**Unveiling Key Predictors of Academic Success:**

**A Comprehensive Analysis of High School Student Performance**

Gabriela Bermudez, Daylin Hernandez, Jessica Krumm, Momisola Odeyemi, and Taylor Sain

**Introduction**

The educational success of high school students is influenced by a variety of factors, both academic and non-academic. Understanding these factors is crucial for educators, policymakers, and parents alike, as it can help in creating strategies that support students' academic growth and personal development. Over the years, various studies have focused on metrics like GPA and attendance, but these often provide only a partial picture of what contributes to student success.

In recent years, there has been an increasing recognition of the importance of considering a broader range of factors, including extracurricular activities, parental involvement, and students' study habits. These elements, although sometimes seen as secondary to academic metrics, play a significant role in shaping a student's ability to thrive in an academic environment. By examining these aspects, we can gain deeper insights into what drives academic success and how students can be better supported in their educational journeys.

The value of understanding these relationships extends beyond just improving grades. It encompasses enhancing student engagement, satisfaction, and overall well-being, which are critical components of a holistic educational experience. This study aims to explore these dimensions by analyzing a comprehensive dataset of high school students, with the goal of identifying key predictors of academic success and offering actionable insights for educators and stakeholders.

**Problem Statement**

Despite the extensive research on factors influencing student performance, there remains a need for a deeper understanding of how different variables interact to impact academic outcomes. Traditional metrics such as GPA and attendance rates provide limited insights into the complexities of student success. As emphasized by Schreiner (2012), student success encompasses not just academic achievement but also engagement in educational activities, satisfaction, and personal development (Schreiner, 2012, p. 3). This study aims to utilize the Student Performance Dataset to identify patterns and associations that can predict student success, using a combination of data mining techniques including classification, clustering, and association rule mining.

**Analysis**

*Subsection 1: The Data*

This section delves into the details of the dataset used for analysis, including the variables, their characteristics, and the data cleaning process. The dataset includes detailed records of individual students, characterized by multiple attributes related to personal demographics, educational history, and performance metrics. Key features include:

* **Student ID**: A unique identifier for each student.
* **Age**: Ranging from 15 to 18 years.
* **Gender**: 0 for Male, 1 for Female.
* **Ethnicity**: 0 for Caucasian, 1 for African American, 2 for Asian, 3 for Other.
* **Parental Education**: Education level of the parents, ranging from 0 (None) to 4 (Higher).
* **Study Time Weekly**: Weekly study time in hours (0 to 20).
* **Absences**: Number of absences during the school year (0 to 30).
* **Tutoring**: 0 for No, 1 for Yes.
* **Parental Support**: Level of parental support, ranging from 0 (None) to 4 (Very High).
* **Extracurricular Activities**: 0 for No, 1 for Yes.
* **Sports**: 0 for No, 1 for Yes.
* **Music**: 0 for No, 1 for Yes.
* **Volunteering**: 0 for No, 1 for Yes.
* **GPA**: Grade Point Average on a scale from 2.0 to 4.0.
* **Grade Class**: Classification of grades based on GPA:
  + 0: 'A' (GPA >= 3.5)
  + 1: 'B' (3.0 <= GPA < 3.5)
  + 2: 'C' (2.5 <= GPA < 3.0)
  + 3: 'D' (2.0 <= GPA < 2.5)
  + 4: 'F' (GPA < 2.0)

The dataset was carefully prepared to ensure its suitability for subsequent analysis, particularly for association rule mining (ARM) and clustering analysis. The following steps outline the data cleaning and preparation process:

* Checking for Missing Values: The dataset was examined for any missing values (NAs) to ensure data integrity. This step involved determining the extent of missing data. Where necessary, appropriate strategies were applied, such as removing rows or imputing missing values to maintain the quality of the dataset.
* Identification of Continuous Variables: continuous variables were identified within the dataset, such as GPA, age, and admission grades. These variables, which can take a wide range of values, required discretization to convert them into categorical data suitable for ARM.
* Discretization of Continuous Variables: Continuous variables were converted into categorical variables by binning their values into meaningful intervals. For example:

A screen shot of a computer code

Description automatically generated

* Creation of Dummy Variables: To prepare the data for analysis, particularly ARM, categorical variables—including those resulting from discretization—were converted into binary (dummy) variables. Each category within a variable was represented as a separate binary variable (0/1). This step was essential to meet the requirements of the Apriori algorithm used in ARM.
* Preparation for Association Rule Mining: The final step for ARM involved converting the cleaned and prepared data into a "transactions" format, a specific structure required by the Apriori algorithm. This made sure that the data was fully prepared for discovering interesting patterns and associations.
* Standardization/Normalization: Since clustering algorithms like K-means are sensitive to the scale of the data, normalizing the continuous variables to make sure that each variable was represented in the clustering process. This was achieved using standardization, where each variable was rescaled to have a mean of 0 and a standard deviation of 1.
* Handling Outliers: Outliers were identified and addressed to prevent them from distorting the clustering results. This involved detecting outliers using the Z-score method and considering their removal or treatment.
* Dimensionality Reduction (Principal Component Analysis (PCA)): was applied for dimensionality reduction, helping to identify the most important features while also reducing noise in the data. This approach addressed two common challenges in clustering—handling high-dimensional data and minimizing noise—ultimately improving the effectiveness and accuracy of the clustering results.

These comprehensive data cleaning and preparation steps, including the important checks for missing values, outlier handling, and standardization, ensured that the dataset was well-structured, consistent, and ready for advanced analysis techniques, such as association rule mining and clustering analysis. This thorough preparation laid the foundation for reliable and insightful exploration of relationships and patterns within the data.

*Subsection 2: Exploratory Data Analysis (EDA):*

Exploratory Data Analysis (EDA) is a crucial step in gaining a comprehensive understanding of the dataset. In this process, key features such as demographics, academic performance, and student activities are examined. Visual tools like histograms, count plots, and correlation matrices are used to identify patterns, explore relationships, and detect any unusual data points. The insights gained from EDA inform decision-making in subsequent stages of analysis and modeling. This initial exploration ensures a thorough understanding of the data before progressing to more complex analyses.

Distribution Plots

A group of graphs showing different types of data

Description automatically generated

* Distribution of Age:
  + Most students are aged between 15 and 17 years.
  + This histogram shows the distribution of ages in the dataset.
  + The ages are evenly distributed across the range with some peaks and valleys, indicating the data might have been collected across a fixed range of ages.
* Distribution of StudyTimeWeekly:
  + The distribution shows that a large number of students study less than 10 hours per week.
  + This histogram shows how much time students spend studying per week.
  + The distribution is relatively uniform, indicating that study times vary widely among students without any significant peaks.
* Distribution of Absences:
  + This histogram displays the number of absences.
  + The absences show a uniform distribution, with a peak around 20 absences, indicating a common occurrence.
* Distribution of GPA:
  + This histogram shows the distribution of GPA scores.
  + The distribution is roughly normal, centered around a mean value, indicating most students have an average GPA.

Summary Statistics

**A screenshot of a computer screen

Description automatically generated**

* Age: The average age is approximately 16.47 years with a standard deviation of 1.12 years.
* StudyTimeWeekly: Students spend an average of about 9.77 hours per week on studying, with a standard deviation of 5.65 hours.
* Absences: The average number of absences is around 14.54 days, with a standard deviation of 8.47 days.
* GPA: The average GPA is approximately 1.91, with a standard deviation of 0.92.
* Categorical Variables: Gender, Ethnicity, ParentalEducation, Tutoring, ParentalSupport, Extracurricular, Sports, Music, Volunteering, and GradeClass.

**Count Plots:**

A screenshot of a graph

Description automatically generated

* Count of Gender:
  + The count plot shows the distribution of gender.
  + The counts are nearly equal, showing a balanced gender distribution in the dataset.
* Count of Ethnicity:
  + This plot shows the distribution of different ethnicities.
  + The majority of students belong to the first category (0), with fewer students in other categories.

* Count of Parental Education:
  + This plot shows the education levels of the parents.
  + Most parents have an education level of 2, with a few at level 0 and level 4.
* Count of Tutoring:
  + This plot shows whether students receive tutoring.
  + A significant number of students do not receive tutoring.
* Count of Parental Support:
  + This plot shows the level of parental support.
  + Most parents provide level 2 or level 3 support.
* Count of Extracurricular:
  + This plot shows whether students participate in extracurricular activities.
  + Many students do not participate in extracurricular activities.
* Count of Sports:
  + This plot shows whether students participate in sports.
  + A significant number of students do not participate in sports.
* Count of Music:
  + This plot shows whether students participate in music.
  + A significant number of students do not participate in music.
* Count of Volunteering:
  + This plot shows whether students participate in volunteering.
  + A significant number of students do not volunteer.
* Count of GradeClass:
  + This plot shows the grade class distribution.
  + Most students are in the highest grade class (4 = F grade).

Correlation Matrix:

A screenshot of a graph

Description automatically generated

* + This matrix shows the correlation coefficients between all pairs of features.
  + Positive correlations are in red, negative correlations in blue.
  + The intensity of the color shows the strength of the correlation.
  + Key observations:
    - There is a strong negative correlation between GPA and Absences (-0.92). So as absences go up the students GPA will go down.
    - There is a strong positive correlation between GradeClass and Absences (0.73). So as GradeClass increases so do absences.
    - Most other features show weak correlations with each other.

The exploratory data analysis reveals several key insights into the dataset. The features of age, study time, absences, and GPA are relatively uniformly distributed, indicating a diverse range of data points across these variables. The dataset is balanced in terms of gender but shows a skew towards Caucasians. Parental education and support levels are generally high, which may positively influence student performance. However, participation in extracurricular activities, including sports, music, and volunteering, is relatively low, which could potentially impact overall student development. The correlation analysis highlights a strong negative relationship between GPA and absences, suggesting that students with more absences tend to have lower GPAs. Additionally, there is a strong positive correlation between grade class and absences, indicating that students in higher-grade classes may have more absences. These visualizations provide a comprehensive understanding of the distributions, relationships, and potential impacts of various features on student performance.

*Subsection 3: Data Mining Methods*

Various data mining algorithms will be applied to extract patterns and insights:

* Association Rule Mining: To discover interesting relationships between variables, such as study habits associated with high GPA.
* Clustering: Grouping similar data points to identify segments with similar characteristics.
* Classification: Using decision tree models, Naïve Bayes classifiers, SVM, and Random Forest to predict outcomes based on attributes like study time and extracurricular participation.

**Results:**

The results section presents the findings from each analysis method, highlighting the strengths and weaknesses of each approach. For example, the SVM and Random Forest models were found to be the most effective in predicting student grades, as evidenced by their high cross-validation accuracy and well-defined ROC curves. In contrast, models like kNN and decision trees showed lower performance, particularly in distinguishing between certain grade categories.

Each method's performance was supported by visualizations, such as confusion matrices and ROC curves, which provided a clear depiction of how well the models classified students across different grade classes. The results also identified specific patterns, such as the strong impact of absences on student performance, reinforcing the importance of regular attendance for academic success

*Subsection 1: Association Rule Mining (ARM) Analysis*

The analysis involves examining the student performance dataset through various data mining techniques, specifically association rule mining (ARM) and decision trees. The dataset contains comprehensive information on 2,392 high school students, detailing their demographics, study habits, parental involvement, extracurricular activities, and academic performance. The primary objective is to identify patterns and associations that can predict student success.

Association rule mining was applied to identify relationships between various variables in the dataset. The results are presented, sorted by support, confidence, and lift, to emphasize the most significant and noteworthy rules.

**Top Rules by Support**

**A diagram of a network

Description automatically generated with medium confidence**

* Volunteering = No:
  + Support: 0.8428
  + Confidence: 0.8428
  + Lift: 1.0
  + A significant portion of students do not participate in volunteering.
* Music = No:
  + Support: 0.8030
  + Confidence: 0.8030
  + Lift: 1.0
  + A large number of students do not participate in music-related activities.
* Tutoring = No:
  + Support: 0.6986
  + Confidence: 0.6986
  + Lift: 1.0
  + A significant section of students do not receive tutoring.

**Top Rules by Lift**

A diagram of a school

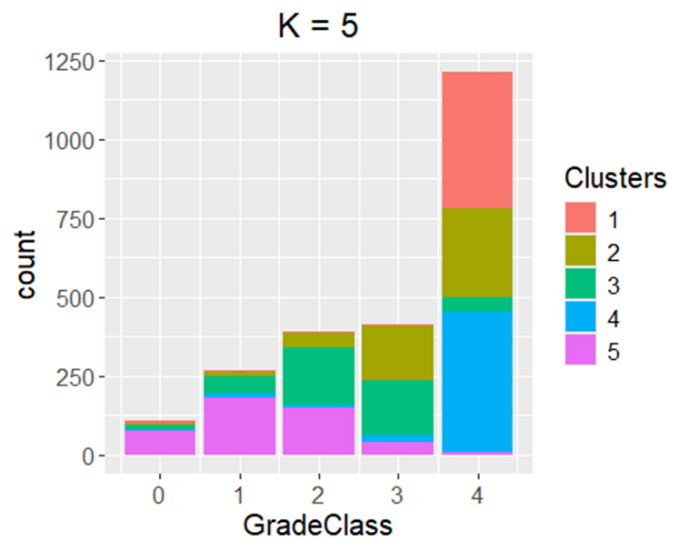
Description automatically generated with medium confidence

* GPA Category = High, Absence Range = Very Low, Tutoring = Yes, Parental Support = Moderate, Volunteering = No -> Grade Class = B:
  + Support: 0.0117
  + Confidence: 0.7568
  + Lift: 6.7292
  + Students with a high GPA, very low absences, who receive tutoring and moderate parental support, and do not volunteer, are highly likely to get a 'B' grade.
* Absence Range = Very Low, Tutoring = Yes, Parental Support = Moderate -> Grade Class = B:
  + Support: 0.0117
  + Confidence: 0.7368
  + Lift: 6.5521
  + Students with very low absences, who receive tutoring and have moderate parental support, are very likely to get a 'B' grade.
* GPA Category = High, Absence Range = Very Low, Ethnicity = Caucasian, Tutoring = Yes, Sports = No, Volunteering = No -> Grade Class = B:
  + Support: 0.0117
  + Confidence: 0.7179
  + Lift: 6.3841
  + Caucasian students with a high GPA, very low absences, who receive tutoring, do not participate in sports or volunteering, are very likely to get a 'B' grade.

*Subsection 2: Clustering Analysis*

The clustering analysis conducted on the student performance dataset provided valuable insights into the factors that influence academic outcomes. The dataset, comprising 2392 observations with both categorical and numerical variables, was carefully prepared and preprocessed to ensure accurate clustering results. Categorical variables such as gender, ethnicity, and parental education were appropriately categorized, while numerical variables like age, study time, and GPA were scaled.

K-Means:



GradeClass Key:

**0: 'A' (GPA >= 3.5)**

**1: 'B' (3.0 <= GPA < 3.5)**

**2: 'C' (2.5 <= GPA < 3.0)**

**3: 'D' (2.0 <= GPA < 2.5)**

**4: 'F' (GPA < 2.0)**

The clustering results from the K-means analysis, with k=5 clusters based on the sampled dataset, showed that clusters 1 and 4 are predominantly associated with students who failed at least one class. In contrast, students who achieved grades A to D are distributed across clusters 2 through 5. Notably, none of the students who earned an A were present in cluster 1, suggesting that this cluster is linked to poorer academic outcomes. The analysis of cluster centers showed that absences have the highest average values in clusters 1 through 4, suggesting a significant relationship between frequent absences and failing grades. Conversely, students who received an A were predominantly in cluster 5, which showed the highest average for parental support, emphasizing the importance of parental involvement in fostering academic success.

A graph of a cluster center

Description automatically generated with medium confidence

The bar chart shows how different student characteristics, or features, define the center of each cluster. These clusters group students based on similarities in features such as absences, GPA, parental support, and extracurricular activities. The y-axis represents the value of each feature for the cluster centers, while the x-axis lists the features.

One key observation is that absences vary significantly between clusters, with one cluster (Cluster 1) having a much higher number of absences. This suggests that students in this cluster have attendance issues. Other features like parental support, GPA, and study time per week also show some differences between clusters, although these variations are less extreme than absences.

Understanding these clusters and the key features that define them can help educators develop more targeted strategies to support students. For instance, focusing on improving attendance in the cluster with high absences or providing extra academic help to those with lower GPAs.

Hierarchical Agglomerative Clustering (HAC)

A diagram of a city

Description automatically generated

The dendrogram shows how students or their grades are grouped based on similarities. Each branch (leaf) at the bottom represents an individual student, and as you move up, similar students are merged into clusters. The height where two clusters merge indicates how different they are—the higher the merge, the more dissimilar the clusters. In this case, there appear to be four main clusters.

These clusters highlight natural groupings of students based on their grades, showing patterns of academic performance. Larger clusters suggest that many students share similar grade profiles, while smaller ones might show unique or outlier profiles. Understanding these groupings can help educators tailor support or interventions, like providing extra help to struggling students or offering advanced challenges to high achievers. This visual tool makes it easier to identify patterns and guide educational strategies.

The findings of this analysis highlight the critical role of absenteeism and parental support in determining student grades. The K-means clustering method successfully identified distinct groups of students with varying academic achievements, providing actionable insights for educators and policymakers. In contrast, HAC proved less effective in distinguishing patterns related to academic outcomes, indicating that K-means clustering is better suited for this type of analysis when the goal is to predict student success based on multiple attributes.

*Subsection 4: Decision Tree Model*

Decision trees are a popular and intuitive machine learning method used for classification and regression tasks. In the context of this study, decision trees were employed to classify students into different grade categories based on various attributes such as their study habits, parental involvement, and extracurricular activities.

A decision tree works by recursively splitting the dataset into subsets based on the values of input features. Each node in the tree represents a feature, and each branch represents a decision rule based on that feature. The process continues until the data is sufficiently partitioned into homogenous subsets, which correspond to the final decision classes (in this case, grade categories).

One of the key strengths of decision trees is their ability to handle both numerical and categorical data, making them versatile for a wide range of datasets. They are also easy to interpret, as the decision rules can be visualized in a tree structure, providing clear insight into the factors that influence the classification decisions.

However, decision trees are prone to overfitting, especially when there is in increase in complexity and capture noise in the data. To address this, pruning techniques can be applied to remove less important branches, simplifying the model and improving its generalization to new data.

In this project, decision tree models were developed both with and without GPA as a feature to predict student grades. When GPA was excluded, the decision tree achieved an initial accuracy of 67.13% with a Kappa statistic of 0.5023. In this model, absences emerged as the most critical predictor of student performance, followed by study time and parental support.

A diagram of a tree

Description automatically generated

Despite efforts to tune the model, the tuned version showed only a slight decrease in accuracy to 66.02%, with a marginally lower Kappa value of 0.4817. Tuning did not significantly improve the performance of the decision tree without GPA, although the tuned model did exhibit better balanced accuracy when predicting certain grade categories, such as high performers (Class F).

A diagram of a tree

Description automatically generated

A graph with a red and blue bar

Description automatically generated

In contrast, the decision tree that included GPA as a feature outperformed the model without GPA. The initial decision tree with GPA demonstrated an accuracy of 74.51% and a Kappa value of 0.6252. In this case, GPA was the most influential predictor, followed by absences and tutoring. The inclusion of GPA as a predictor allowed the model to achieve higher accuracy across different grade categories

A diagram of a tree

Description automatically generated

After tuning, the model with GPA maintained its accuracy and Kappa value, suggesting that the initial model was already well-optimized. The stability in performance after tuning highlights the significant role GPA plays in predicting student grades, overshadowing other factors such as absences and parental support.

A diagram of a tree

Description automatically generated

A graph with a bar graph

Description automatically generated

A diagram of a tree

Description automatically generated

A diagram of a tree

Description automatically generated

A graph with red and blue bars

Description automatically generated

In summary, the decision tree model that included GPA as a feature performed significantly better than the model without GPA in terms of both accuracy and Kappa statistics. Including GPA enhanced the model’s predictive ability, particularly for high-performing students, and allowed for more stable and accurate classification across grade categories. In contrast, the decision tree without GPA, although effective, showed lower accuracy and struggled to classify certain grade categories accurately, particularly those with high or very low grades. This comparison underscores the importance of including key academic indicators, such as GPA, when modeling student performance.

*Subsection 5: Regression Model*

Initial model

A screenshot of a computer

Description automatically generated

The initial regression model, which was developed using only a few variables that were considered as important factors impacting GPA, produced an R squared value of 0.8804, indicating that approximately 88% of the variance in GPA is explained by the predictors included in the model. This high R squared value suggests that the model fits the data well, capturing a substantial portion of the variability in GPA. However, the model's Root Mean Squared Error (RMSE) was 0.3491, which implies that, on average, the model's predictions deviate from the actual GPA values by about 0.35 points. Given that GPA is typically measured on a 4-point scale, an average prediction error of ±0.35 could be considered moderate, indicating room for improvement in prediction accuracy. To enhance the model's performance, additional variables or alternative modeling approaches should be explored.

In an effort to improve the prediction results, a new linear regression model was developed using all the available variables. The updated model includes predictors such as Gender, Ethnicity, ParentalEducation, Tutoring, ParentalSupport, Extracurricular Activities, Sports, Music, Volunteering, Absences, StudyTimeWeekly, and Age.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with blue dots

Description automatically generated

* Intercept: The intercept of -0.6221 represents the predicted GPA when all predictors are at their reference levels.
* Gender: The coefficient for GenderFemale is 0.01685, but it is not statistically significant, indicating that gender might not have a strong effect on GPA in this model.
* Ethnicity: The coefficients for ethnicity categories (African American, Asian, Other) are small and not statistically significant, suggesting that ethnicity does not have a strong impact on GPA after controlling for other variables.
* Parental Education: None of the parental education levels are statistically significant, indicating that, in this model, parental education might not be a strong predictor of GPA.
* Tutoring: The coefficient for TutoringYes is 0.2685, which is highly significant. This suggests that students who receive tutoring are predicted to have a GPA that is, on average, 0.2685 points higher than those who do not, holding other factors constant.
* Parental Support: The coefficients for all levels of parental support are significant and positive, with higher levels of support associated with higher GPAs. For example, students with "Very High" parental support are predicted to have a GPA that is 0.6737 points higher than those with "None" parental support.
* Extracurricular Activities: Participation in extracurricular activities is associated with a 0.2043 increase in GPA, and this is statistically significant.
* Sports: Participation in sports is associated with a 0.2078 increase in GPA, and this is also statistically significant.
* Music: Participation in music is associated with a 0.1496 increase in GPA, which is statistically significant.
* Volunteering: The coefficient for volunteering is very small and not significant.
* Absences: The coefficient for absences is -0.9198, which is highly significant. This means that each additional absence is associated with a decrease in GPA by approximately 0.92 points, highlighting the strong negative impact of absences on academic performance.
* Study Time Weekly: The coefficient is 0.1781, and it is statistically significant. This indicates that each additional unit of weekly study time is associated with a 0.178 increase in GPA, holding other factors constant.
* Age: The coefficient for age is -0.0108, which is statistically significant. This negative coefficient suggests that older students might have slightly lower GPAs, though the effect is small.

This regression model appears to be very strong in predicting students' GPAs based on the given variables. Key predictors such as tutoring, parental support, extracurricular activities, sports, music, absences, and study time are statistically significant and contribute meaningfully to predicting GPA. The Multiple R-squared value of 0.9542 indicates that approximately 95.42% of the variance in GPA is explained by the model, and the low RMSE of 0.2178 suggests a high level of accuracy in the model's predictions.

Final model:

A screenshot of a computer

Description automatically generated

A graph showing a graph of a graph

Description automatically generated with medium confidence

This improved regression model demonstrates strong predictive power, with key variables such as tutoring, parental support, extracurricular activities, sports, music, absences, and study time significantly contributing to GPA predictions. The high Rsquared value and low RMSE reflect the model's effectiveness in accurately predicting GPA based on these factors. The significant coefficients reinforce the importance of these predictors, while non-significant predictors have been removed, improving model simplicity and interpretability.

**Comparison and Interpretation**

Both regression models provide a robust prediction of GPA, with nearly identical performance metrics. The improved model, which excludes non-significant variables, simplifies the analysis without sacrificing accuracy. This refinement makes the model more interpretable, highlighting the most impactful factors such as study time, absences, parental support, and participation in tutoring, sports, and music.

The scatter plot of actual vs. predicted GPA for the improved model shows a strong linear relationship, reinforcing the model's accuracy in predicting GPA. The data points are closely aligned with the diagonal line, indicating that the model’s predictions are highly consistent with the actual GPA values.

The regression analysis demonstrates that GPA is strongly influenced by study habits, attendance, parental support, and engagement in extracurricular activities. The exclusion of non-significant variables in the refined model results in a more streamlined and interpretable analysis without compromising predictive power. This refined model can be a valuable tool for educators and policymakers aiming to understand and improve academic performance based on these key factors.

Confusion Matrix Analysis (Non-tuned Models):

The confusion matrices for each model provide a detailed breakdown of their performance across grade classifications (A, B, C, D, and F).

**A screenshot of a graph

Description automatically generated**

The SVM model demonstrates strong performance, particularly in predicting Class F, with 343 correct predictions. It also performs well for Class D and Class C. However, the model struggles with Class A and Class B, often misclassifying them as Class C or Class D. The majority of correct classifications are concentrated in the higher classes (D and F), suggesting that the model is more effective at distinguishing between these categories.

**A graph with different colored rectangular bars

Description automatically generated with medium confidence**

Not all SVM models are the same, so it is important to examine which kernels perform the best on this dataset. The barplot compares the accuracy of SVM models using different kernels: Sigmoid, RBF (Radial Basis Function), Polynomial, Linear, and the best-tuned SVM model. The Sigmoid kernel, with an accuracy of 64.2%, performs the worst, suggesting it struggles to handle the data's complexity.

The RBF kernel performs better, achieving 77.4% accuracy, showing it handles non-linear relationships well. However, it’s not the top performer. The Polynomial kernel, with 74.9% accuracy, is slightly less effective, possibly due to its sensitivity to hyperparameters like the degree of the polynomial.

The Linear kernel and the best-tuned SVM model both achieve the highest accuracy of 78%. This indicates that the data likely benefits from a linear decision boundary. The best-tuned model likely found an optimal setup through careful adjustments, matching the Linear kernel's performance

**A screenshot of a graph

Description automatically generated**

The Random Forest model also excels in predicting Class F, with 334 correct predictions, and shows strength in classifying Class C and Class D. However, like the SVM, it struggles with Class A, with most instances being incorrectly classified as Class B. While its performance is slightly better than the SVM for Class B, it still misclassifies a significant number of instances in this category.

**A grid of squares with numbers

Description automatically generated**

The Naive Bayes model performs well in predicting Class F with 341 correct classifications and shows solid performance in predicting Class D and Class C. However, the model encounters challenges with Class A and Class B, frequently misclassifying them as Class C or Class D. This pattern indicates that the model may have difficulties with class balance or distribution.

**A screenshot of a graph

Description automatically generated**

The kNN model is particularly effective in predicting Class F, with 340 correct predictions, and performs reasonably well with Class D. However, it struggles with the lower classes, especially Class A, where predictions are often scattered across Class B and Class C. This tendency to confuse neighboring classes is typical of kNN models, especially when the class boundaries are not well-defined in the feature space.

**A white grid with purple squares and numbers

Description automatically generated**

The Decision Tree model shows strength in predicting Class F, with 335 correct classifications, and performs adequately with Class C and Class D. However, it faces significant challenges with Class A, where most instances are misclassified as Class B. The performance for Class B is also inconsistent, with frequent misclassifications into Class C.

Across all models, Class F is consistently well-predicted, while Classes A and B pose challenges, often being misclassified into neighboring classes.

Tuning Process:

The models were tuned using grid search to optimize their hyperparameters:

* SVM Tuning: SVM was tuned for the C (cost) and sigma (kernel width) parameters. The best results were obtained with C = 10 and sigma = 0.1, leading to an accuracy of 77.99%. The tuned model achieved a good balance of regularization and performance, but the non-tuned version performed slightly better, suggesting the initial parameters were already effective.
* Random Forest Tuning: Random Forest was tuned by adjusting the mtry parameter (number of variables considered at each split). The optimal value was mtry = 2, with an accuracy of 78.82%. However, the non-tuned model slightly outperformed the tuned version, indicating the original settings were effective.
* Naive Bayes Tuning: Naive Bayes was tuned using fL = 0 and usekernel = TRUE, leading to an accuracy of 80.31%. The kernel-based Naive Bayes model performed well, capturing non-linear relationships effectively.
* kNN Tuning: The kNN model was tuned by selecting the number of neighbors (k), with k = 7 giving the best accuracy of 73.92%. However, the non-tuned model performed better, suggesting that a different k value or tuning method might improve performance.
* Decision Tree Tuning: The Decision Tree model was tuned using the complexity parameter (cp = 0.01), which improved its accuracy to 80.19%. The tuning reduced overfitting and led to better splits and generalization.

Comparing Tuned vs Non-Tuned Models:

The table below compares the accuracy and kappa of the non-tuned models (before tuning) and the tuned models (after hyperparameter adjustments).

A screenshot of a computer

Description automatically generated

* The non-tuned SVM slightly outperformed the tuned version, indicating that the initial hyperparameters were already near optimal. The slight drop in accuracy post-tuning suggests potential overfitting.
* Random Forest like SVM, the non-tuned Random Forest model showed slightly better results than the tuned model, with marginal differences in accuracy and kappa scores.
* Naive Bayes also performed better in the non-tuned version. This suggests that the default hyperparameters were well-suited for the data.
* The non-tuned k-NN model performed better than the tuned version, indicating that the selected number of neighbors (k = 7) might not have been optimal.
* The tuned decision tree model showed a significant improvement in accuracy and kappa, demonstrating the effectiveness of tuning the complexity parameter.

Multi-Model Comparison (Based on Tuned Models):

In this analysis, the performance of several classification models—Support Vector Machines (SVM), Random Forest, Naive Bayes, k-Nearest Neighbors (kNN), and Decision Trees—was evaluated after tuning. The results include insights from ROC curves, cross-validation accuracies, confusion matrices, and a table comparing non-tuned and tuned model performances.

ROC Curves Analysis (Tuned Models):

A graph of different models

Description automatically generated

The ROC (Receiver Operating Characteristic) curves illustrate the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate) for each model.

* SVM (Blue Curve): The SVM model shows a smooth and steep ROC curve, indicating a strong balance between sensitivity and specificity. This makes it one of the top performers in distinguishing between classes.
* Random Forest (Green Curve): Similar to SVM, the Random Forest model performs very well, with its ROC curve closely approaching the top-left corner, indicating high sensitivity and specificity.
* Naive Bayes (Red Curve): Naive Bayes performs slightly worse than SVM and Random Forest, with a less steep ROC curve, but still does better than kNN and Decision Tree models.
* kNN (Purple Curve) & Decision Tree (Orange Curve): These models have ROC curves closer to the diagonal line, indicating that they are less effective at distinguishing between classes.

In summary, the ROC curves suggest that the SVM and Random Forest models are the most effective at distinguishing between classes, followed by Naive Bayes, while kNN and Decision Tree lag behind.

Cross-Validation Accuracies (Tuned Models):

The boxplot of cross-validation accuracies provides insights into the consistency and average performance of the models.

**A graph showing different colored bars

Description automatically generated with medium confidence**

* Decision Tree: The Decision Tree model has the lowest median accuracy with a wider range, indicating variability in performance. While capable of achieving high accuracy in some cases, it is less reliable overall.
* kNN: The kNN model shows a narrow range of accuracies but with a lower median than the other models, indicating consistent but lower performance.
* Naive Bayes: Naive Bayes has a higher median accuracy compared to kNN and Decision Tree, with moderate variability in accuracy across folds.
* Random Forest: The Random Forest model shows one of the highest median accuracies with a small interquartile range, indicating consistent and strong performance.
* SVM: Like Random Forest, the SVM model also exhibits high accuracy and low variability across cross-validation folds, highlighting its robustness and reliability.

These cross-validation results further confirm that SVM and Random Forest are the best-performing models, while Naive Bayes offers decent performance, and kNN and Decision Tree are less reliable.

General Observations Across Models:

**A graph showing different colored rectangular shapes

Description automatically generated**

Top of Form

In this project, several methods were used to analyze student performance, each providing distinct insights. Association Rule Mining (ARM) helped uncover relationships between student characteristics and academic outcomes. For instance, ARM revealed that students who did not participate in volunteering or music activities were more likely to have lower grades, whereas those with low absenteeism and moderate parental support were more likely to achieve higher academic performance. ARM is effective for finding patterns in data but does not predict outcomes directly.

Clustering techniques, including K-Means and Hierarchical Agglomerative Clustering (HAC), were used to group students based on similarities in their attributes. K-Means successfully identified meaningful clusters, such as students with high absenteeism and lower grades, and those with strong parental support and better academic outcomes. However, HAC was less effective in detecting clear patterns, making K-Means more useful for segmenting students for targeted interventions. Clustering, like ARM, is beneficial for grouping and analysis but does not directly predict student performance.

Decision Trees were employed to classify students into grade categories based on their attributes, such as absenteeism and parental support. This method highlighted key predictors of academic success but faced challenges in distinguishing between higher grades. While decision trees are easy to interpret, they have a tendency to overfit the data, reducing their reliability for classification tasks in some cases.

The most accurate models were Support Vector Machines (SVM- specifically the linear and RBF kernels) and Random Forests. SVM excelled due to its ability to handle complex relationships in the data, while Random Forests improved accuracy by combining multiple decision trees and identifying important factors for prediction. These models offer a significant improvement in accuracy but operate at a higher complexity level, making them harder to interpret compared to simpler models like decision trees.

Overall, SVM and Random Forest were the most accurate models for predicting student performance. Naive Bayes is a solid alternative due to its simplicity and efficiency, while kNN and Decision Tree require further optimization to achieve better results. Combining the predictive power of these models with context-driven insights from ARM and Decision Trees offers a comprehensive understanding of the factors influencing student success. The analysis underscores the importance of considering both individual factors and their interactions to develop effective academic intervention strategies.

Bottom of Form

**Conclusion**

This study explored the key factors influencing high school student performance and found that attendance and parental support play a vital role in academic success. Students who missed more school consistently had lower grades, emphasizing the importance of regular attendance. Additionally, students with strong parental support were more likely to perform well academically.

The findings suggest that improving attendance and increasing parental involvement can significantly enhance student outcomes. Educators and policymakers can use this information to develop strategies that make it easier for parents, especially working ones, to stay involved, and implement programs to boost student attendance.

The analysis also revealed that non-academic activities, like volunteering or music, positively impact academic performance. Clustering students based on similar performance levels highlighted patterns, such as the link between certain extracurricular activities and better grades.

While some advanced methods accurately predicted student grades by considering complex factors like study time and attendance, simpler methods helped isolate specific influences. These simpler techniques offered clearer insights into what drives academic success without the confusion caused by overlapping factors.

In summary, this study confirmed the importance of attendance and parental support while uncovering new trends in academic achievement. These results offer valuable guidance to parents, educators, and school boards on how to best support students. By understanding these critical factors, schools can take proactive steps to help students succeed, and parents can focus their efforts on the most effective ways to support their children’s academic progress.

**References**

Schreiner, L. A. (2012). Introduction: Thriving in Transitions. In L. A. Schreiner, M. C. Louis, & D. D. Nelson (Eds.), *Thriving in Transitions: A Research-Based Approach to College Student Success* (pp. 3-11). Columbia, SC: University of South Carolina, National Resource Center for The First-Year Experience and Students in Transition.

Schreiner, L. A. (2012). Chapter 2: The Thriving Quotient: A New Vision for Student Success. In L. A. Schreiner, M. C. Louis, & D. D. Nelson (Eds.), *Thriving in Transitions: A Research-Based Approach to College Student Success* (pp. 35-65). Columbia, SC: University of South Carolina, National Resource Center for The First-Year Experience and Students in Transition.