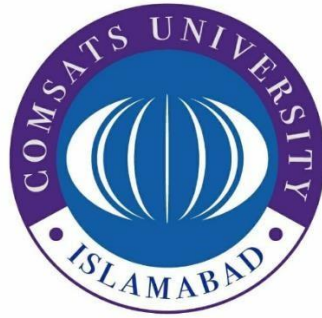


COMSATS UNIVERSITY ISLAMABAD



DATA MINING

BDS-5A

LAB PROJECT

Academic Performance Analysis with Predictive Modelling

Submitted by:

SUMAMA BIN TAHIR (FA22-BDS-046)

AHMED NAWAZ ABBASI (FA22-BDS-005)

Submitted to:

Ma'am Hilal Jan

Date: 29th December 2024

Table of Contents

Problem Description	4
1.1. Project Objective.....	4
1.2. Significance of Problem	5
1.3. Why does this problem require a Data Mining approach?	5
Dataset Selection	6
2.1. Dataset Size and Composition	6
2.2. Key Attributes	7
Preprocessing Steps	8
3.1. Data Cleaning.....	8
3.2. Data Transformation	8
3.3. Feature Selection and Engineering	9
Why was Dimensionality Reduction not necessary?	10
Data Mining Technique Selection	11
4.1. Justification for Using Classification.....	11
4.2. Model Selection	11
Decision Tree Classifier	12
Random Forest Classifier	14
Why both Decision Tree and Random Forest?.....	14
Support-Vector Machine.....	16
Reason we chose SVM for our dataset:	16
K-Nearest Neighbor (KNN):.....	18
Why KNN ?	18
Conclusion	21
Naïve Bayes:.....	22
Results and Insights	25
Conclusion	25
Gradient Boosting:	26
Results and Insights	29
Conclusion	31
Artificial Neural Network (ANN)	32
Results and Insights	32

Model Accuracy	33
Results.....	34
What are the most important factors that determine the academic performance of students?	34
How does a guardian's role (whether mother or father) impact a child's academic performance?.....	35
Is there a specific attendance threshold where a student can be classified as high-performing or low-performing?	37
How does student attendance correlate with their likelihood of being categorized as high-performing or low-performing?.....	38
How do student attendance and participation correlate with their likelihood of being categorized as high-performing or low-performing?	39
What impact does parental influence ("ParentAnsweringSurvey" or "ParentschoolSatisfaction") has on student performance?	41

Problem Description

In modern educational systems, there is a growing need to monitor and improve student performance proactively. Predicting academic performance early enables educators, administrators, and parents to intervene and provide support to students who may face academic challenges.

Academic performance is influenced by a wide array of factors, ranging from personal demographics to behavioral metrics, such as participation in class discussions, attendance, and resource utilization.

The goal of this project is to find out in what ways these factors influence academic performance of students.

1.1. Project Objective

This project aims to predict students' academic performance based on features such as study behaviors, parental involvement, and demographics.. By analyzing these variables, we can develop models that categorize students into different performance levels such as Low, Medium or High. These results will allow educators to take a data-driven approach to identify students who may need additional support, reinforcement or guidance in their studies.

Hence we have six research questions/problem statements:

- 1. What are the most important factors that determine the academic performance of students?**
- 2. How does a guardian's role (whether mother or father) impact a child's academic performance?**
- 3. Is there a specific attendance threshold where a student can be classified as high-performing or low-performing?**
- 4. How does student attendance correlate with their likelihood of being categorized as high-performing or low-performing?**
- 5. How do student attendance and participation correlate with their likelihood of being categorized as high-performing or low-performing?**
- 6. What impact does parental influence has on student performance?**

1.2. Significance of Problem

Student performance prediction is critical for several reasons:

1. Early Intervention

Identifying students with potential academic struggles early enables timely intervention, which can help prevent these students from falling further behind.

2. Resource Allocation

Schools often operate with limited resources, such as tutors or supplemental programs. By pinpointing at-risk students, resources can be strategically allocated where they are most needed, optimizing educational outcomes within existing constraints.

3. Parental Involvement

Parental engagement is widely recognized as an essential factor in student success. By including parental satisfaction and involvement as features in this analysis, schools can better understand how family support correlates with academic performance, helping to create family-inclusive support programs.

4. Long-term Academic Planning

Performance prediction can assist in long-term academic planning, especially for curriculum adjustments, course counseling and advice.

By using historical data to forecast academic outcomes, schools can implement support systems that enhance student readiness for future courses or academic challenges.

1.3. Why does this problem require a Data Mining approach?

The diverse and multi-dimensional nature of educational data necessitates a data mining approach to reveal meaningful patterns that may not be apparent through manual analysis.

Complexity of Educational Data – Educational data is diverse and multi-dimensional, making it difficult to analyze effectively through manual methods.

Benefits of Data Mining Techniques – Data mining techniques like classification can uncover hidden relationships, generate predictive insights, and support data-driven

decision-making, which manual analysis may overlook. Data mining techniques allow us to:

1. [Uncover Hidden Relationships](#)
2. [Generate Predictive Insights](#)
3. [Support Data-Driven Decision Making](#)

Predicting academic performance provides a pathway to enhance student success and improve educational outcomes. By leveraging this approach educational institutions can apply data driven methodologies to support students in a personalized, proactive and effective manner.

Dataset Selection

In this project, we use the **xAPI-Edu-Data** dataset a publicly available educational dataset designed to support educational research and analytics specifically in predicting student performance based on behavioral and demographic features.

- **Dataset Name:** xAPI-Edu-Data
- **Source:** This dataset was acquired from an open-access educational repository “Kaggle” with contributions from multiple institutions. The data has been structured according to the xAPI standard to capture students' interactions and outcomes.
- **Purpose of the Dataset:** It is tailored for predictive analysis of student performance. It includes various attributes that track students' engagement in educational activities, interactions with resources, and parental feedback, making it ideal for training classification models.

2.1. Dataset Size and Composition

- **Number of Rows:** 480
- **Number of Features:** 17
- **Attributes:** Each row represents an individual student's record, capturing a wide array of features.

- **Target Variable:** The “**Class**” attribute, which categorizes students into three academic performance levels: Low, Medium, and High. This attribute serves as the output variable for classification.

2.2. Key Attributes

The dataset contains several key attributes, grouped into four main categories:

1. Demographic Information:

- **Gender:** Represents the gender of the student (e.g., Male, Female).
- **Nationality:** Specifies the nationality of the student with values representing different countries or regions.
- **Place of Birth:** Indicates the student's place of birth, potentially useful in understanding geographic influences on academic performance.

2. Academic and Behavioral Metrics:

- **StageID:** Represents the student's education level such as elementary or secondary school.
- **GradeID:** Specifies the student's grade capturing their current academic standing.
- **raisedhands:** Tracks the number of times a student raised their hand in class indicating active participation and engagement.
- **Visited Resources:** Counts the frequency of online resources accessed by the students reflecting their independent learning efforts.
- **Announcements View:** Measures the number of announcements the student has viewed providing insight into their attentiveness to school communications.
- **Discussion:** Represents the student's engagement in classroom discussions.

3. Parental Involvement: • **Relation:** Identifies whether the student's guardian is the mother or father. This attribute is crucial for analyzing family influences on academic performance.

- **ParentAnsweringSurvey:** Indicates whether the parent or guardian completed a school survey representing their level of involvement in school activities.
- **ParentschoolSatisfaction:** Captures parental satisfaction with the school an attribute that could influence or reflect the student's academic experiences.

4. Attendance and Absence:

- **StudentAbsenceDays:** Categorical data that specifies whether a student has been absent for more than seven days ("Above-7") or fewer ("Under-7"). Attendance is a crucial factor influencing academic performance.

Preprocessing Steps

The dataset required several preprocessing steps to ensure data quality and compatibility with machine learning algorithms. The following steps outline the data cleaning, transformation and preparation processes applied:

3.1. Data Cleaning

Although the data didn't require extensive cleaning, we still needed to prepare it for further processes.

- **Handling Missing Values:** The dataset was examined for missing values across all columns. Since any missing values could impact model accuracy, rows containing null values were removed using `data.dropna()`.
- **Addressing Inconsistencies:** The dataset was inspected for potential inconsistencies in categorical values, such as misspelled labels or duplicated entries, to maintain data integrity. No notable inconsistencies were found.

3.2. Data Transformation

To prepare the data for machine learning, several transformations were applied to ensure compatibility and optimize model performance.

- **Label Encoding of Categorical Features:** The dataset includes several categorical features, such as gender, Nationality, Place of Birth, StageID, GradeID, Semester,

Relation, ParentAnsweringSurvey, ParentschoolSatisfaction, and StudentAbsenceDays. These categorical values were converted into numerical forms using LabelEncoder from ScikitLearn, making them compatible with classification models that expect numerical inputs.

Example: The Relation feature, which specifies the student's guardian (Mother or Father), was encoded into numerical values (e.g., 0 for Mother and 1 for Father).

This transformation allowed the categorical features to be processed efficiently by machine learning algorithms without changing the underlying meaning of the categories.

- **Standardization of Numeric Features:** To ensure that all features contribute proportionately to the model, numeric features were standardized using StandardScaler from Scikit-Learn. This step rescales features to a mean of 0 and standard deviation of 1, which improves model performance, particularly for distance-based algorithms.
- **Affected Features:** The standardized features include raisedhands, Visited Resources, Announcements View, and Discussion, all of which are engagement metrics that can vary significantly in scale.
- **Impact of Standardization:** Standardizing these features prevents large-scale features from disproportionately influencing the model, ensuring that each feature contributes meaningfully to the predictions.

3.3. Feature Selection and Engineering

Feature selection and engineering techniques were explored to reduce noise and enhance the model's predictive power. Though we ended up not applying any of those techniques.

For this dataset, dimensionality reduction was assessed but ultimately not applied due to the manageable number of features and the importance of retaining all features for interpretability.

- **Dimensionality Reduction:** Although dimensionality reduction techniques, such as Principal Component Analysis (PCA) were considered at first, they were not applied since the dataset contained only 17 features. Reducing dimensions was deemed

unnecessary as it could lead to a loss of interpretability, which is valuable for understanding which specific features influence academic performance.

- **Feature Engineering:** No additional features were engineered as the existing features were sufficient for capturing the students' academic and behavioral metrics.

Why was Dimensionality Reduction not necessary?

1. Number of features being manageable.
2. For our goal, interpretability is very important, since techniques like PCA produce fewer features (principal components) as linear combination of original features, it was adamant that we would lose interpretability, which is valuable for understanding features that influence academic performance.

This approach was validated when initial attempts at dimensionality reduction resulted in skewed results, further confirming that retaining all features was the right choice for this project.

Data Mining Technique Selection

To address the problem of predicting academic performance, classification techniques were chosen. The goal of classification is to predict discrete labels (in this case, performance categories: Low, Medium, or High) based on various input features.

4.1. Justification for Using Classification

Classification is a supervised learning technique suited for predicting categorical outcomes. In this project, the target variable (Class) represents distinct academic performance levels (Low, Medium, High), making classification the ideal choice. Using classification techniques allows us to:

- **Predict Performance Levels:** By training on historical data, the models can learn the relationships between student attributes (such as engagement and demographic information) and performance levels.
- **Identify Key Influencers:** Classification models provide insights into which features most significantly impact predictions helping educators understand the factors influencing student outcomes.
- **Support Decision-Making:** Predictive insights allow educators to proactively address students' needs, aligning with the project's goal of improving academic outcomes.

4.2. Model Selection

Seven classification algorithms were selected: Each has unique characteristics and strengths, making them well-suited for our dataset.

Decision Tree Classifier

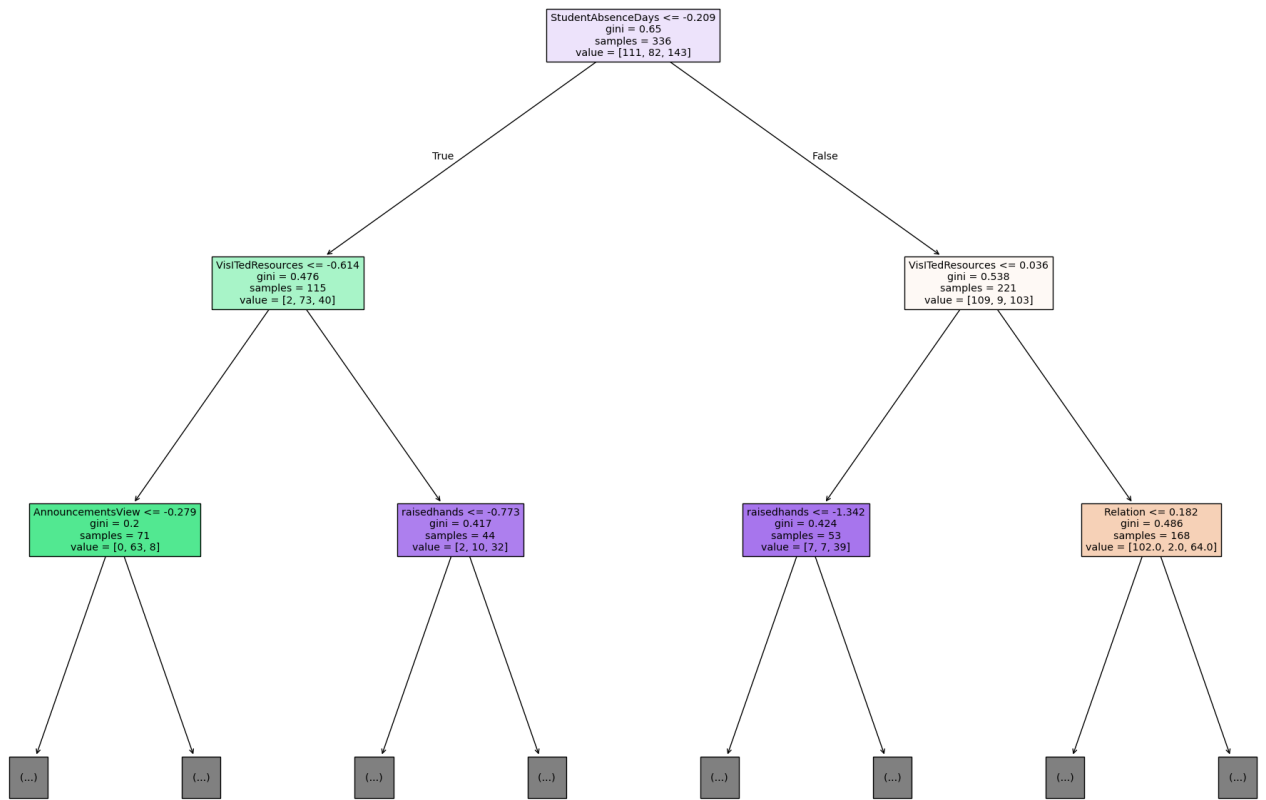
A Decision Tree is a flowchart-like model that makes decisions by splitting the dataset into smaller subsets based on feature values. Here's why Decision Tree Classifier was chosen:

- **Interpretability:** Decision Trees are highly interpretable, allowing us to visualize the decision-making process at each node. This transparency enables educators to understand which features and thresholds contribute to performance classifications, providing actionable insights.
- **Feature Importance:** Decision Trees provide feature importance scores, highlighting which attributes contribute most to predictions. This information can help educators focus on specific factors that influence academic outcomes.

Advantages

The Decision Tree model is particularly useful in this educational context because it clearly shows which variables (e.g., resource access or class participation) lead to different performance outcomes. This model would help us identify intervention points for students with low engagement or high absenteeism.

Simplified Decision Tree for Academic Performance Prediction



Random Forest Classifier

Random Forest is an ensemble technique that combines multiple Decision Trees to improve predictive accuracy and robustness. It generates a “forest” of trees, each trained on random samples of the dataset, and aggregates their predictions. Here’s why Random Forest Classifier was selected:

- **Improved Accuracy:** Random Forests reduce the risk of overfitting that can occur in a single Decision Tree by averaging the predictions of multiple trees. This enhances model generalization, resulting in higher accuracy.
- **Robustness to Noise:** By averaging results across many Decision Trees, Random Forests are more resistant to noise in the data. This robustness is crucial for educational datasets, which may have some variance due to differences in student behavior or recording errors.

Advantages

The higher accuracy of Random Forest allows for more reliable predictions, making it a strong choice for identifying at-risk students.

Why both Decision Tree and Random Forest?

Random Forest, being an ensemble of multiple decision trees, generally offers better accuracy than a single decision tree. This was the case in our scenario, as the Decision Tree model showed lower accuracy.

Therefore, we applied the Random Forest classification technique, which resulted in improved overall accuracy.

Since Random Forest provided higher accuracy, we’ll use it to address our research question later:

What are the most important factors that determine the academic performance of students?

Decision Tree Classification Report:

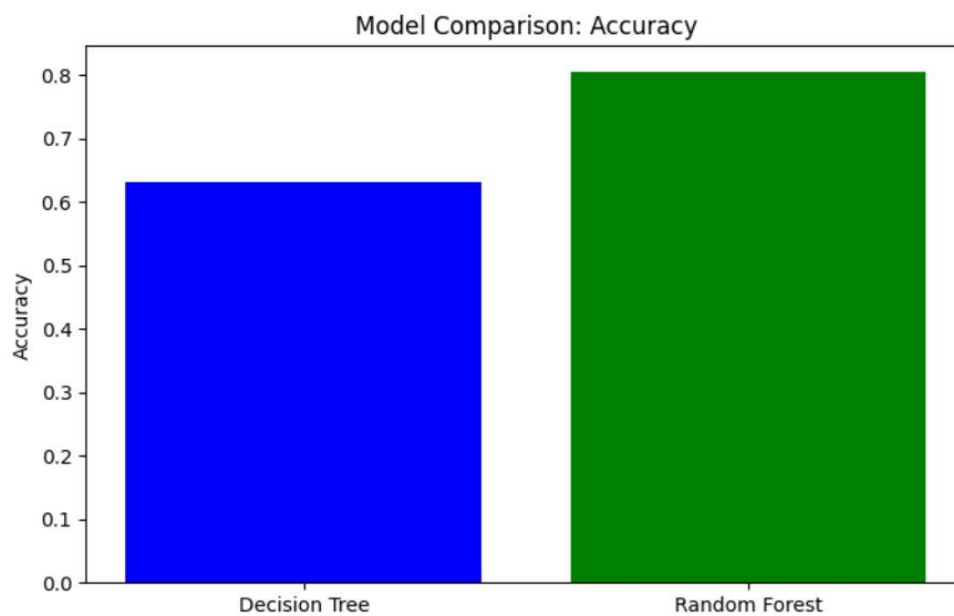
	precision	recall	f1-score	support
0	0.62	0.58	0.60	31
1	0.68	0.67	0.67	45
2	0.61	0.63	0.62	68
accuracy			0.63	144
macro avg	0.64	0.63	0.63	144
weighted avg	0.63	0.63	0.63	144

Decision Tree Accuracy: 0.6319444444444444

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.78	0.68	0.72	31
1	0.87	0.87	0.87	45
2	0.78	0.82	0.80	68
accuracy			0.81	144
macro avg	0.81	0.79	0.80	144
weighted avg	0.81	0.81	0.80	144

Random Forest Accuracy: 0.8055555555555556



Support-Vector Machine

SVM is a supervised machine learning model that can be used for both classification and regression.

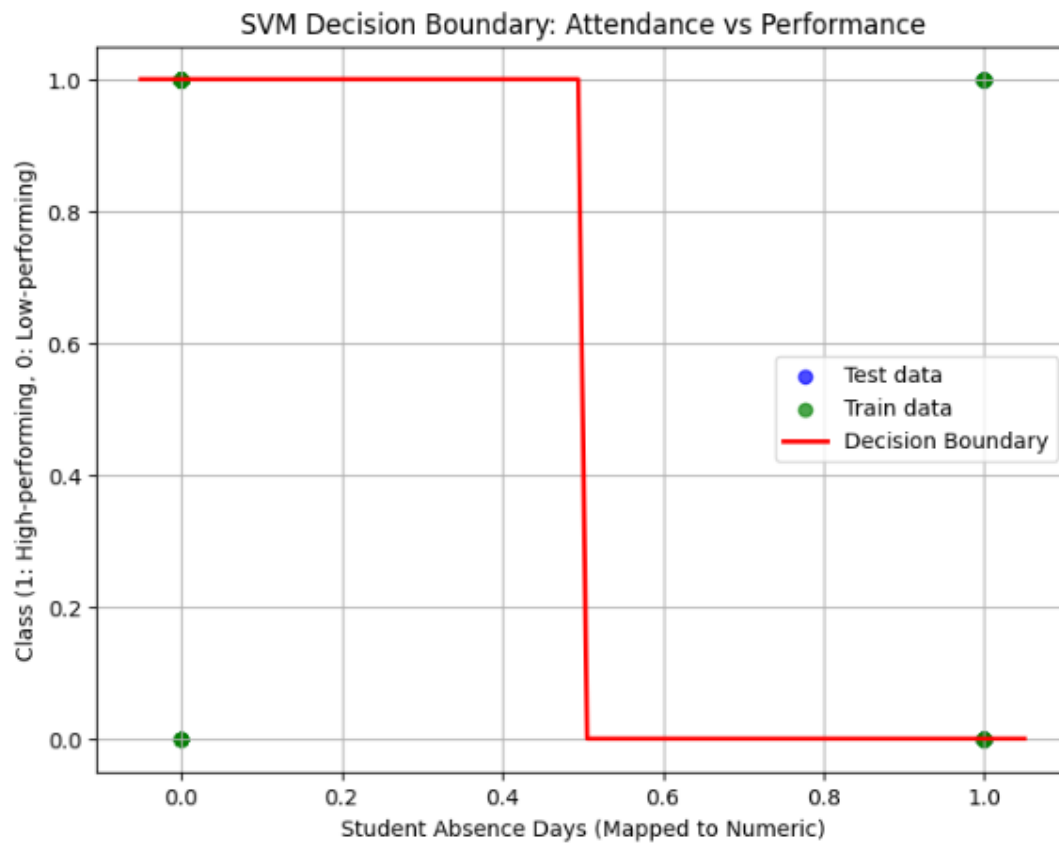
Reason we chose SVM for our dataset:

Even though it's computationally intensive, it is effective on high dimensionality data. Since our dataset has a lot of dimensions(columns), Support Vector Machine is a well-suited model.

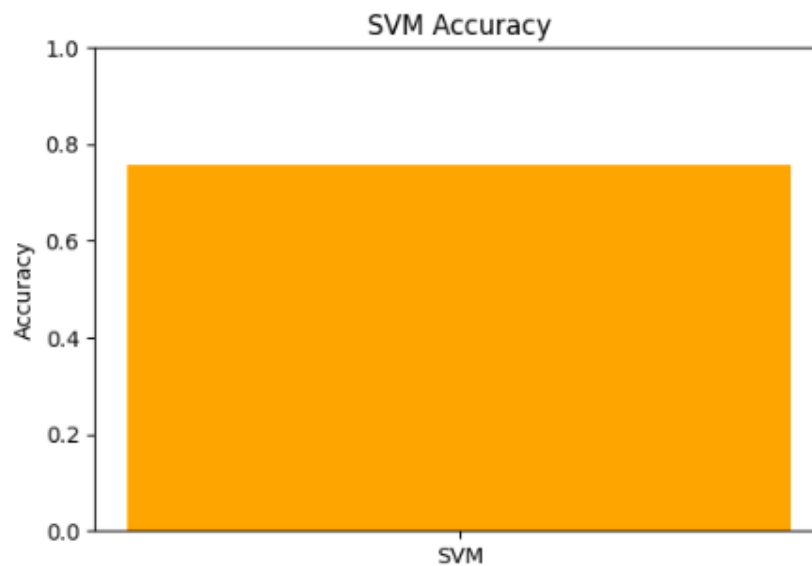
My main motive to apply SVM was to answer the question:

Is there a specific attendance threshold where a student can be classified as high-performing or low-performing?

The model had an accuracy of 76% and showed ample insights.



Decision Boundary (Threshold for Attendance): 0.5
SVM Accuracy: 0.7569444444444444



The answer is clear that:

The **threshold** is:

- **7 absences:** If a student has fewer than 7 absences (mapped as Under-7), they are likely to be classified as high performing.
- Students with **7 or more absences** (mapped as Above-7) are classified as low-performing.

This finding suggests that attendance plays a critical role in student performance, with **7 absences** being the key tipping point identified by the model.

K-Nearest Neighbor (KNN):

Why KNN ?

1. Simplicity:

KNN is easy to implement and interpret, making it a good choice for educational datasets.

2. Handling Non-Linear Relationships:

KNN does not make assumptions about the distribution of data or relationships between features and outcomes. This is useful for a dataset that includes behavioral features like attendance (StudentAbsenceDays) and participation (raisedhands).

KNN in our context is used to address the following question.

How do student attendance and participation correlate with their likelihood of being categorized as high-performing or low-performing?

The KNN model uses **StudentAbsenceDays** (attendance) and **raisedhands** (class participation) as predictors to determine whether a student is **high-performing (Class = 1)** or **low-performing (Class = 0)**.

Steps Taken to Address the Question

1. Data Preparation:

- Features (StudentAbsenceDays and raisedhands) were selected for training the model.
- The target variable (Class) was cleaned using imputation for any missing values.

2. Train-Test Split:

- The data was split into training (75%) and testing (25%) sets to evaluate the model on unseen data.

3. KNN Model Implementation:

- A KNN model was initialized with **k=3** neighbors. This means the class label of a test point is determined based on the majority vote of its 3 nearest neighbors.

4. Model Training:

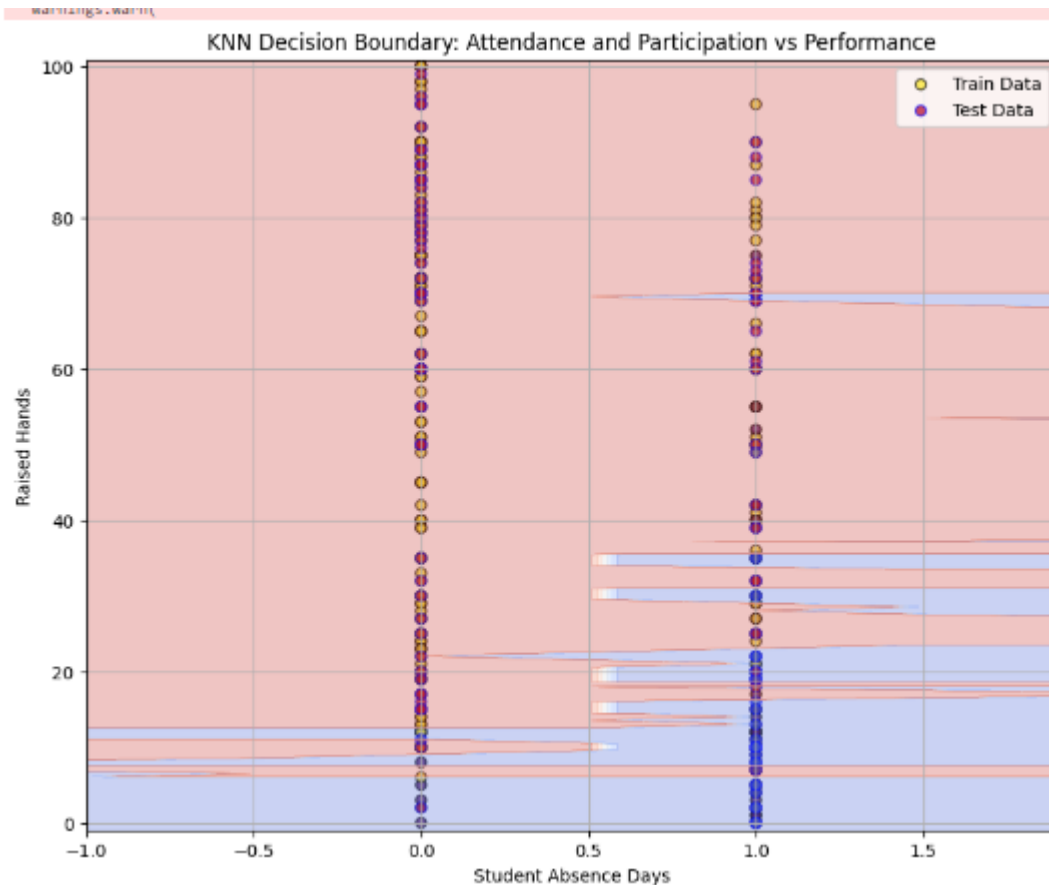
- The model was trained using the training set (X_train, y_train).

5. Predictions and Evaluation:

- The trained model predicted the performance (y_pred) of the students in the testing set (X_test).
- The accuracy of the model was calculated using **accuracy_score**. This metric measures how many predictions the model got correct.

6. Visualization:

- A **decision boundary plot** was created to visualize the classification regions for high and low-performing students, using the two features (StudentAbsenceDays and raisedhands).
- Training and testing data points were plotted on the decision boundary to show how the model classifies them.
- A **bar chart** was used to represent the accuracy of the KNN model, providing a clear view of its performance.



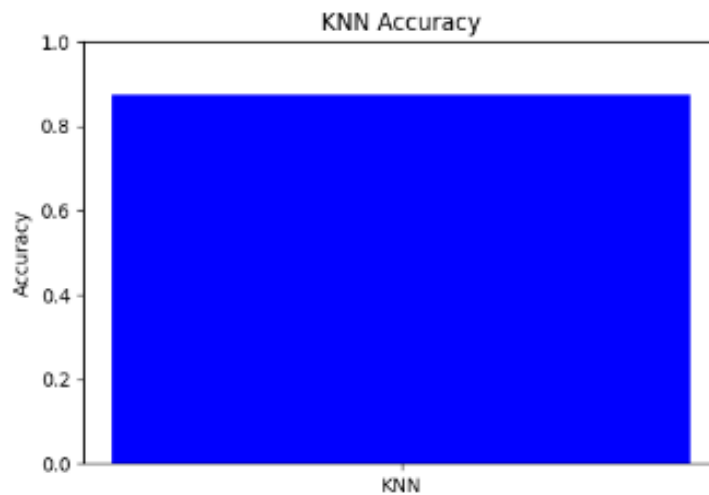
The KNN model demonstrated that:

1. Students with **lower absence days** and **higher participation (raised hands)** are more likely to be categorized as **high-performing**.
2. Students with **higher absence days** and **lower participation** are more likely to be categorized as **low-performing**.

The decision boundary plot visually confirmed these trends:

- High-performing students clustered in regions with low absence and high participation.
- Low-performing students clustered in regions with high absence and low participation.

The model achieved an accuracy of **87.50%** when **k=3** the test dataset, which indicates its effectiveness in classifying student performance based on attendance and participation.



Conclusion

The KNN model effectively addressed the question:

- Students who are more engaged in class (measured by raisedhands) and have fewer absences are more likely to perform well academically.
- Attendance and participation are strong predictors of student performance.

The use of KNN provided a simple yet powerful approach to classify student performance, with clear visualizations supporting the analysis. These insights can help educators focus on improving attendance and class engagement to boost student outcomes.

Naïve Bayes:

We used Naïve Bayes for the same use case as KNN to see if it provided better insights than KNN.

Naive Bayes is used because of its

1. Probabilistic Nature:

- Naive Bayes is a probabilistic classifier that predicts the probability of each class given the features.
- It is particularly effective when features are independent, which aligns well with the assumption of NB.

2. Efficiency:

- NB is computationally efficient and suitable for relatively small datasets.

3. Well-Suited for Categorical Data:

- Features like StudentAbsenceDays and raisedhands can be treated as categorical or discrete, making NB a strong candidate.

The Naive Bayes classifier predicts whether a student is **high-performing (Class = 1)** or **low-performing (Class = 0)** based on:

- **StudentAbsenceDays:** The number of days a student has been absent.
- **raisedhands:** The number of times a student has raised their hand in class, indicating participation.

Naive Bayes Model Training

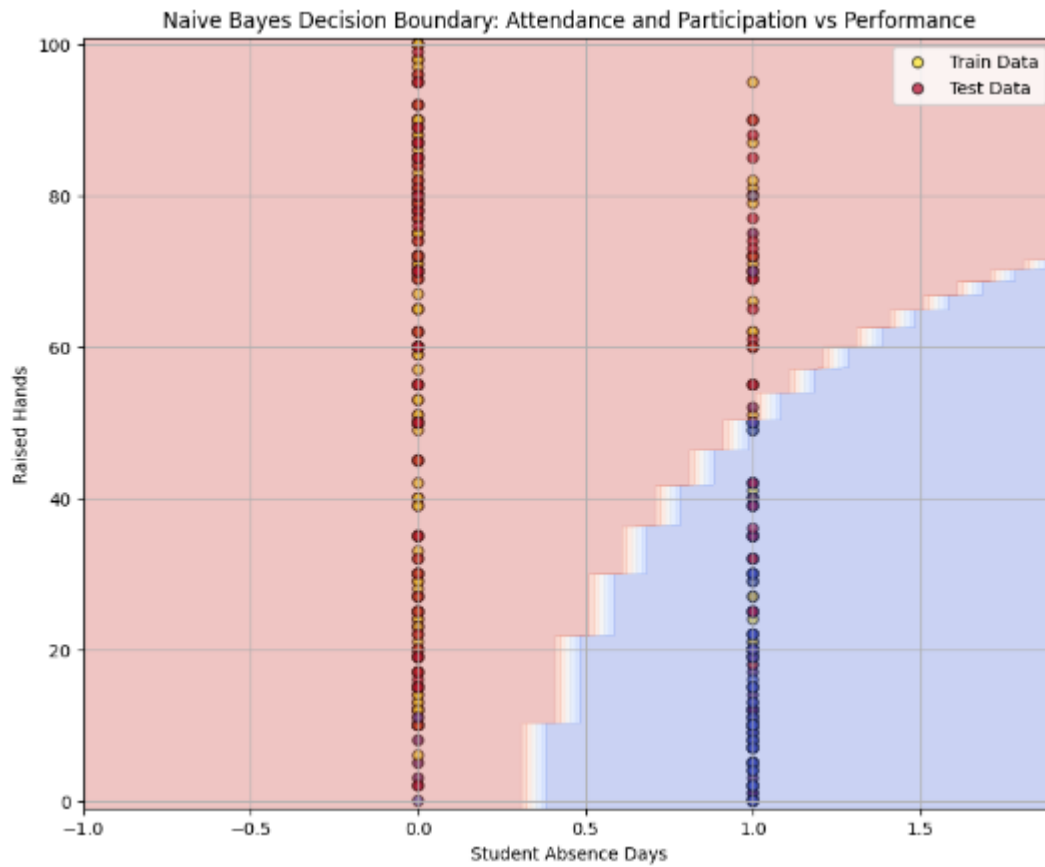
- A **Gaussian Naive Bayes** model was used:
 - Assumes a normal distribution for continuous features like StudentAbsenceDays and raisedhands.
- The model was trained on the **training dataset (X_train, y_train)**.

Predictions and Evaluation

- The model predicted the performance of students in the **test dataset (X_test)**.
- **Evaluation Metrics:**
 - **Accuracy:** The percentage of correctly classified students.
 - **Classification Report:**
 - Precision, Recall, F1-score for each class (high-performing and low-performing).

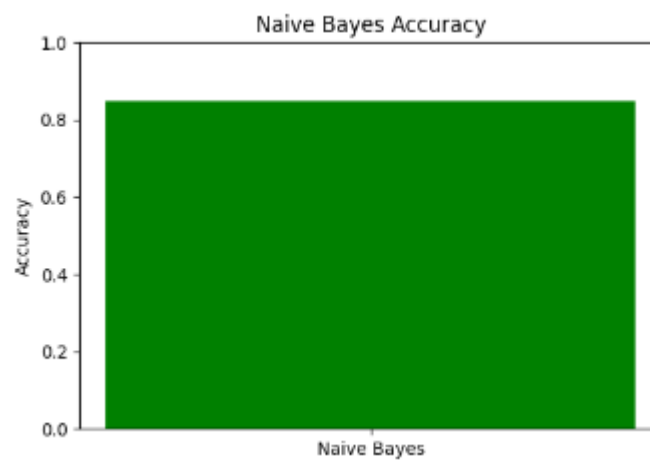
Visualization

- **Decision Boundary:**
 - A mesh grid was created to visualize the decision boundary for the two features (StudentAbsenceDays and raisedhands).
 - Predicted regions for high and low-performing students were plotted.
 - Training and testing data points were overlaid to show their classification.



- **Accuracy Bar Chart:**

- The accuracy of the Naive Bayes model was represented as a bar chart.



Results and Insights

1. Accuracy:

- The Naive Bayes model achieved an accuracy of **84.72%** indicating its effectiveness in predicting student performance.

2. Decision Boundary:

- Students with **fewer absences** and **higher participation** were more likely to be classified as **high-performing**.
- Students with **more absences** and **lower participation** were more likely to be classified as **low-performing**.

3. Classification Report:

- Precision and recall for both classes highlighted how well the model identified high and low-performing students.

Conclusion

Even though Naive Bayes model effectively addressed the research question, it could not prove more accurate than KNN.

Gradient Boosting:

Gradient Boosting is used because of

1. Boosting Strength:

- Gradient Boosting combines multiple weak learners (decision trees) into a strong learner by iteratively reducing the prediction error.

2. Feature Importance:

- Gradient Boosting naturally identifies the importance of features like StudentAbsenceDays, making it useful for understanding the relationship between attendance and performance.

3. Versatility:

- It handles both binary and multi-class classification effectively.

4. Improved Accuracy:

- Gradient Boosting often provides higher accuracy compared to simpler models like Logistic Regression or Naive Bayes, especially for moderately complex datasets.

In our context it is used to answer the following question:

How does student attendance correlate with their likelihood of being categorized as high-performing or low-performing?

The Gradient Boosting model predicts whether a student is **high-performing (Class = 1)** or **low-performing (Class = 0)**

Steps

1. Data Preparation

- **Features:**
 - Focused on StudentAbsenceDays, which indicates the level of attendance.
- **Target Variable:**
 - Class represents performance, where 1 is high-performing and 0 is low-performing.
- **Mapping and Encoding:**
 - StudentAbsenceDays was mapped to numerical values.
 - The target variable Class was encoded for binary classification.
- **Missing Data Handling:**
 - Imputed missing values in the target variable using the most frequent value.

2. Train-Test Split

- The dataset was split into:
 - **Training set (70%).**
 - **Testing set (30%).**

3. Scaling Features

- Since Gradient Boosting benefits from well-scaled features, the **StandardScaler** was applied to ensure uniformity.

4. Model Training

- A **Gradient Boosting Classifier** was built with:
 - **Number of estimators (n_estimators):** 100 (number of decision trees).

- **Learning rate:** 0.1 (controls how much each tree contributes to the final prediction).
- **Maximum depth (max_depth):** 3 (to prevent overfitting while capturing complexity).
- The model was trained on the **scaled training data**.

5. Predictions and Evaluation

- **Evaluation Metrics:**
 - **Accuracy:** Percentage of correctly classified students.
 - **Classification Report:**
 - Precision, Recall, F1-score for each class (high-performing and low-performing).

6. Visualization

- **Decision Regions:**
 - A mesh grid was created to visualize the decision regions for the feature (StudentAbsenceDays).
 - Predicted regions for high and low-performing students were plotted using the Gradient Boosting model.
- **Scatter Plot:**
 - Sampled training and testing data points were plotted to show their distribution in the decision regions.
- **Axis and Labels:**
 - The x-axis represents StudentAbsenceDays (attendance level).

- The y-axis represents the performance class (1 for high-performing, 0 for low-performing).

Results and Insights

1. Accuracy:

- The Gradient Boosting model achieved an accuracy of **75.6%** indicating its effectiveness in predicting student performance.

2. Classification Report:

- Precision, Recall, and F1-score provided detailed insights into how well the model distinguished between high-performing and low-performing students.

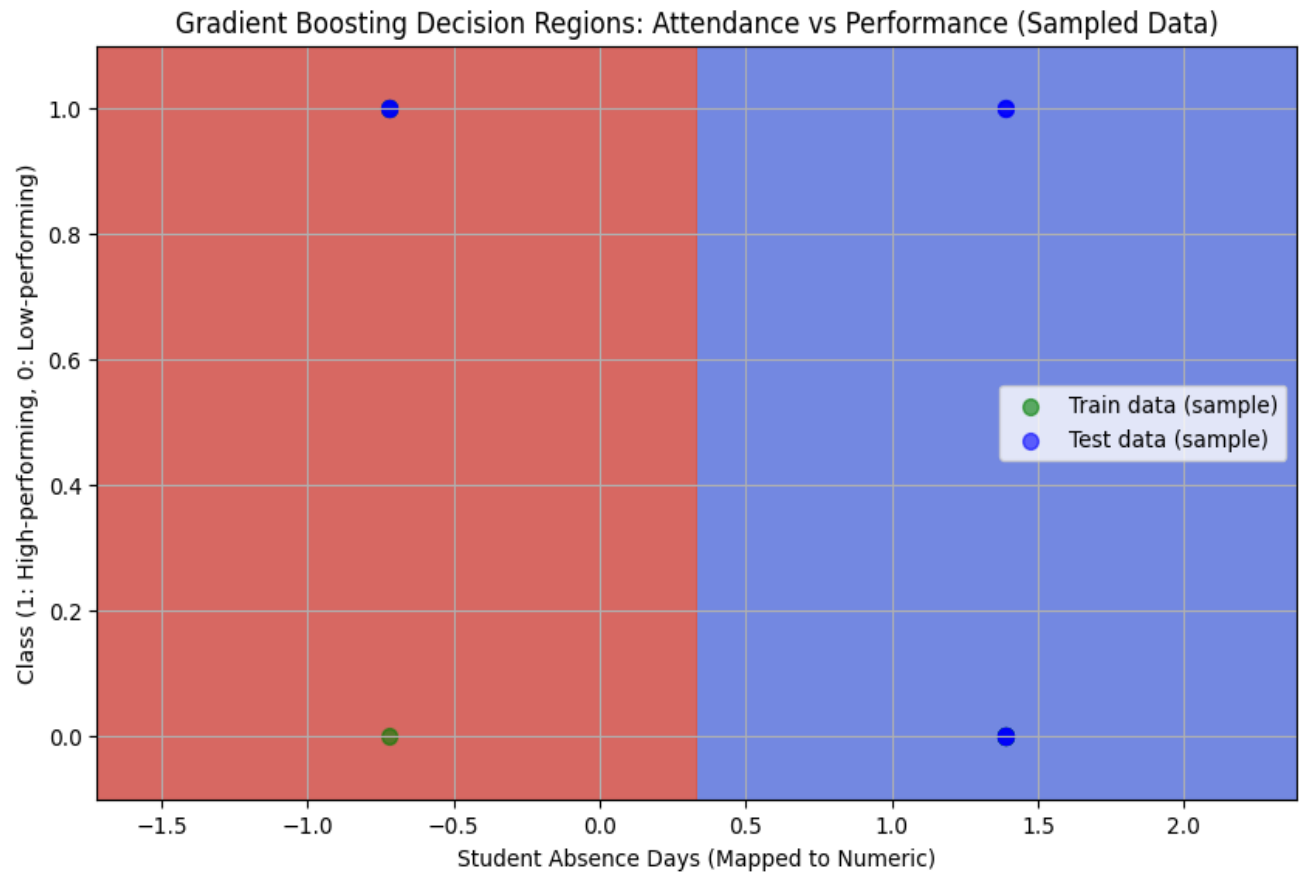
Gradient Boosting Model Classification Report:				
	precision	recall	f1-score	support
0.0	0.57	0.96	0.71	45
1.0	0.97	0.67	0.79	99
accuracy			0.76	144
macro avg	0.77	0.81	0.75	144
weighted avg	0.84	0.76	0.77	144

3. Decision Regions:

- The decision region visualization demonstrated how the model separates students into high-performing and low-performing groups based on attendance.

4. Sample Visualization:

- Sample data points from the training and testing sets were plotted, confirming the model's predictions align with the decision regions.



Advantages

1. High Predictive Power:

- Gradient Boosting provided higher accuracy compared to simpler models due to its ability to minimize prediction errors iteratively.

2. Feature Interpretability:

- Attendance was clearly identified as a significant factor in predicting performance.

3. Robustness:

- The model handled binary classification effectively, even with limited features.

Conclusion

The Gradient Boosting model effectively addressed the research question:

- **High-performing students** are more likely to have fewer absences (Under-7 days).
- **Low-performing students** are more likely to have more absences (Above-7 days).

The decision region visualization and accuracy metrics highlighted the robustness of the model in predicting student performance based on attendance. These insights can guide educational institutions in focusing on attendance policies to improve overall performance.

Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is well-suited for our dataset because it can learn complex, non-linear relationships between the various features, such as parental involvement, student engagement, and class performance.

By using multiple layers and neurons, ANNs can capture intricate patterns that simpler models might miss. With features like 'ParentAnsweringSurvey', 'ParentschoolSatisfaction', and others, the ANN can effectively predict student class performance, especially when interactions between different features are important.

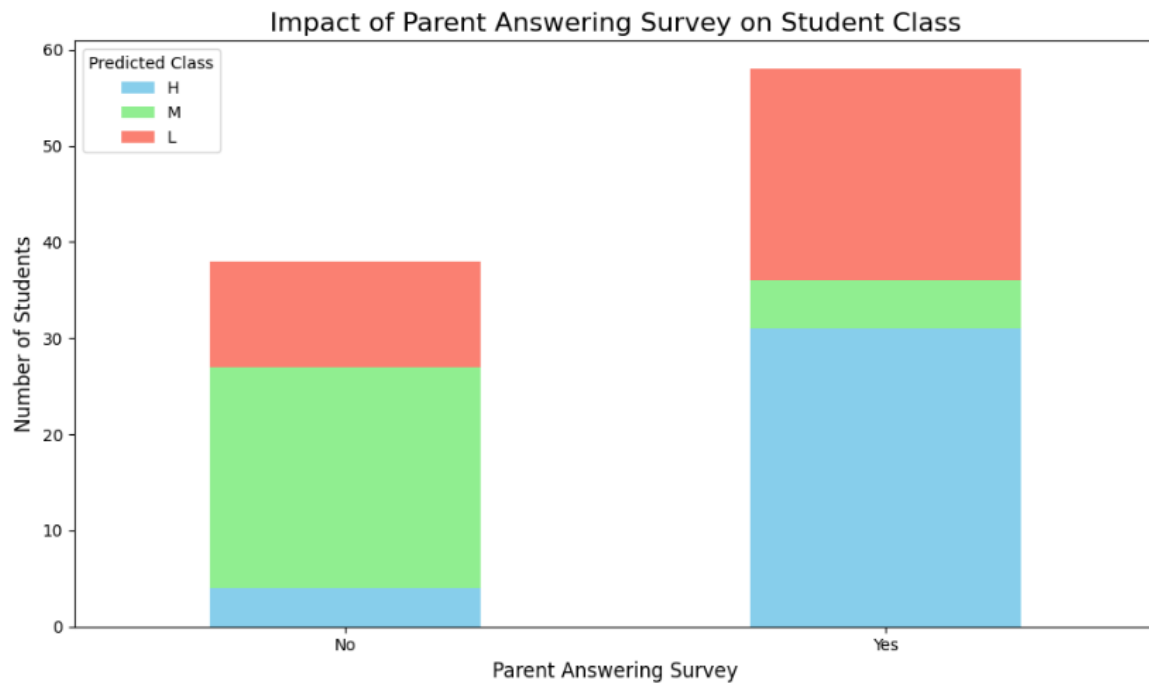
Its flexibility in handling both categorical and continuous data makes it ideal for this type of dataset.

We want to address the following problem statement by using ANN:

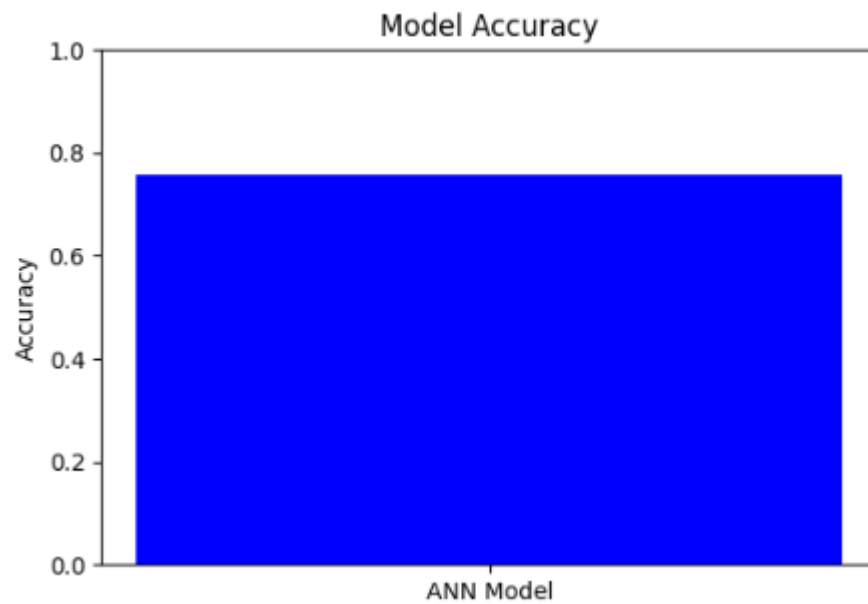
What impact does parental influence has on student performance?

Results and Insights

ANN Accuracy: 0.7569444444444444				
ANN Classification Report:				
	precision	recall	f1-score	support
0.0	0.57	0.96	0.71	45
1.0	0.97	0.67	0.79	99
accuracy			0.76	144
macro avg	0.77	0.81	0.75	144
weighted avg	0.84	0.76	0.77	144



Model Accuracy



Results

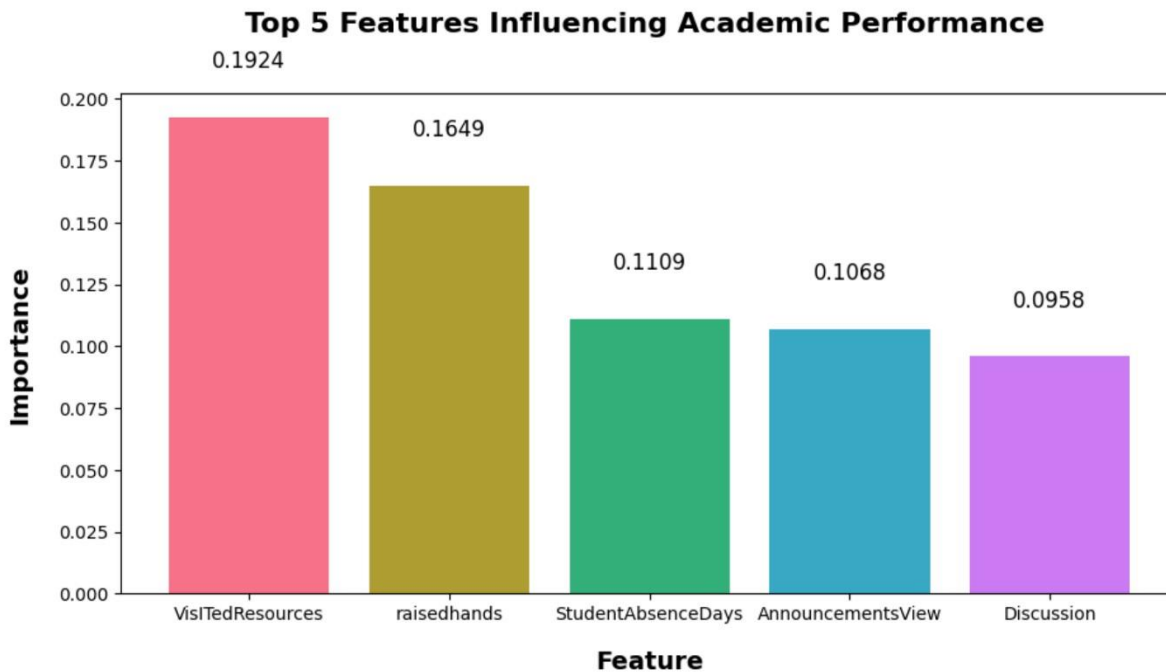
Now, to answer our problem statements.

What are the most important factors that determine the academic performance of students?

Through Random Forest Classification, the following is a sorted list of the most important features that contribute to academic performance of a student:

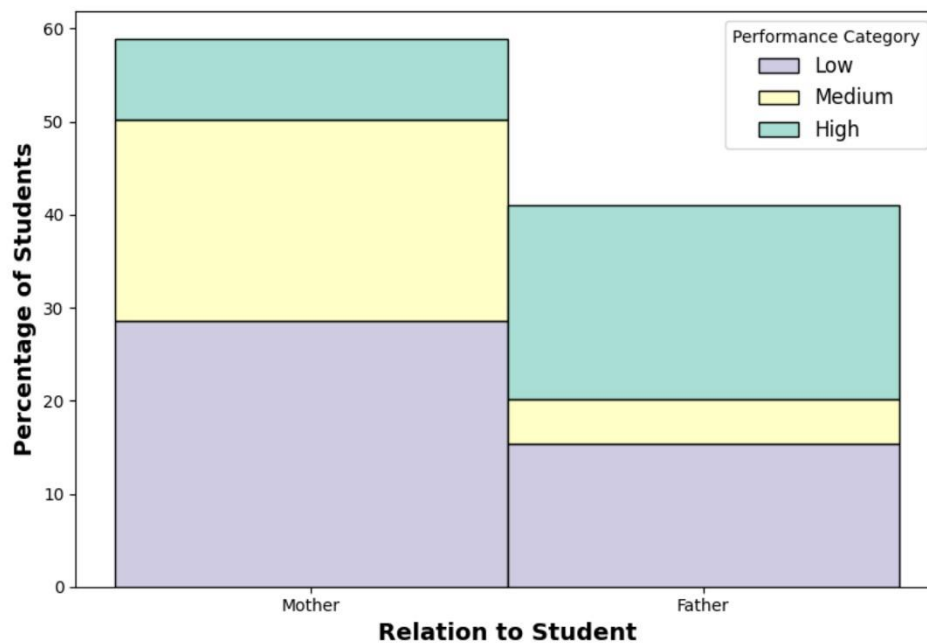
Factors influencing academic performance (Most important to least important):

- VisITedResources (Importance: 0.1924)
- raisedhands (Importance: 0.1649)
- StudentAbsenceDays (Importance: 0.1109)
- AnnouncementsView (Importance: 0.1068)
- Discussion (Importance: 0.0958)
- Relation (Importance: 0.0439)
- Topic (Importance: 0.0419)
- ParentAnsweringSurvey (Importance: 0.0405)
- NationalITY (Importance: 0.0377)
- GradeID (Importance: 0.0331)
- gender (Importance: 0.0330)
- PlaceofBirth (Importance: 0.0311)
- ParentschoolSatisfaction (Importance: 0.0233)
- SectionID (Importance: 0.0190)
- StageID (Importance: 0.0136)
- Semester (Importance: 0.0120)



How does a guardian's role (whether mother or father) impact a child's academic performance?

Percentage Distribution of Student Performance by Relation (Mother vs Father)



- Guardians play a significant role in the academic performance of students, ranking 6th overall.
- Although there were more entries for students with a mother as their guardian in the dataset, our analysis shows that students with a father as their guardian are more likely to perform well (High performance category).

Is there a specific attendance threshold where a student can be classified as high-performing or low-performing?

Through SVM, with 75.6% accuracy we can conclude that the threshold is:

- 7 absences: If a student has fewer than 7 absences (mapped as Under-7), they are likely to be classified as high performing.
- Students with 7 or more absences (mapped as Above-7) are classified as low-performing.

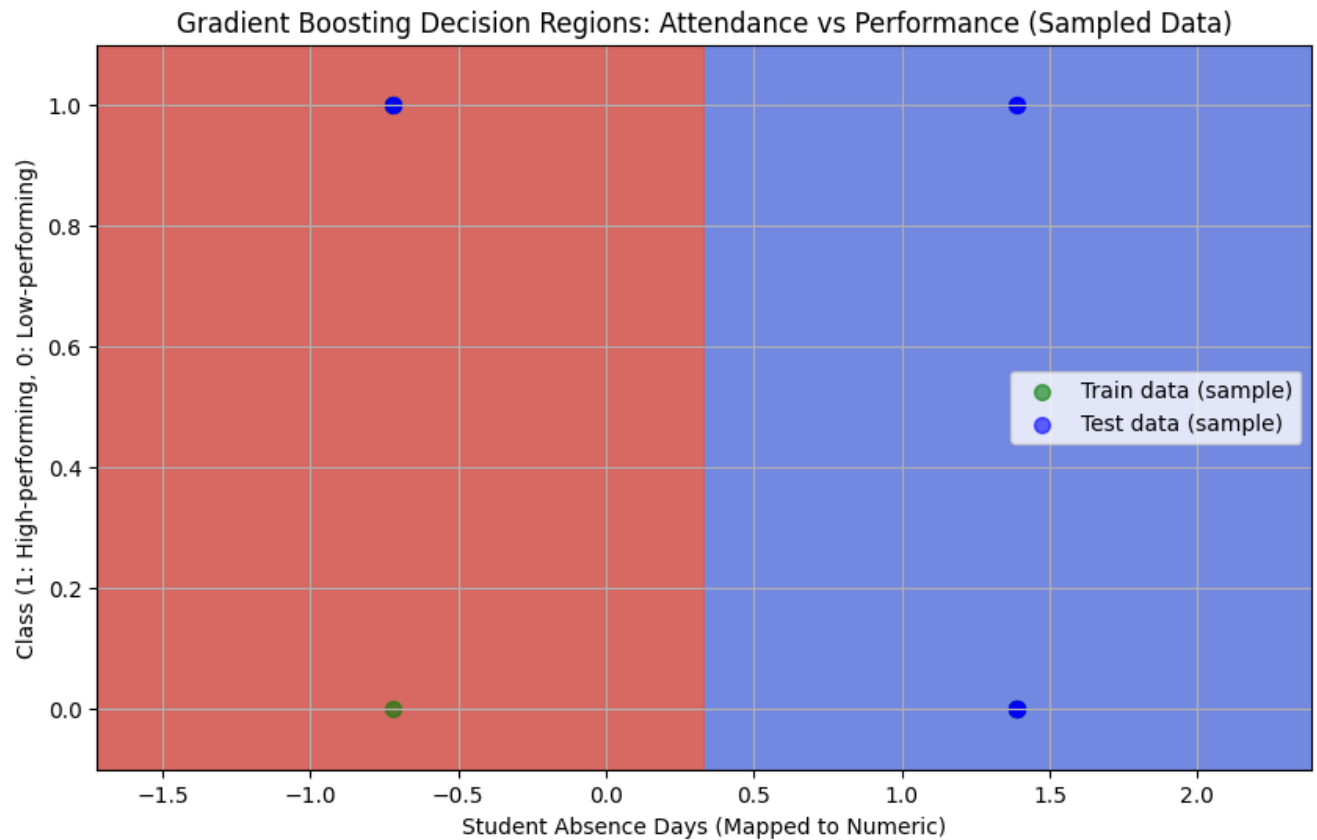
This finding suggests that attendance plays a critical role in student performance, with 7 absences being the key tipping point identified by the model.



Decision Boundary (Threshold for Attendance): 0.5
SVM Accuracy: 0.7569444444444444

How does student attendance correlate with their likelihood of being categorized as high-performing or low-performing?

According to Gradient Boosting model:



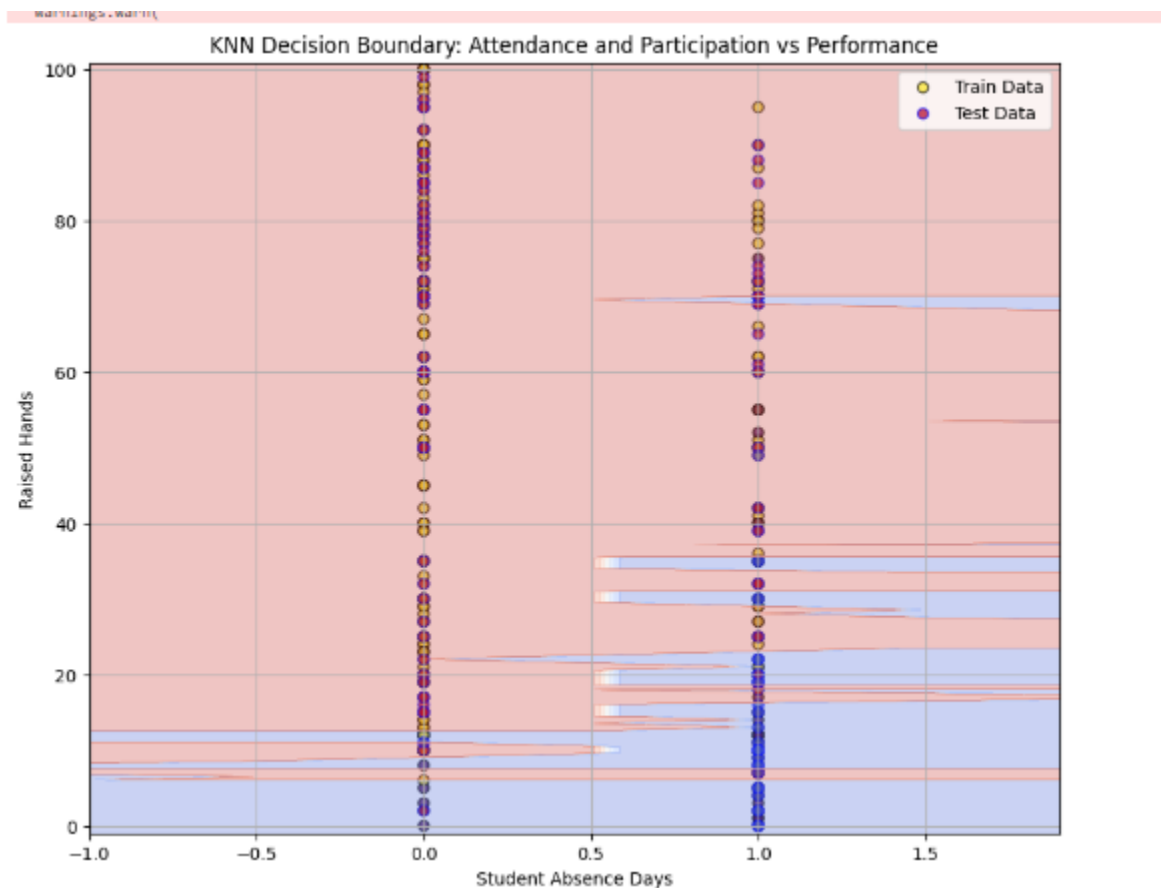
Conclusion

- **High-performing students** are more likely to have fewer absences (Under-7 days).
- **Low-performing students** are more likely to have more absences (Above-7 days).

How do student attendance and participation correlate with their likelihood of being categorized as high-performing or low-performing?

To answer this, we used KNN and Naïve Bayesian Classifiers. KNN achieved an accuracy of 87.50 while Naïve Bayes achieved 84.72%. Clearly KNN provides better insights.

Hence, using KNN results to answer the question:



The model demonstrated that:

- Students with **lower absence days** and **higher participation (raised hands)** are more likely to be categorized as **high-performing**.
- Students with **higher absence days** and **lower participation** are more likely to be categorized as **low-performing**.

The decision boundary plot visually confirmed these trends:

- High-performing students clustered in regions with low absence and high participation.
- Low-performing students clustered in regions with high absence and low participation.

Conclusion

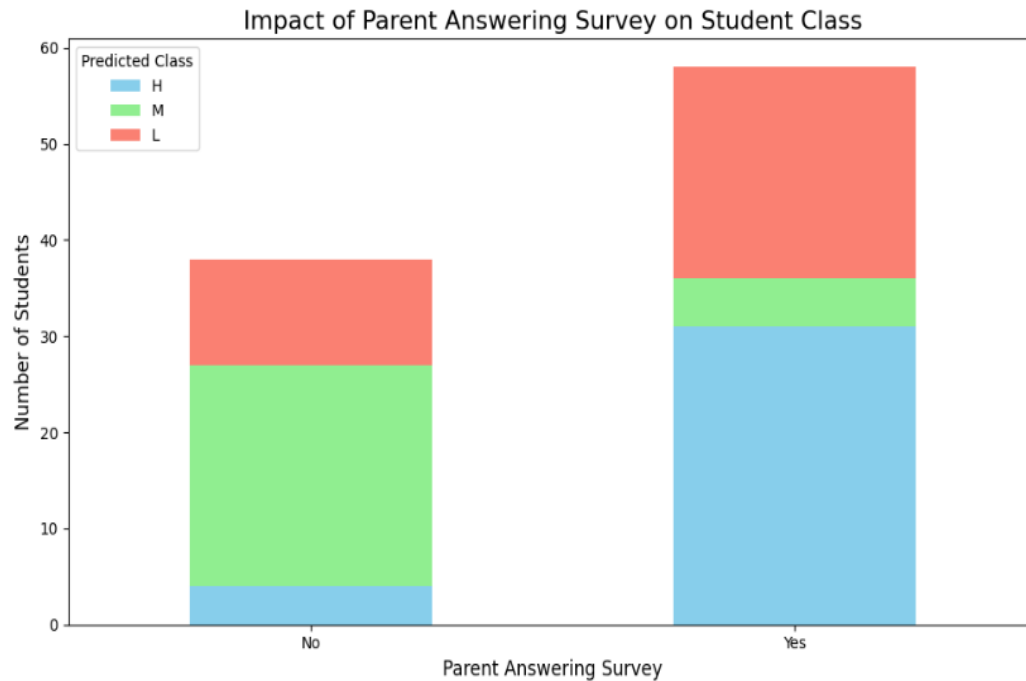
The KNN model effectively addressed the question:

- Students who are more engaged in class (measured by raisedhands) and have fewer absences are more likely to perform well academically.
- Attendance and participation are strong predictions of student performance.

The use of KNN provided a simple yet powerful approach to classify student performance, with clear visualizations supporting the analysis.

What impact does parental influence ("ParentAnsweringSurvey" or "ParentschoolSatisfaction") has on student performance?

ANN results clearly show that:

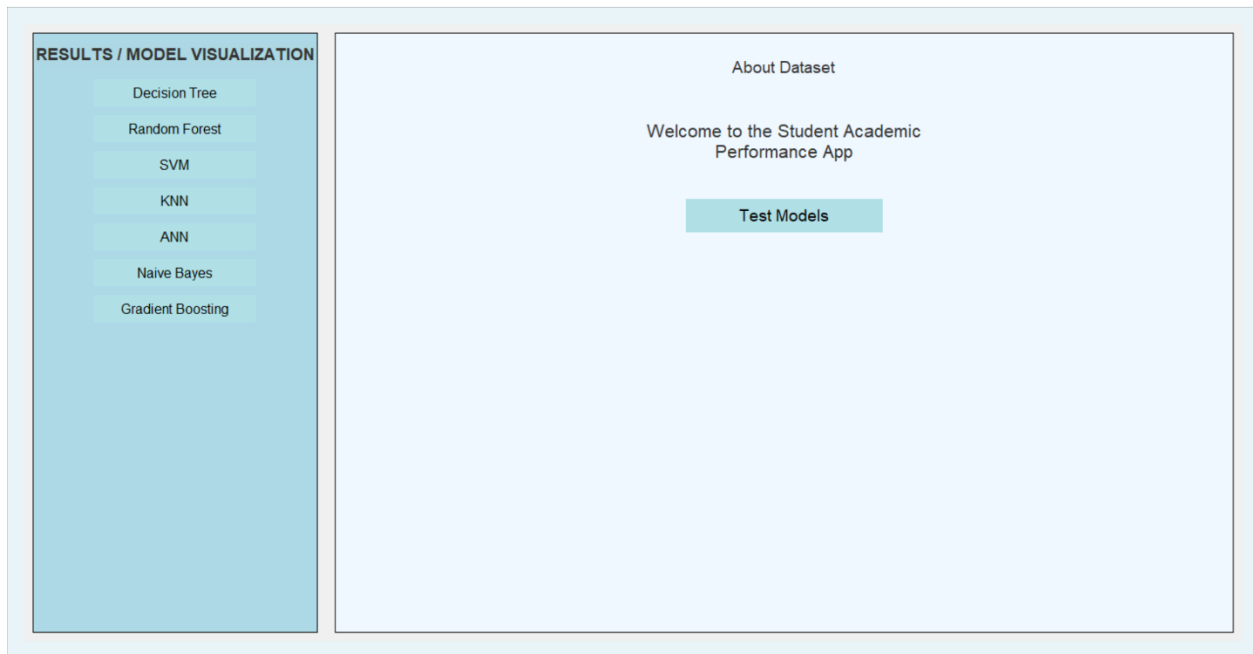


- When parents didn't answer the survey ("No"), the number of Medium and Low-performing students is higher, and there are fewer High-performing students.
- When parents answered the survey ("Yes"), there are more High-performing students and fewer Medium and Low-performing students.

This suggests that students whose parents participate in the survey are more likely to perform better academically.

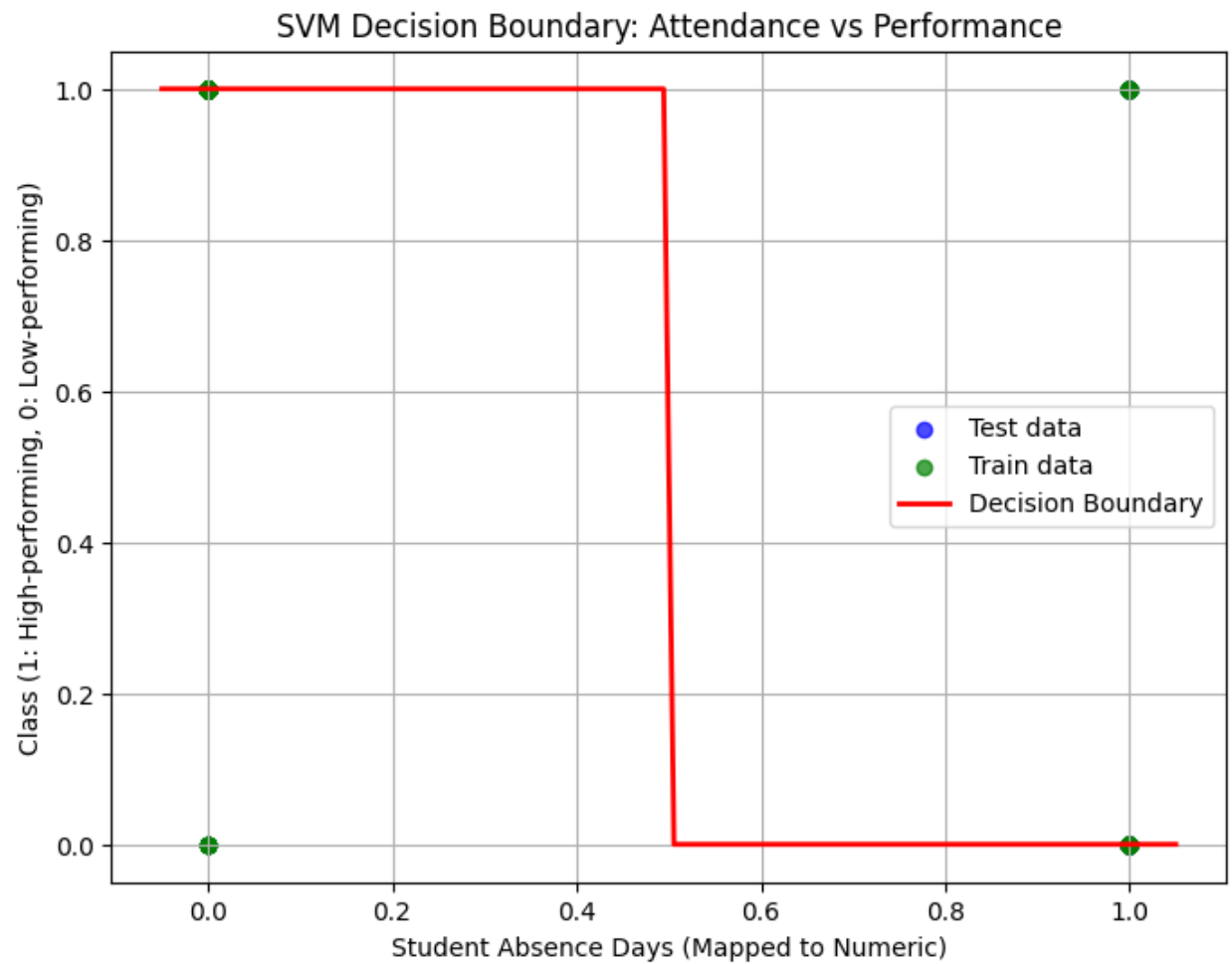
Graphical User Interface

We have used **Tkinter** as it is the standard GUI library for Python, providing a simple way to create desktop applications with graphical elements such as buttons, labels, and windows. Tkinter is easy to use, with widgets for handling user input and displaying outputs in a visually organized manner.

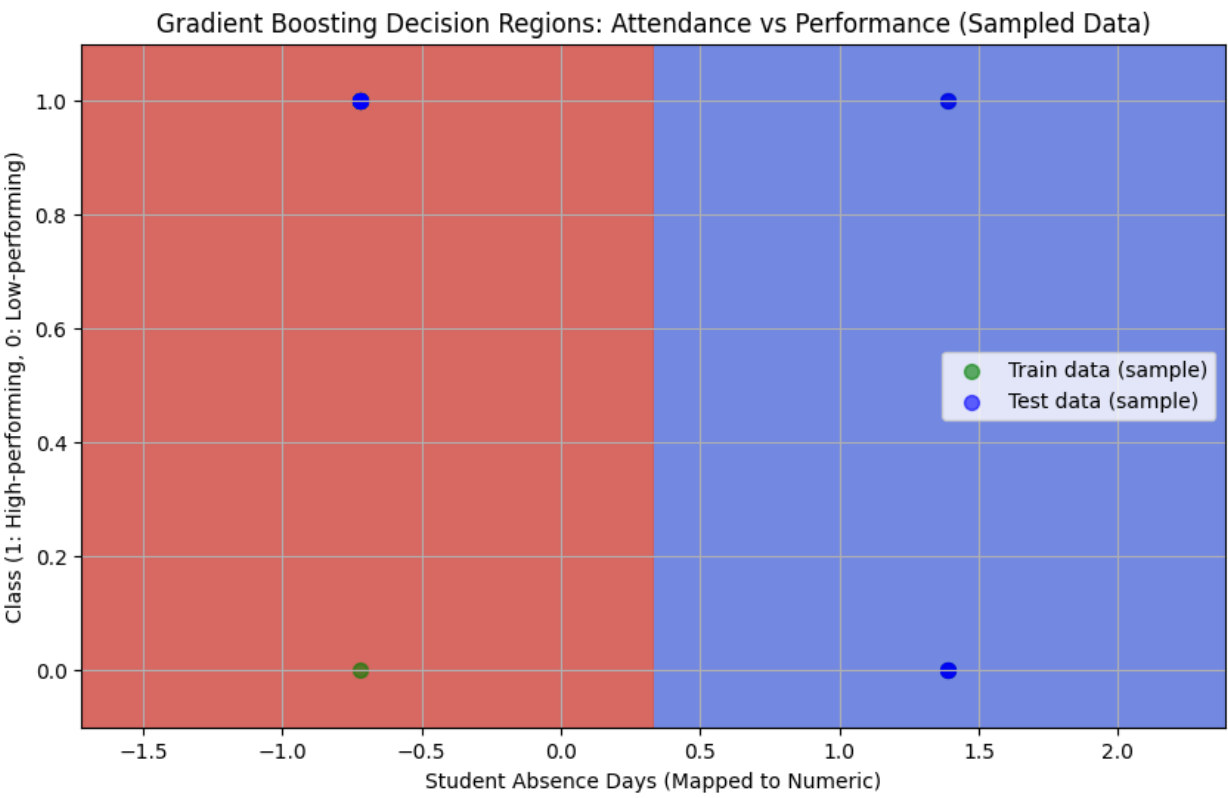


Results/Model Visualization

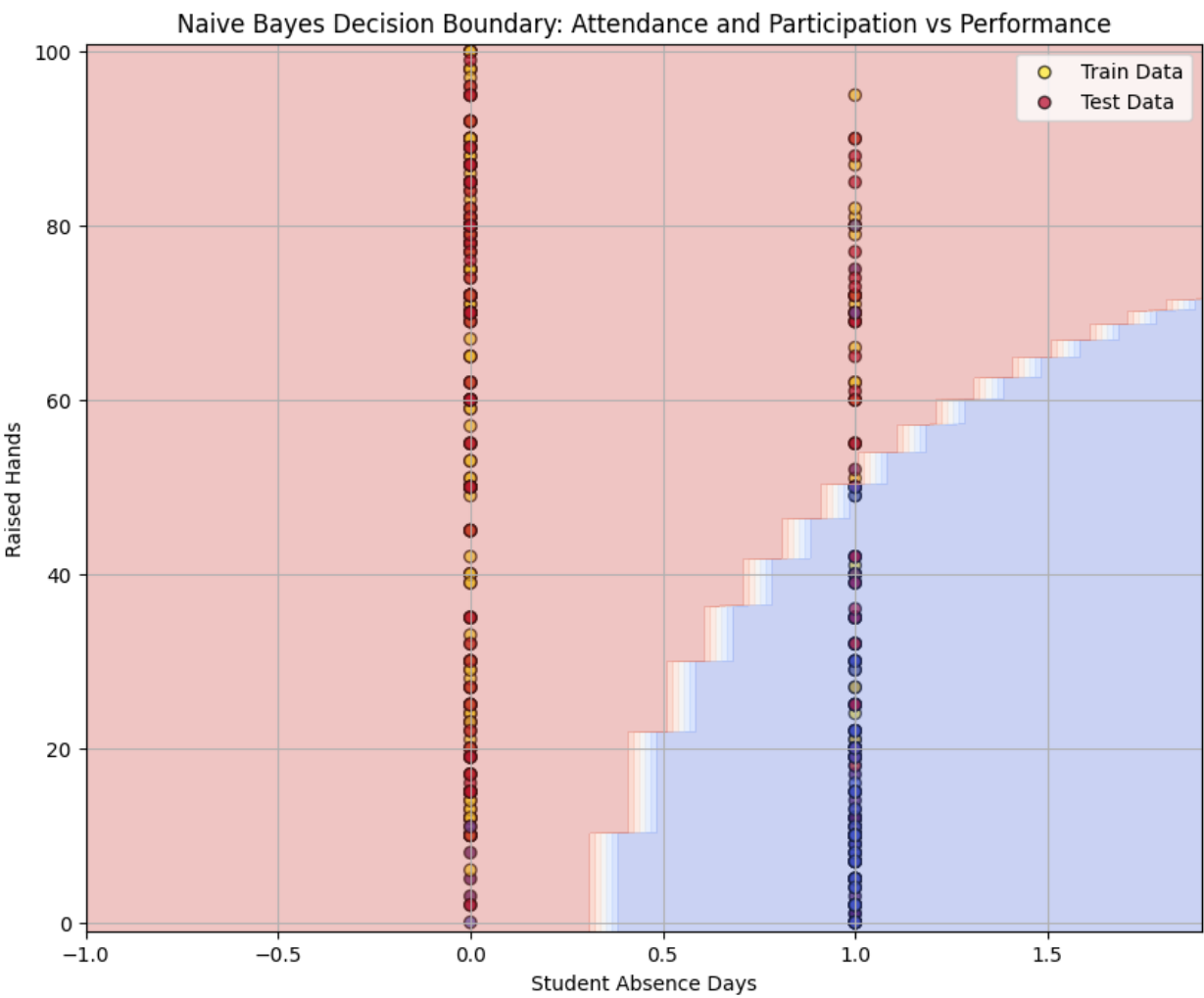
SVM



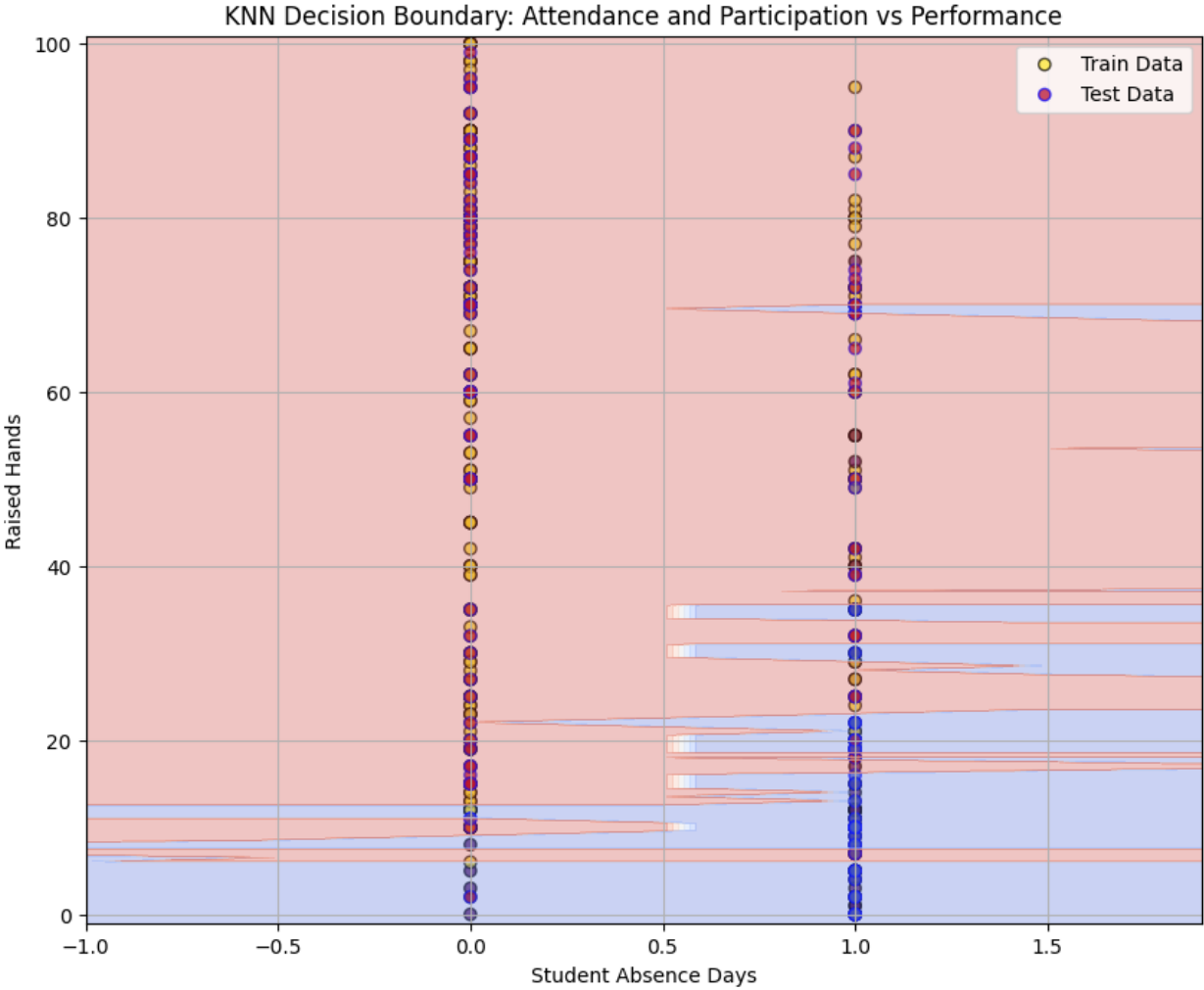
Gradient Boosting



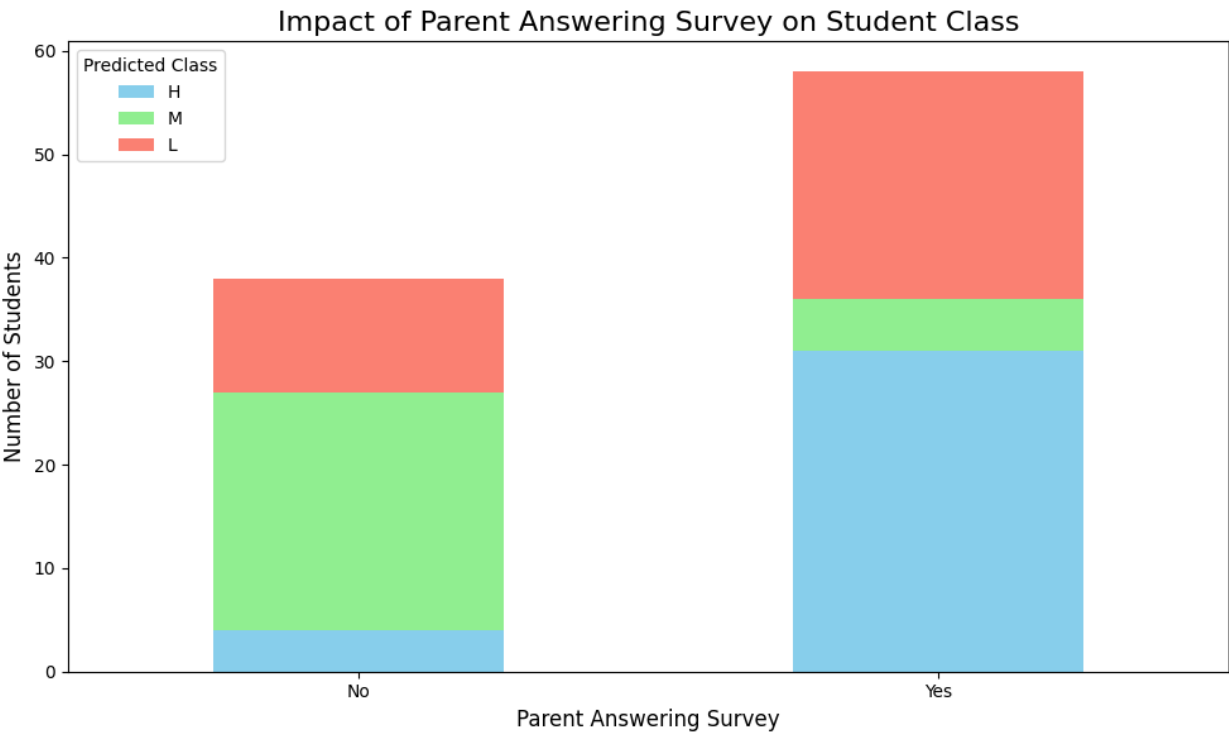
Naïve Bayes



K-Nearest Neighbor (KNN)



Artificial Neural Network (ANN)



Test Models

Decision Tree

Naive Bayes

Gradient Boosting

Decision Tree

Random Forest

SVM

KNN

ANN

Naive Bayes

Gradient Boosting

Decision Tree

Performance Prediction

Enter Raised Hands:

Enter Visited Resources:

Enter Announcements View:

Prediction: Low-performing (0)

Naïve Bayes

Naive Bayes Performance Predictor


Enter Student Absence Days and Raised Hands:

Student Absence Days:

Raised Hands:

Prediction: Low-performing (0)

KNN

 KNN Performance Predictor


Enter Student Absence Days and Raised Hands:

Student Absence Days:

Raised Hands:


Prediction: High-performing (1)

Gradient Boosting

 Performance Predictor

Enter Student Absence Days:

Prediction: Low-performing (0)

 Performance Predictor

Enter Student Absence Days:

Prediction: High-performing (1)

Random Forest

 Random Forest Performance Prediction


Enter Raised Hands:

Enter Visited Resources:

Enter Announcements View:

Prediction: High-performing (1)

ANN

 Performance Prediction

Select Parent Answering Survey: ☒ Yes ☐ No

Select Parent School Satisfaction: ☒ Good ☐ Bad


Enter Raised Hands:

Enter Visited Resources:

Enter Announcements View:

The impact of parents answering survey is low on student performance.

SVM

 SVM Attendance Classifier

Enter Student Absence Days (Mapped to Numeric):

Student Absence Days:

Predict

Prediction: High-performing (1)