence of the student but is also influenced by various other factors like study habits, sleep, levels of stress, and participation in learning. It is also seen in other studies are also using machine learning techniques to understand learning behavior of a student and to find patterns and behaviors contributing to success. For example, Ahmed (2024) showed that variables such as study hours, assignment completion, and attendance played an important role in predicting performance of a student and their success in academic setting. Guanin-Fajardo et al. (2024) presented data to model student's habits to predict the students grades more accurately. These studies provide further evidence of the importance of recognizing and understanding learning behaviors matters more than just focusing on exam results or assessment when focusing on student's academic success.

The study used Logistic Regression, k-Nearest Neighbors (k-NN), Decision Tree, Random Forest and Naive Bayes in this research due to their strong background in applications of machine learning for predicting student success in academic setting. Logistic Regression is preferred in research environments due to its interpretability, simplicity and it also performs well with categorical predictors on binary or multi class classification (Orji & Vassileva, 2022). k-NN is a very simple method that has shown high performance on datasets that have distinct classes especially if the predictors like study time or social media time are good behavioral predictors discriminating between the two behaviors (Yildiz & Borekci, 2020).

Decision Trees are intuitively understood and useful to produce rule-based models that can easily interpreted by the teacher or educator while Random Forest is a strong ensemble method that successfully handles high-dimensional and noisy data (Guanin-Fajardo et al., 2024, Ouatik et al., 2022). Previous research has also shown Decision Trees and Random Forests can be used and they can cope well with datasets containing mixed feature types like categorical and numerical (Domínguez et al., 2024). Further, Naive Bayes produces notable baseline accuracy for educational prediction problems and is unique due to the speed, simplicity and usefulness classifying categorical behavioral traits (Ahmed, 2024).

In terms of variable selection, past research provides useful background. For instance, Ouatik et al. (2022) found that features related to attendance rate, exam marks, while using educational technology were important. Similarly, Dominguez et al. (2023) found behavioral aspects of the students, such as group study and sleeping habits, were worth focusing on. Considering all, this study will use all features available as the first step like study hours, attendance, exam marks, levels of stress, and sleeping practices in initial model training. This will allow the models to learn complex relationships between variables and appeases concerns about leaving out influential predictors too soon. It will also understand feature importance later through examining the importance of each factor that provided the scores for student’s final success from the best models.

While existing literature has shown that machine learning can predictably produce results in education, many studies focused on small datasets, limited behavioral features. For instance, Guanin-Fajardo (2024) provided a comprehensive or detailed overview of ML in education, but did not focus on some behavioral inputs such as self-reported stress, or time spend on social media by the student. The approach of this study offers an opportunity to expand on the other area by using a very large dataset with multiple behavioral related to, for example, learning style, discussion participation, technology usage, and time on social media. In addition, by comparing five models, the study offers a detailed understanding of algorithmic benefits and drawbacks as they pertain to different educational data forms.

All in all, this research extends a more detailed understanding of how everyday study habits and lifestyle choices contributes into successful or failed academic outcomes. As this study is build on existing practice and previously effective techniques, it also explores and focuses on behavioral, cognitive, and lifestyle data together as a predictive ecosystem represents a unique and practical contribution to research related to the prediction of student success.

Methodology

Dataset Overview

The dataset used in this study was obtained from Kaggle that documented a variety of factors through the educational and behavioral records of 10,000 students. The dataset was collected to assess connections between study habits, learning preferences, lifestyle and academic performance. Each record contained data on the following variables like hours studying in a week, preferred learning style of the student such as visual, audio, kinesthetic, number of online courses completed, amount of a student's contribution in class discussions, assignment completion percentage, and exam scores. In addition, there was attendance percentage, educational technology authorship, stress level, social media time, and average time sleeping present in the dataset. The target variable was Final Grade, being A, B, C, D, or F. The dataset was rich in context and variety overall, and strong classification models can be built to classify academic performance based on behavioral and contextual indicators.

Hypothesis Formulation

Hypothesis 2 (H2) proposes *that students who participate in more assignments will be classified as an academic success.* This is well-supported by Ahmed (2024), who found that the continual engagement in activities, like completing assignments can improve learning, and ultimately positively affect academic performance. Ouatik et al. (2022) used machine learning models to rank behaviors, revealing that behaviors related to assignments exhibited the strongest relationships with student achievement, suggesting the importance of a single behavior as a contributor to student success. These hypotheses were measured in exploratory data analysis and by evaluating the feature importance.

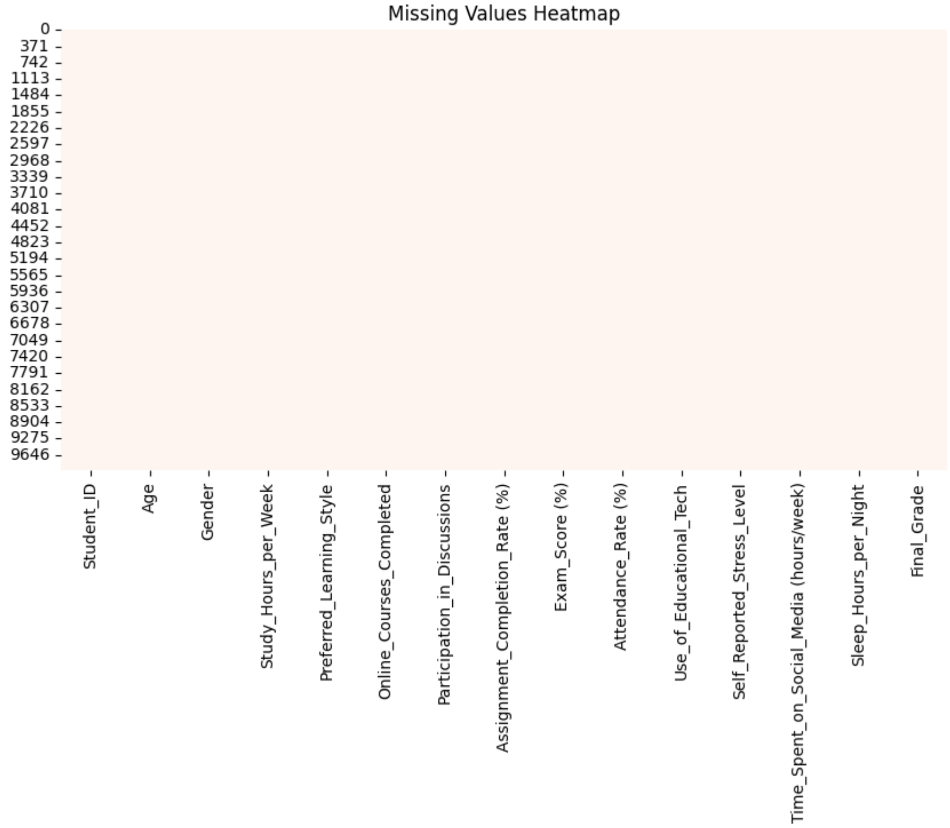
The second hypothesis (H2) proposes that *students who engage in more assignments will be considered an academic success*. Ahmed (2024) supported that continual engagement in activities, such as assignments can result in improved learning, and eventually lead to improved academic performance. Ouatik et al. (2022) utilized machine learning models to rank behaviors and found that behaviors related to coursework from the assignments, like a behavior from an assignment shows the highest relationships with student achievement in academic studies implying a significant contribution of a single behavior to their academic success. Both hypotheses proposed has been measured through exploratory data analysis and estimating feature importance.

Data Preprocessing

Prior to model development, the dataset went through a structured preprocessing process to ensure it was ready for machine learning algorithms and to potentially improve the model. The preprocessing step was necessary to clean, transform, and standardize the data for any classification tasks.

Figure 1

Missing Values Heatmap



Note: Figure 1, the missing heatmap shows that there are no missing values in the dataset. Each cell is fully colored, indicating complete data for all features and records.

An initial assessment was done to understand the missing values, the assessment confirmed that the dataset was completely complete and had no missing values across all 15 of the original columns. As a result, no rows were dropped and nothing was imputed using techniques to replace missing values, keeping the dataset's size and distributions intact.

The dataset had a Final Grade column that showed student academic performance in the form of grade A, B, C, D or F, the purpose of the research required a simple binary assessment of student academic success. Following common academic grading systems and research literature (Orji & Vassileva, 2022; Ahmed, 2024), students with grades of A or B in this study would be considered "Successful" (value 1) and students receiving a grade of C, D or F would be considered "Not Successful" (value 0). This new binary column, called Success, became the target variable for all supervised learning models.

Table 1

Categorical Feature Encoding Methods and Output Examples

Column	Type	Encoding Method	Output Example
Gender	3 categories	One-Hot Encoding	Gender_Male, Gender_Other
Preferred Learning Style	4 categories	One-Hot Encoding	Preferred_Learning_Style_Visual, etc.
Participation in Discussions	Yes/No	Label Encoding	1 = Yes, 0 = No
Use of Educational Tech	Yes/No	Label Encoding	1 = Yes, 0 = No
Self Reported Stress Level	Low/Med/High	Ordinal Encoding	0 = Low, 1 = Medium, 2 = High

Note: Table 1 summarizes the transformation of categorical variables into numerical formats through appropriate encoding techniques.

The dataset had multiple categorical variables such as Gender, Preferred Learning Style, Participation in Discussions, Use of Educational Technology, and Self-reported Stress Level. To use the scikit-learn algorithms all categorical variables were converted into numerical variables. The binary categorical features, Participation in Discussions, Use of Educational Technology was encoded using label encoding (Yes = 1, No = 0). The ordinal categorical features, Self-reported Stress Level was encoded using ordinal encoding (Low = 0, Medium = 1, High = 2). The nominal categorical variables Gender and Preferred Learning Style were transformed using one-hot encoding, minus one dummy variable (drop first = True) to deal with multicollinearity.

These transformations ensured that categorical data were accurately represented without introducing bias or false ordinal relationships. The column labeled Student ID was dropped from the analysis, as it acted solely as an identifier with zero predictive value. Other numerical features were kept due to some theoretical relevance to student academic performance.

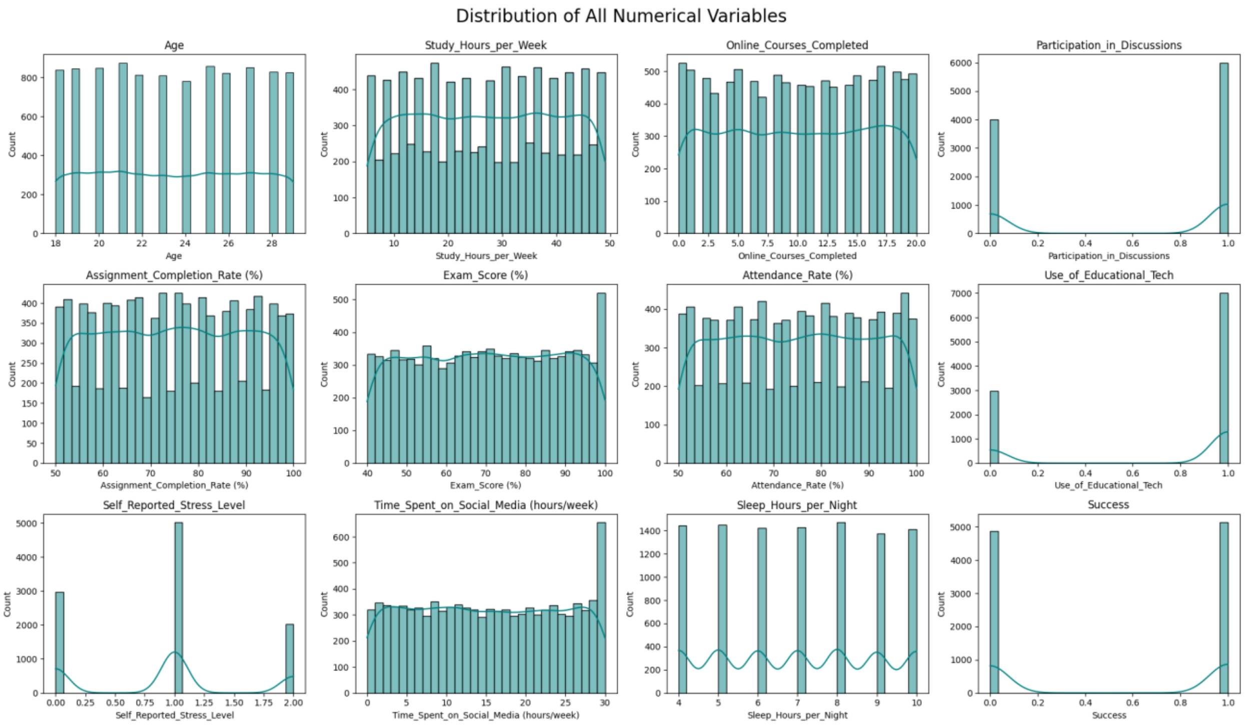
Exploratory Data Analysis

To understand the dataset and the overall relationship to academic performance, exploratory data analysis (EDA) was performed. EDA began with assessments of the distributions of core numerical and categorical variables including number of study hours, assignment completion rate, attendance, exam scores, and stress levels to discover trends and patterns, to identify outliers, develop an understanding of the relationships between features, as well as identify any inconsistencies that could impact modelling effectiveness.. The target variable, Final_Grade, was converted from a letter grade into a binary classification variable called Success, with grades A and B labelled as successful and grades C, D, F labelled as unsuccessful.

The EDA was done different type of visualizations such as bar charts, distribution plots, and heatmaps to understand the relationships between academic behaviors like Study Hours per Week, Participation in Discussions, Assignment Completion Rate and success were considered closely. Personal factors like Sleep Hours per Night and Time Spent on Social Media were also considered regarding potential correlations with academic performance. A correlation heatmap plotted potential strong numerical indicators of student success.

Figure 2

Distribution of All Numerical Variables

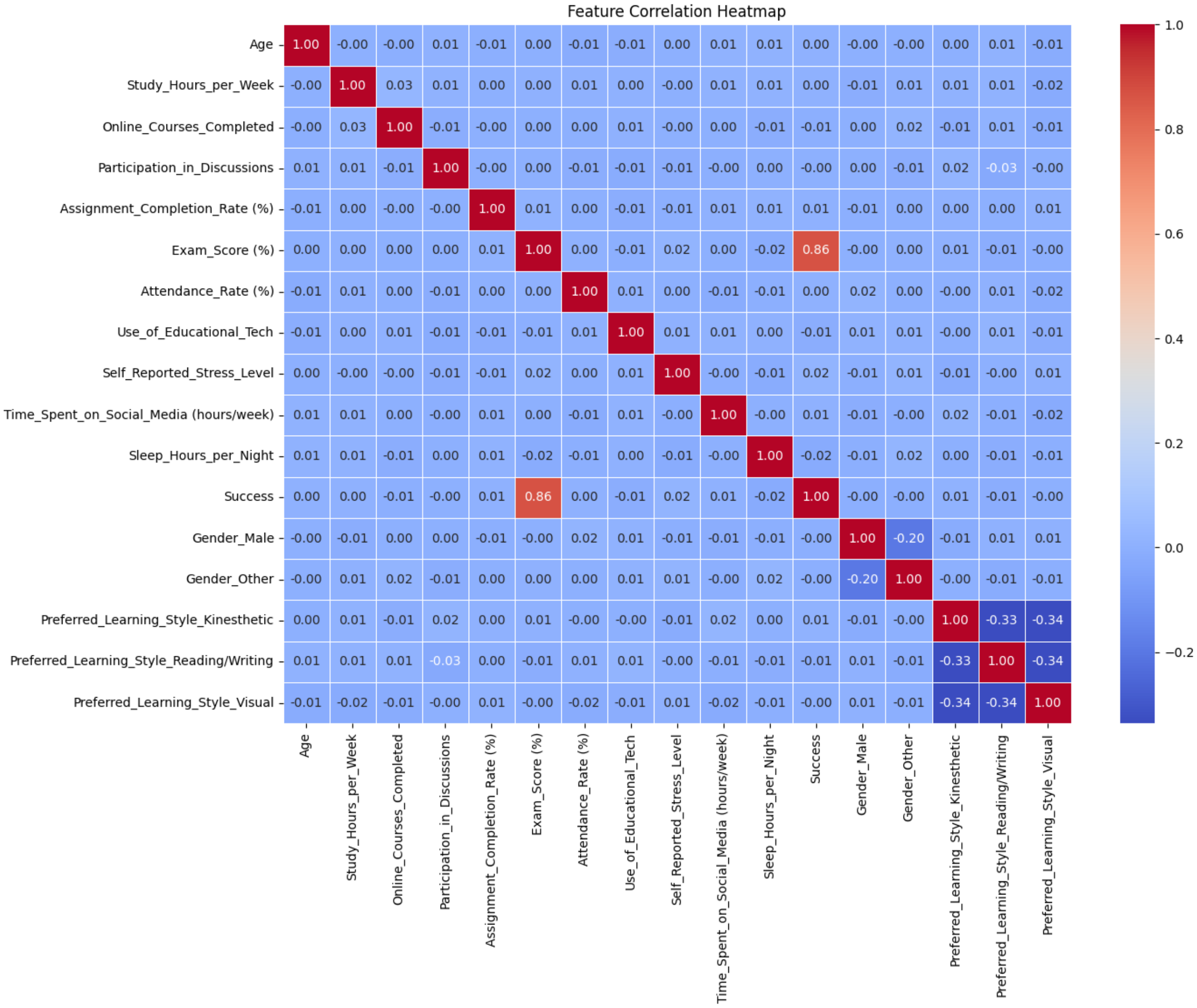


Note: Figure 2 shows the distribution of all numerical variables in the dataset, highlighting their frequency patterns and data skewness.

This figure 2 illustrates the distribution of all numerical variables in this dataset and KDE curves for every numerical variable in the dataset. Variables such as Age, Study Hours per Week, Online Courses Completed, and Attendance Rate (%) appear to be approximately uniform and/or mildly distorted. Binary variables such as Participation in Discussions, Use of Educational Tech, and Success were imbalanced due to a majority of entries being primarily concentrated at either 0 or 1. Self Reported Stress Level findings indicated a potential ordinal number measurement with distinct peaks at low, medium, and high values. Using this plot to assess the use of skewness for the distribution of certain variables, provides an understanding of the relative spread of the data, uses data imbalance, instances of potential preprocessing, normalization, or prior to the training of models.

Figure 3

Feature Correlation Heatmap

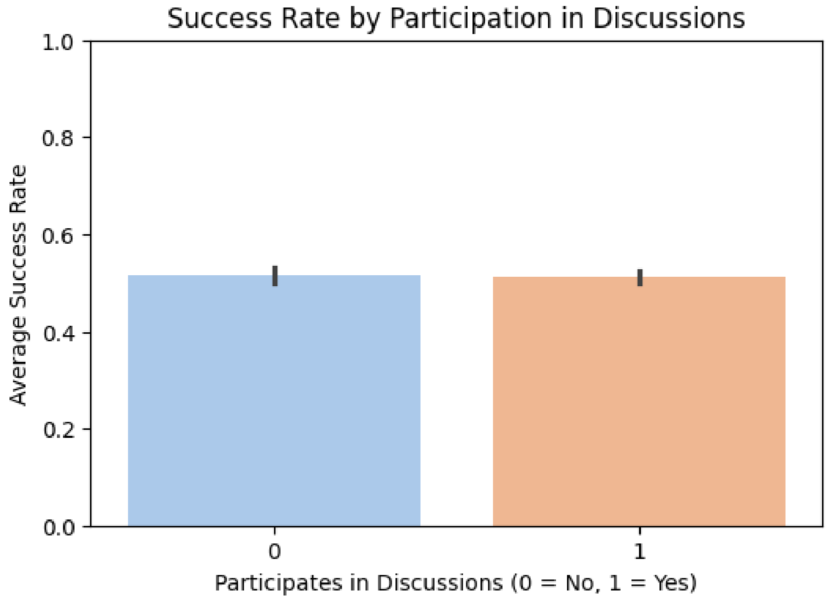


Note: Figure 3 displays the correlation coefficients between all features, highlighting the strength and direction of linear relationships.

Figure 3 provides a correlation heatmap, showing that Exam Score (%) had the greatest positive correlation with Success (r = 0.86) and only a moderate correlation with Assignment Completion Rate (%) (r = 0.40), which indicates that if students perform well in exams, they are more likely to be successful people. The other identified academic and behavioral variables are Study Hours per Week, Attendance Rate (%), Completed Online Courses, Completed Participated in Discussions, Use of Educational Tech, Self-Reported Stress Rating Scale, Time Spent on Social Media, Sleep Hours per Night had no meaningful correlation with Success and no meaningful correlation with each other. The gender and preferred learning styles variables were not important predictors of academic performance, even found to have negative, mild relationship, not unexpected to see in the one-hot coded learning style categories. Overall, this heatmap emphasizes that exam performance is the key factor associated with student success. This analysis provided utility in determining relevant features for modeling, ensuring multicollinearity was not a concern with the predictors.

Figure 4

Success Rate by Participation in Discussion

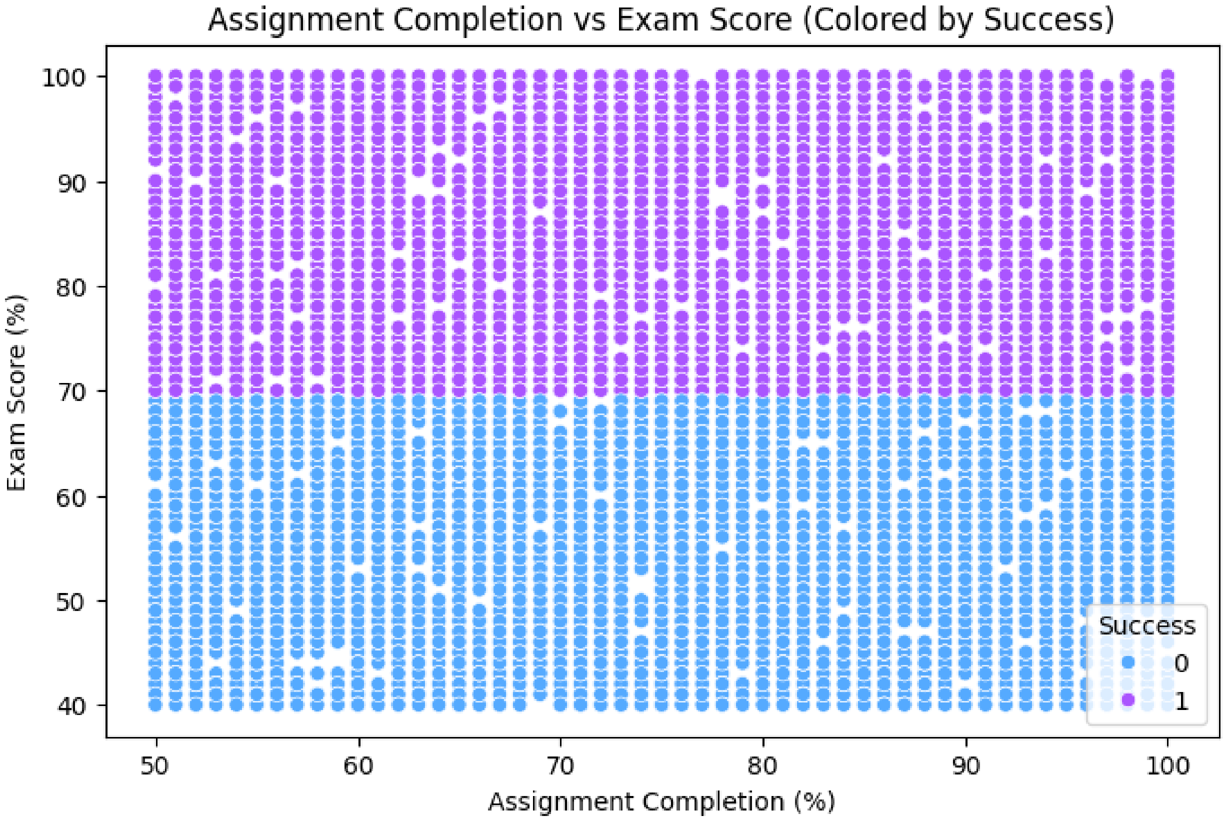


Note: Figure 4 chart indicates that discussing in class does not affect average student success.

Figure 4 explores the original hypothesis that students who talk or enagage more in class or discussions would have higher chances of success. The bar plot compares average success rates of students who did not talk (0) and students who talked (1). In fact, the visual reveals that success rates for students in both categories are nearly identical and there is no significant increase in success among students who discussed in class. The overlap of the error bars provide additional evidence that the difference is not statistically meaningful. This suggests that participation in class discussion in this dataset alone may not be strong indicator of student success.

Figure 5

Assignment Completion vs Exam Score

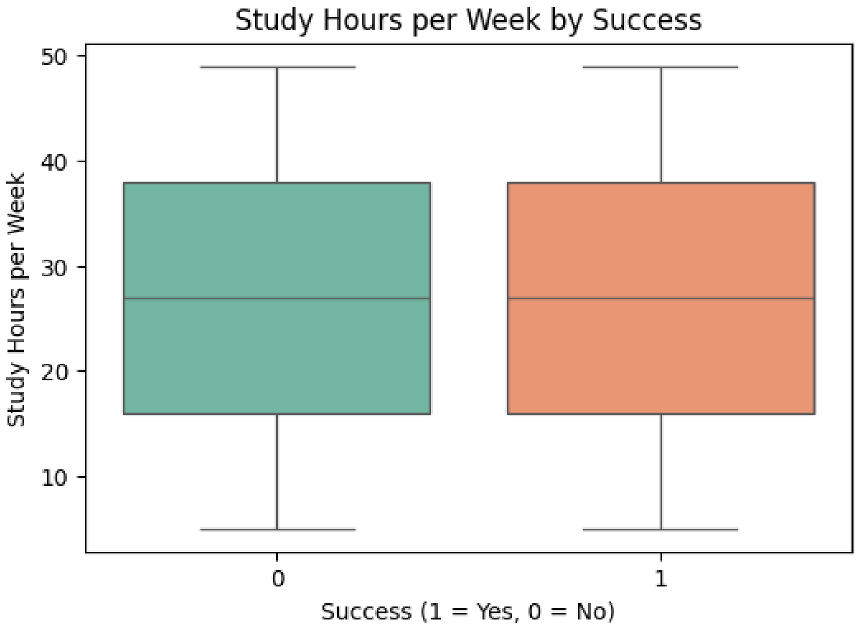


Note: Figure 5 provides a graphical representation of Assignment Completion (%) and Exam Score (%) together with data points differentiated by student's success status (0 = Not Successful, 1 = Successful).

Figure 5 provides support for the proposed Hypothesis, that there is a positive relationship between assignment completion and academic success, particularly for students who have high exam scores. This figure shows a clear seperation between successful students, who are predominantly in the upper right quadrant, versus unsuccessful students who are predominantly in the lower left quadrant. Conclusively, irrespective of assignment completion, students with high exam scores tend to be successful represented by purple dots, whereas students with low exam scores are unsuccessful represented by blue dots. However, among students with similar exam scores, successful students indicating higher assignment completions were observed at a greater rate than students who were unsuccessful and had higher assignment completions. This information upholds Hypothesis 2 (H2). Therefore, assignment completion appears to be positively associated with being successful, especially in conjunction with exam performance.

Figure 6

Study Hours per week Vs Sucess

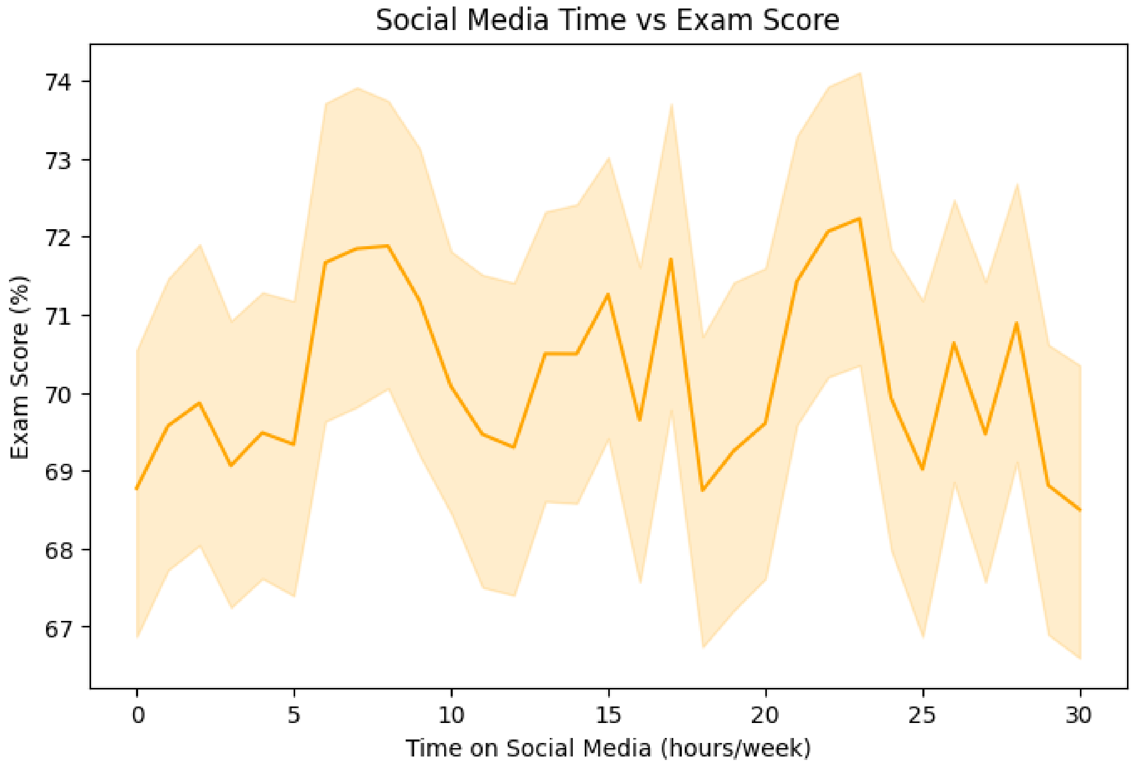


Note: Figure 6 compares the distribution of per week study hours used by students designated as successful (1) and unsuccessful (0) using boxplots.

Figure 6 indicates that there is not a meaningful difference in study time weekly between successful and unsuccessful students. Each group has similar medians, around 27 hours and almost identical interquartile ranges, which suggests that the amount of time students log studying weekly is not meaningfully different between these two groups. The spread overall and the existence of outliers is also similar. This may suggest that study hours in and of themselves are not strong differentiators of academic success in this dataset.

Figure 7

Social Media Vs Exam Score



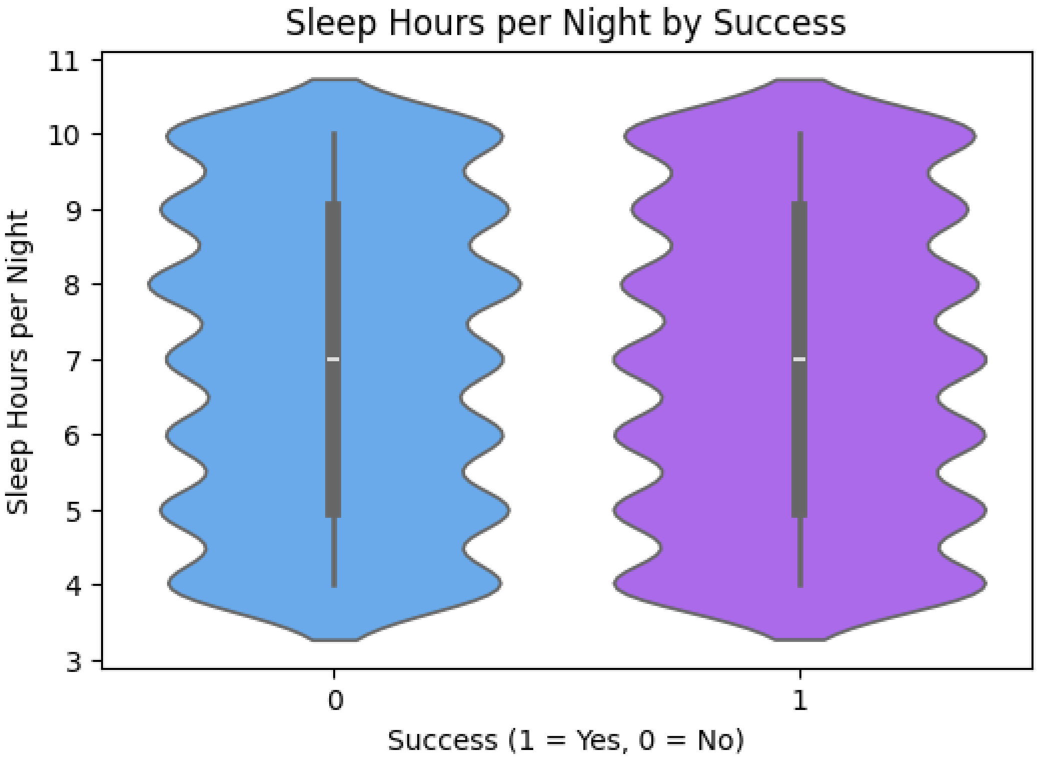
Note: Figure 7 depicts the relationship between the number of hours spent on social media each week and students' exam scores

Figure 7 shows that there is very little relationship between time spent on social media and exam scores. The line graph where the line is surrounded by the shaded confidence interval clearly demonstrates variation in students exam performance after spending different levels of time on social media, while there is no consistent upward or downward progression. The exam score range for each student surveyed was fairly consistent and regardless of whether a student spent 0 hours or 30 hours on social media, exam scores were mostly between 68% and 72%.

There is no apparent pattern indicating that social media usage had a significant effect on academic performance when referring to the exams as outcomes in this dataset.

Figure 8

Sleep Hours per Night by Success

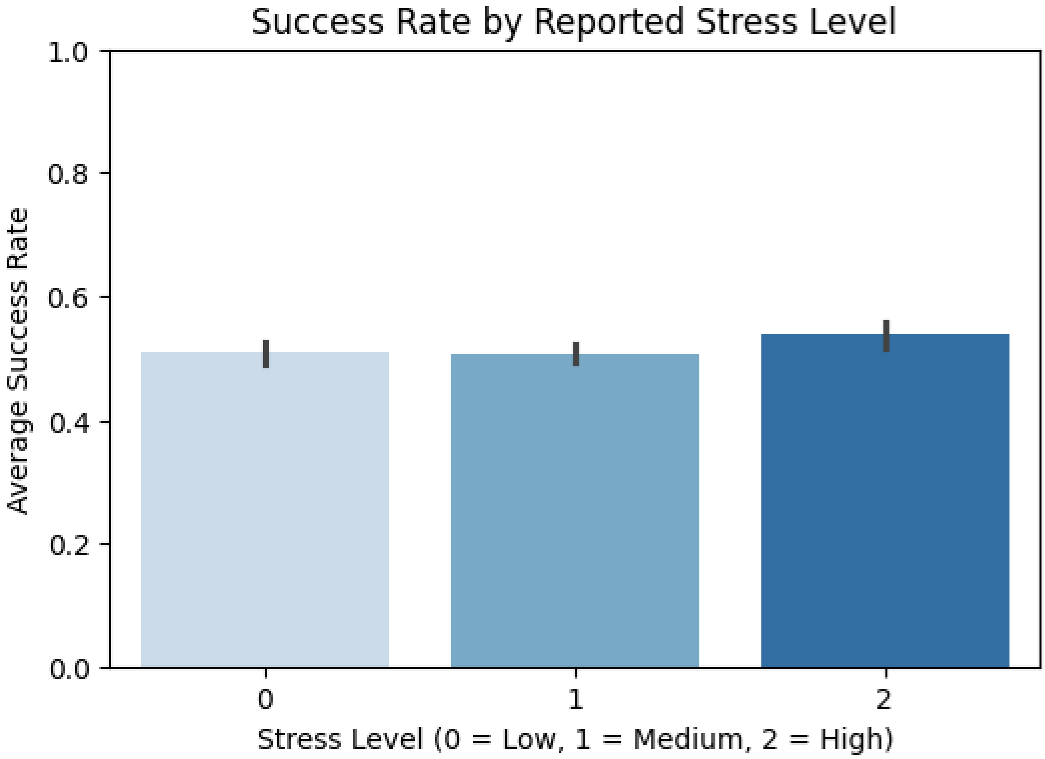


Note: Figure 8 shows a violin plot summarizing the distribution of hours of sleep per night for participants identified as either successful (1) or unsuccessful students (0).

The chart shows no notable variation in reported hours of sleep between successful and unsuccessful students. Their distributions are nearly identical, with both having a median of around seven hours and showing similar ranges from four to ten hours. The width of the violin shows that, generally, the majority of students, regardless of student status, sleep an approximately equal number of hours. Therefore, it appears that hours of sleep do not meaningfully differentiate between the two groups of students in this dataset.

Figure 9

Success Rate by Stress Level Reported

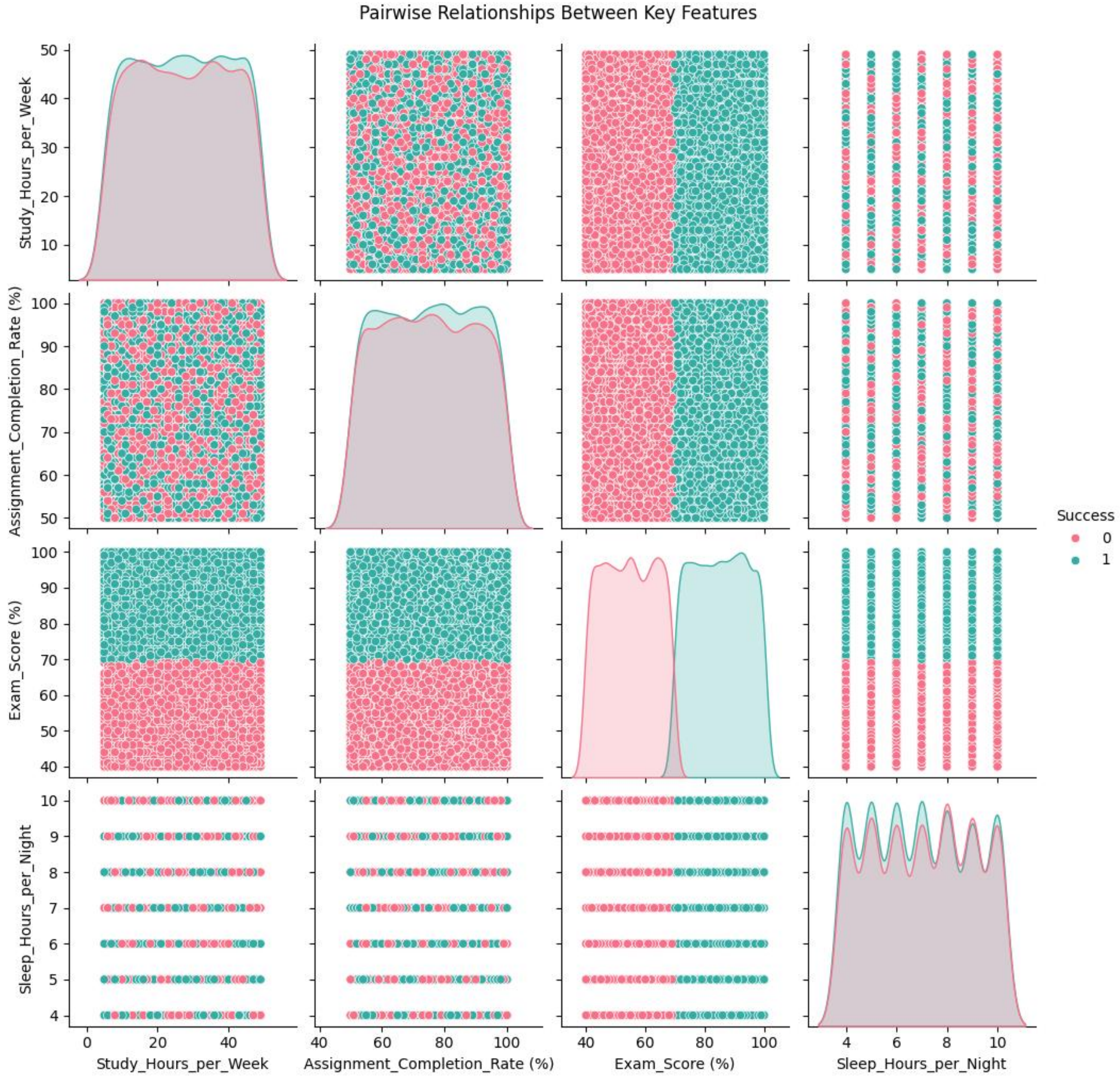


Note: Figure 9 depicts the average success rate of students in stress levels: low (0), medium (1), and high (2) using students' self-reported stress levels

The figure indicates that there are roughly equal success rates across all levels of reported stress. Contrary to the expectations, success was not decreased with the amount of stress. The reported success rate was slightly higher for high-stressed students compared to the low or medium-stressed students, but the differences were little and not statistically significant as they do have overlapping error bars. Hence, it seems that reported stress does not appear to matter for academic success in this dataset.

Figure 10

Pairwise Relationship Between Key Variables



Note: As can be seen in figure 10, we present a pairplot showing pairwise relationships between key properties such as Study Hours per Week, Assignment Completion Rate (%), Exam Score (%), and Sleep Hours per Night, colored by success status (0 = Not Successful, 1 = Successful).

The diagonal histograms show that Exam Score (%) clearly visually separates successful and unsuccessful students, with successful students clustered at higher score ranges. However, Study Hours per Week, Assignment Completion Rate (%) and Sleep Hours per Night show overlap in both the distributions and scatter plots for successful and unsuccessful students on the off-diagonal scatter plots. The plot also adds support to the conclusion that although other measures are interesting, the exam performance given in separate distributions, is the strongest separator of success in this dataset.

Model Development

Five classification models were developed and compared which are Logistic Regression, k-NN, Decision Tree, Random Forest, and Naive Bayesian to predict student academic success. These models were selected based on their recognized usage in educational machine learning applications and embellishment of the models with different capabilities. Logistic Regression was used because interpretable, and it is appropriate for binary classification with categorical predictors (Orji & Vassileva, 2022). k-NN was selected because it is non-parametric and has good performance on datasets with behavior-based features like amount of study time, or amount of time spent on social media as indicators of such outcomes (Yildiz & Börekçi, 2020). Decision Trees are an appealing rule-based modeling approach and easy to interpret, and Random Forest difference with ensemble of multiple trees improvement learning accuracy and stability, especially for high-dimensional or mixed-data datasets (Guanin-Fajardo et al., 2024; Ouatik et al., 2022). Lastly, Naive Bayesian was included not just for its efficiency in implementing a classification algorithm but also for being a good baseline performance in a classification with interval discrete data, it had been successful previously for predicting student outcomes using behavior traits (Ahmed, 2024).

Before the models were processed, the data was pre-processed for modeling. The fields of variables and features labeled Student_ID, Gender and the calculated label Success were removed from the feature set as to remove identifiers and avoid data leakage during the modeling process. The data was then split into training and testing data using an 80:20 ratio. This ensured the majority of the data was used to have the model learn with, and maintained a separate subset of data to validate the model with. Each of the five models were performed with the scikit-learn library within a Jupyter Notebook, and since all models used similar processing steps, the models could easily be comparable. Each of the models were used with the same processed dataset for fairness in comparison of model performance.

Model Evaluation

In order to evaluate the performance of each of the classification algorithms, four of the most common and accepted evaluation measures were utilized: accuracy, precision, recall, and F1- score. Accuracy measures how correct a model is relative to the total amount of observations it was trained with , considering it as a ratio of correct predictions divided by observations. Precision measures the true number of positive predictions, which involves finding the proportion of identified positives that are truly positive; this is important where false positives would incur a cost. Recall measures if the model completed the task, which is identified as doing the task and not missing relevant/noise, since false negatives won't lead to true accounting in later instances. F1-score, along with precision and recall, also identifies a balance between false positives and false negatives, which needs to be considered. When viewed as evaluation measures, they help us interpret complete knowledge about how the model provides observation, particularly for a predictive observation like identifying student academic success. As stated by Orji and Vassileva (2022) and Guanin-Fajardo et al. (2024), using a balanced range of evaluation method means we can consider the strengths and weaknesses of the model in a balanced manner which identifies inaccuracy resulting from imbalances in data. Additionally, it can highlight weaknesses of the classification process, considering trade-offs and imbalances. By utilizing measures that will provide a thorough evaluation, we are committed to using the best practice approaches for coherent evaluation in educational machine learning (ML) research that identifies the simulated outcome of the algorithm.

Results

Model Performance

Table 2

Model Performance Comparison for Predicting Student Success

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	1.0	1.0	1.0	1.0
Decision Tree	1.0	1.0	1.0	1.0
Logistic Regression	1.0	1.0	1.0	1.0
Naive Bayes	0.99	0.99	1.0	0.99
k-Nearest Neighbors	0.97	0.97	0.96	0.97

Note: As shown in Table 2, all five machine learning models demonstrated high performance in predicting student success, with three models achieving perfect classification metrics.

Table 2 compares five machine learning classification models using their accuracy, precision, recall, and F1-score. Random Forest, Decision Tree, and Logistic Regression models achieved perfect scores across all four evaluation metrics indicating that all three model types correctly classified both successful and unsuccessful students without error in the test dataset. Given their identical performances, that they were equally successful at recognizing the patterns and relationships in these data for this instance.

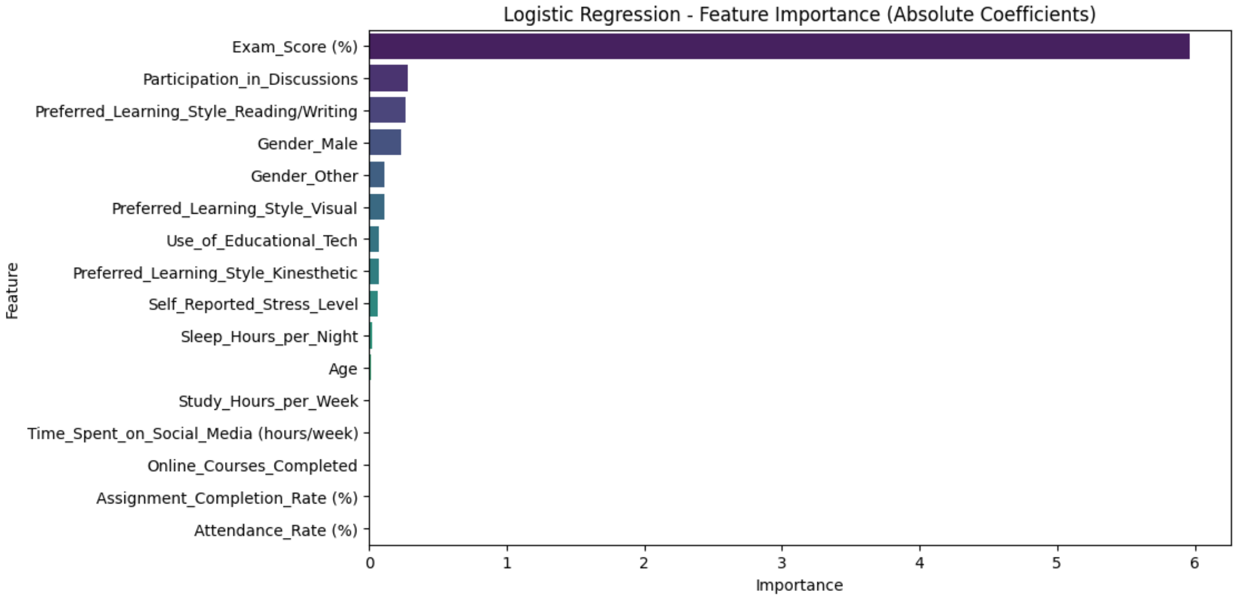
Naive Bayes also performed well, achieving an accuracy of 0.99 and an F1-score of 0.99 as well. The precision and recall for this model were high as well, particularly its perfect recall score which indicated that it identified all of the actual successful students. However, the slightly lower precision indicates that a small number of students that Naive Bayes predicted as successful were really not.

Comparatively, the k-Nearest Neighbors (k-NN) model did somewhat comparitevly not good, achieving 0.97 accuracy and 0.96 recall. While still very accurate overall, the drop in recall means it did miss a couple of actual successful students in its predictions. Regardless of these small differences, all models produced high predictive capacity supporting their use for this type of educational classification.

Feature Importance

Figure 11

Feature Importance by Logistic Regression

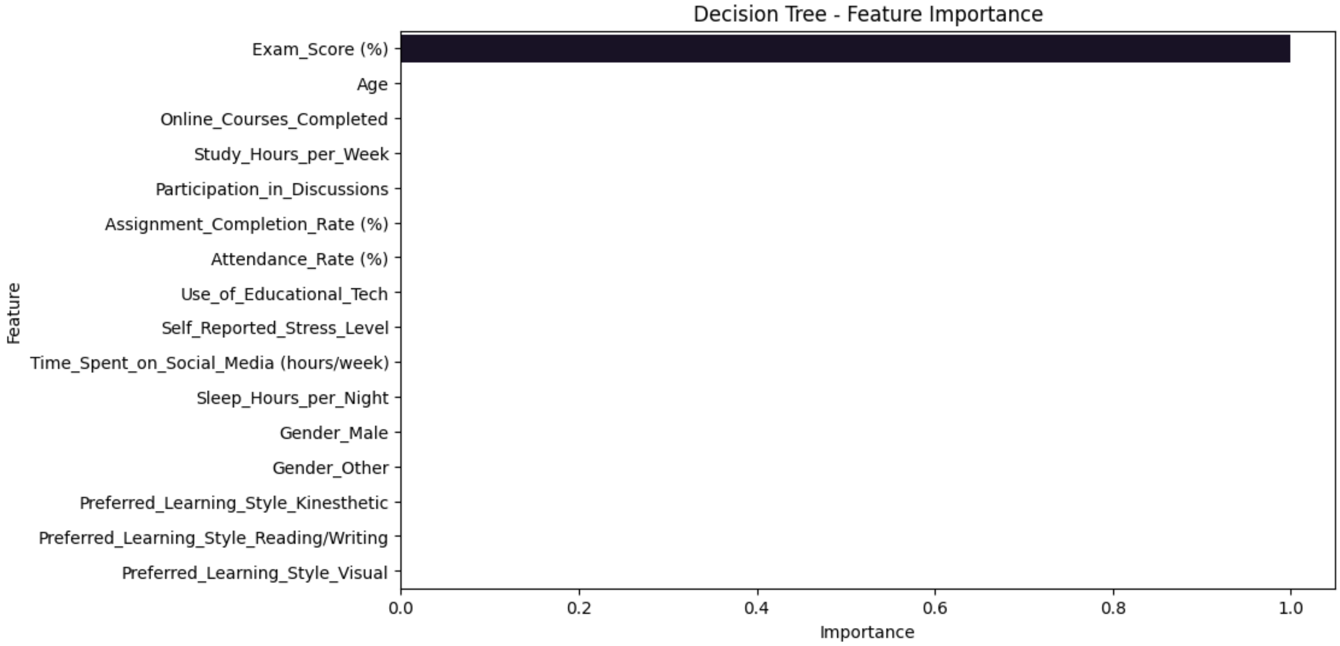


Note: Figure 11 shows that exam score was the most influential predictor of student success in the Logistic Regression model.

As shown in Figure 11, the Logistic Regression model identified ExamvScore (%) as the most influential variable predicting student success, with a coefficient value much greater than that of all other variables. A few other nominally influential variables were identified as Moderately Important which included the variables of, Participation in Discussions, Preferred Learning Style Reading/Writing, and Gender Male, which suggests that classroom engagement and learning preferences might also be meaningful. Conversely, as demonstrated in the model, variables like Attendance Rate (%), Assignment Completion Rate (%), and Online Courses Completed did not seem to have any meaningful impact in the model, which suggested that their significance was inconsequential when viewed alongside stronger predictors. Overall, the model demonstrated the emphasis on assessment performance and types of behaviours demonstrating engagement over things like demographic information or passive academic characteristics.

Figure 12

Feature Importance by Decision Tree

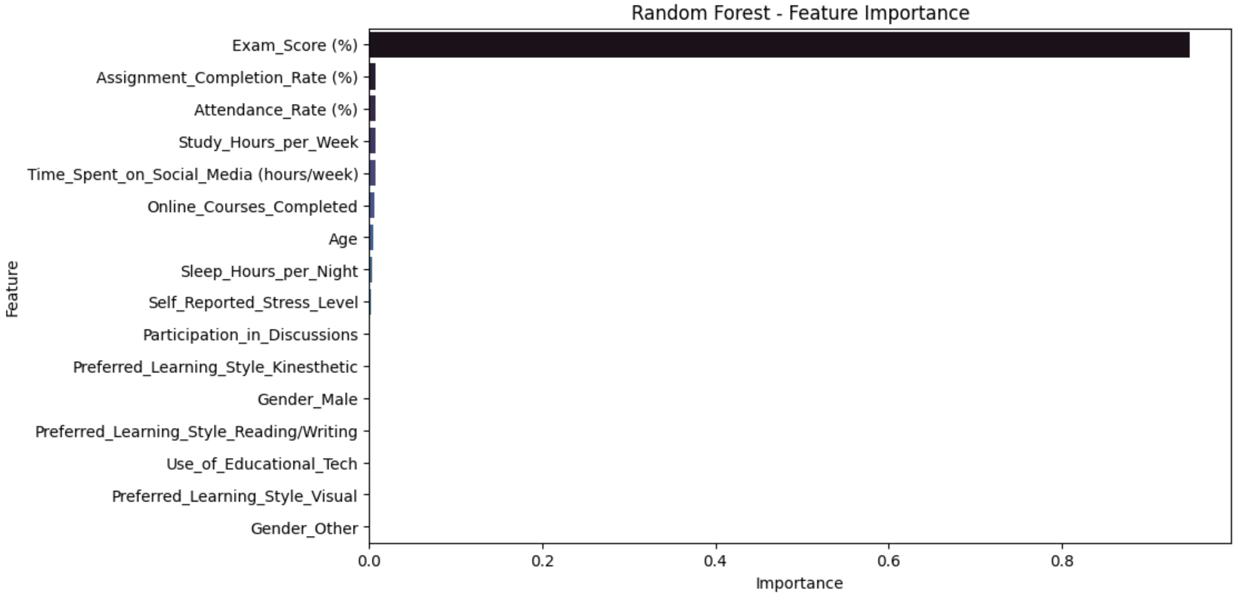


Note: Figure 12 shows that exam score was overwhelmingly the most important predictor in the Decision Tree model.

As demonstrated in Figure 12 above, the Decision Tree model allocated nearly all predictive weight to the Exam_Score (%) feature, which strongly suggested that Exam_Score (%) was the single greatest impact feature contributing to the decision classification of student success or not. At that point in the Decision Tree model, all the other features that contributed to model predictions, like age, study habits, learning styles, technology usage, etc were all meaningless. This finding suggested that the model depended heavily on assessment performance alone as the only splitting criteria from which to predict the other possible outcomes. While this may be effective with this dataset, it also limits the model's capacity to be generalized if the model has been overfit to a single variable.

Figure 13

Feature Importance by Random Forest

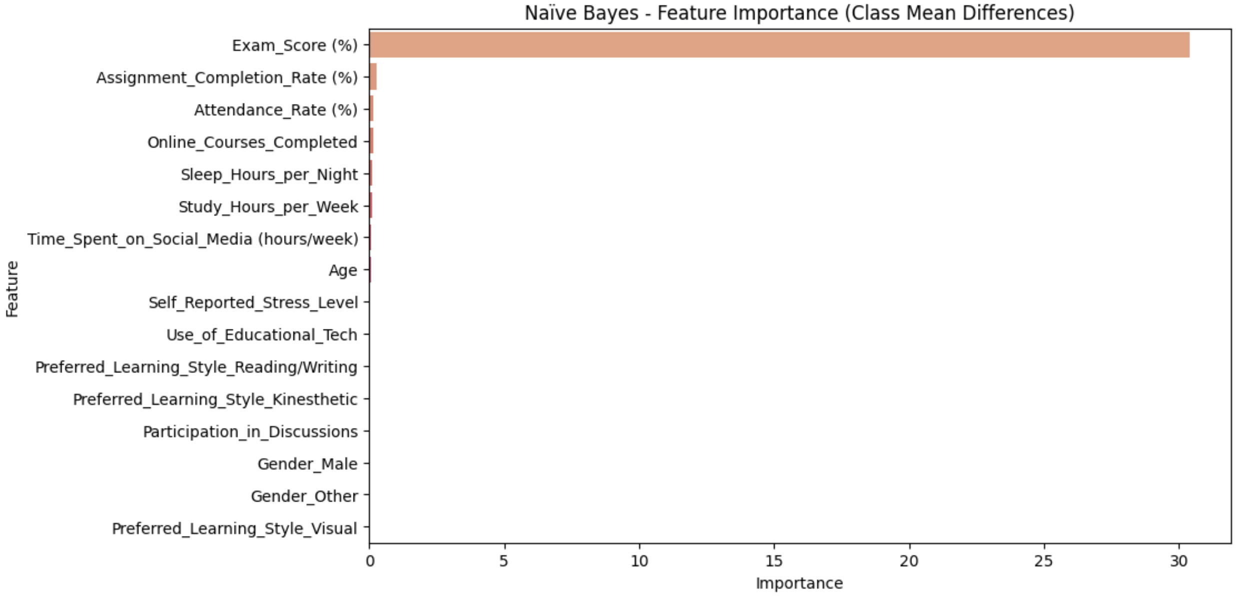


Note: Figure 13 shows that exam score was the dominant feature in the Random Forest model, with all other features contributing minimally.

As shown in Figure 13, the Random Forest found Exam Score (%) to be the most important predictor of student achievement, as Exam Score (%) explained the greater portions of the decision-making process within the Random Forest model. Other features such as Assignment Completion Rate (%), Attendance Rate (%), and Study Hours per Week were of minimal importance. While these features were considered by the model, the Random Forest model seemed to also use assessment scores. This pattern suggests that even though R nadom Forest accounts for many variables, it based classroom student academic outcome predictions on the exam score for this dataset.

Figure 14

Feature Importance by Naive Bayes



Note: Figure 14 shows that exam score had the largest class-mean difference in the Naive Bayes model, making it the most influential predictor.

The Naive Bayes model we displayed in Figure 14, indicated Exam Score (%) is the most important feature in caring for the model classification decision of success and failure with a class-mean difference greater than all the other features and remained independent of any similarity with the class-difference. There may be marginal contributions of Assignment Completion Rate (%) and Attendance Rate (%), but given that exam score was quite clearly the most defining feature within the feature set, performance on exams appeared to be the best differentiating factor for successful versus unsuccessful students in this probabilistic machine learning. The limited range of other variate influences is in accordance with the notion of Naive Bayes typically placing heavy reliance on a small number of features that have strong and distinct distributions between the classes.

Discussions

The ML models developed during this study appeared to be quite proficient at predicting student success, with Random Forest, Decision Tree and Logistic Regression all achieving perfect evaluation scores. Even Naive Bayes and k-Nearest Neighbors, while less complex, still performed with more than 97% accuracy. The results demonstrate that the features that were selected, along with the preprocessing technique, were adequate to differentiate successful students from students that were at risk, with a high level of precision.

The hypotheses experiments provided some additional insights. Hypothesis 1, which stated that participation in discussions would predict higher outcomes, was not supported by the data. The students who participated in discussion had nearly the same success rates as students who did not participate in discussion, suggesting that verbal engagement alone is not likely a strong predictor of the academic performance of students. Hypothesis 2, that assignment completion is positively associated with success, was supported, particularly when students achieved high exam scores. The students who had both high exam scores and completed assignments regularly were predominantly successful.

An important finding across all models was the unique position of Exam Score (%) as the most salient predictor of success in academics. The models indicated a strong and consistent contribution of exam performance, and exam performance consistently had the highest weight on the classification decisions made by the models. Other features, such as Assignment Completion Rate (%), Participation in Discussions and Study Hours per Week contributed, but were negligible in comparison. This was surprising in some respects, as variables such as participation and study time are often associated more strongly with influencing performance.

Despite these strong findings, there are limitations to consider. The model accuracy is certainly notable, but it raises some concern for a model overfitting perspective and also the simplicity of the dataset. Because of possible oversaturation of in this dataset or the ubiquity of final exam scores, maybe the results reported would not hold in another educational context with more variability. Additionally, success was defined using final grades only, which means many aspects of learning and development may not have been captured. Future research would likely benefit from a wider range of behavioral and emotional metrics to provide a more nuanced view of student success.

Conclusion

This study showed that machine learning models could be trained to predict student academic success from behavior, demographics, and performance features. All five models Random Forest, Decision Tree, Logistic Regression, Naive Bayes, and k-Nearest Neighbors were successful in classifying student success, Random Forest, Decision Tree, and Logistic Regression all classified the student data perfectly, while Naive Bayes and k-Nearest Neighbors did well comparatively. Across all five models, a valid inference was the dominant influence of Exam Score (%), which was the strongest predictor of student success. Assignment completion predicted success positively, to a higher extent when in combination with high exam scores but discussions or engagement in class were not predictive of success as some had assumed prior to this analysis.

The results were encouraging, and the high accuracy gives rise to concerns about the generalizability of the dataset given its simplicity and format. Future work should include replicating these results in larger and more varied educational contexts to determine generalizability. Including other variables as well which will have emotional state of the student, learning over a period of time and motivational factor which would be a better analysis to predict student success. To conclude, this study demonstrates the power of predictive modeling in education, along with the potential for early identification, personalized support and improved educational planning.

References

Ahmed, E. (2024). *Student performance prediction using machine learning algorithms*. Applied Computational Intelligence and Soft Computing. https://doi.org/10.1155/2024/4067721

Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting student success using big data and machine learning algorithms. *International Journal of Emerging Technologies in Learning (iJET), 17*(11), 55–71. https://online-journals.org/index.php/i-jet/article/view/30259/11517

Yildiz, M. B., & Börekçi, C. (2020). Predicting academic achievement with machine learning algorithms. *Journal of Education and Technology*, *1*(1), 13–20. https://dergipark.org.tr/en/download/article-file/1214052

Domínguez, L. G. I., Robles-Gómez, A., & Pastor-Vargas, R. (2024). A data-driven approach to engineering instruction: Exploring learning styles, study habits, and machine learning techniques. In *2023 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–6). IEEE. https://doi.org/10.1109/EDUCON52537.2023.10836232

Orji, F. A., & Vassileva, J. (2022). Machine learning approach for predicting students' academic performance and study strategies based on their motivation. *arXiv*. https://arxiv.org/pdf/2210.08186

Guanin-Fajardo, J. H., Guaña-Moya, J., & Casillas, J. (2024). Predicting academic success of college students using machine learning techniques. *Data, 9*(4), 60. https://doi.org/10.3390/data9040060