Capstone Project

Student Performance Prediction Final Report

1. **Problem Statement:**

Provide actionable insights to Schools, teachers and policy makers to plan and design their programs to boost student performance through guidance and mentoring which can result in better academic outcomes.

1. **Model Outcomes or Predictions:**

* This project uses Machine Learning techniques to analyze student performance based on various features like attendance, study habits, prior grades, and socio-economic factors.
* The model predicts the future grade of the students and provides opportunities for parents, policy makers, teachers & tutors to proactively involve and contribute to the student’s success.

1. **Data Acquisition:** The dataset is downloaded from Kaggle in CSV format and included as part of documentation. Careful consideration is given to selecting a dataset that encompasses key factors influencing performance through grades. While visualization and correlation analysis highlight major contributors to predictive outcomes, it is important to recognize that other features can indirectly impact these primary contributors.

Hours\_Studied: Number of hours spent studying per week.

Attendance: Percentage of classes attended.

Parental\_Involvement: Level of parental involvement in the student's education (Low, Medium, High).

Access\_to\_Resources: Availability of educational resources (Low, Medium, High).

Extracurricular\_Activities: Participation in extracurricular activities (Yes, No).

Sleep\_Hours: Average number of hours of sleep per night.

Previous\_Scores: Scores from previous exams.

Motivation\_Level: Student's level of motivation (Low, Medium, High).

Internet\_Access: Availability of internet access (Yes, No).

Tutoring\_Sessions: Number of tutoring sessions attended per month.

Family\_Income: Family income level (Low, Medium, High).

Teacher\_Quality: Quality of the teachers (Low, Medium, High).

School\_Type: Type of school attended (Public, Private).

Peer\_Influence: Influence of peers on academic performance (Positive, Neutral, Negative).

Physical\_Activity: Average number of hours of physical activity per week.

Learning\_Disabilities: Presence of learning disabilities (Yes, No).

Parental\_Education\_Level: Highest education level of parents (High School, College, Postgraduate).

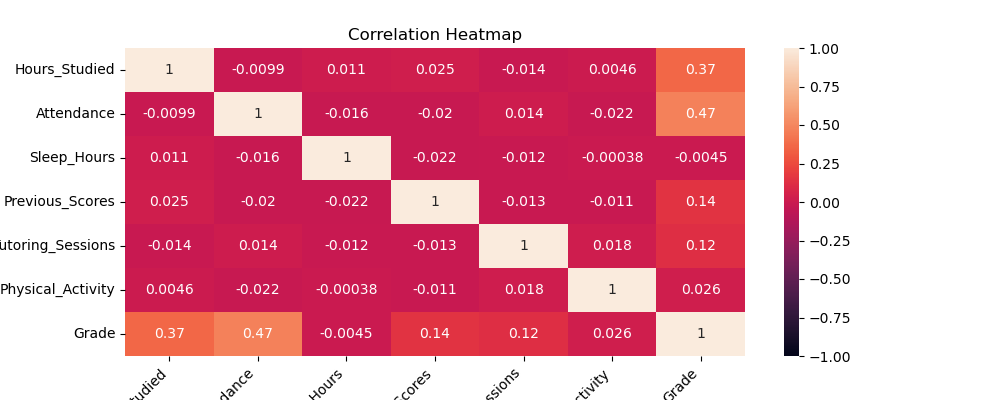
Distance\_from\_Home: Distance from home to school (Near, Moderate, Far).

Gender: Gender of the student (Male, Female).

Grade: Grade of Student: 1 - F; 2 - D; 3 - C; 4 - B; 5 – A

**Visualizations of the acquired dataset:**

* Heat Map to display the correlation between numerical features
* Box Plots to display numerical data distribution with target variable Grade
* Pie Charts to display data distribution of categorical features
* Count Plots to display the categorical features with target variable Grade

****

**A group of boxes with text

AI-generated content may be incorrect.**

**A screenshot of a graph

AI-generated content may be incorrect.A diagram of pie charts

AI-generated content may be incorrect.**

* From the correlation matrix and Plots
* Hours\_Studied and Attendance are highly correlated with Grade; increasing them improves student grades.
* Tutoring\_Sessions and Previous\_Scores have moderate correlation; raising them slightly boosts grades.
* 98% of students fall into Grade 2 & Grade 3, suggesting data bias towards these grades and potential lower accuracy for higher grades.
* Significant variance is observed in Tutoring\_Sessions, Previous\_Scores, Hours\_Studied, and Attendance across grades, while other features show minimal variance.
* These are key factors in Grade classification.

1. **Data Preprocessing/Preparation:** Training a machine learning model needs data to be cleaned, transformed, and structured for better accuracy. This step involves handling missing values, encoding categorical variables, scaling numerical features, and selecting relevant features. Proper preprocessing ensures the dataset is optimized for model training, improving performance and reducing bias.

* Handled missing data using Mode based imputation
* Encoded the categorical features using Ordinal Encoder and One Hot Encoder.
* Numerical data is scaled using Standard Scaler.
* The data is split into train and test sets with 80% training size and 20% test size.

1. **Modeling:** The problem is classification problem as we predict the student grade based on available features. Also, we have multiple classes for prediction. We have considered the following models to solve this problem.
2. Decision Trees
3. KNN
4. Random Forests
5. SVM
6. XGBoost
7. Logistic Regression

We then perform GridSearchCV to perform cross validation and hyper parameter tuning to get better evaluation metrics so that the model performs well for unseen data.

Basic Model Performance Comparison:

***Model Accuracy Precision Recall F1-Score***

Decision Tree 0.838880 0.439180 0.495923 0.459694

KNN 0.842663 0.334571 0.303716 0.313313

Random Forest 0.916793 0.369570 0.352102 0.359083

SVM 0.934191 0.375063 0.365213 0.369403

**XGBoost 0.946293 0.546095 0.462195 0.492561**

Logistic Regression 0.909228 0.492138 0.461658 0.475387

Tuned Model Comparison

A screenshot of a computer

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

1. **Model Evaluation:** Based on the above training, testing, validation and tuning of the models, we consider the below evaluation metrics to pick best model for our problem.

**Evaluation metrics for this classification:**

1. Accuracy – How well the model predicts the student grade.
2. Precision - How well the model predicts the higher-grade students
3. Recall – How well the model predicts the lower grade students
4. F1 score – How well the model classifies both higher and lower grade students.
5. Confusion matrix – Matrix that show case actual and misclassified predictions.

The model with the best accuracy, F1 score, and recall will be selected for these predictions. It is essential to predict student grades accurately and address the performance of lower-performing students without negatively impacting the grades of high-performing students.

* In this classification problem:

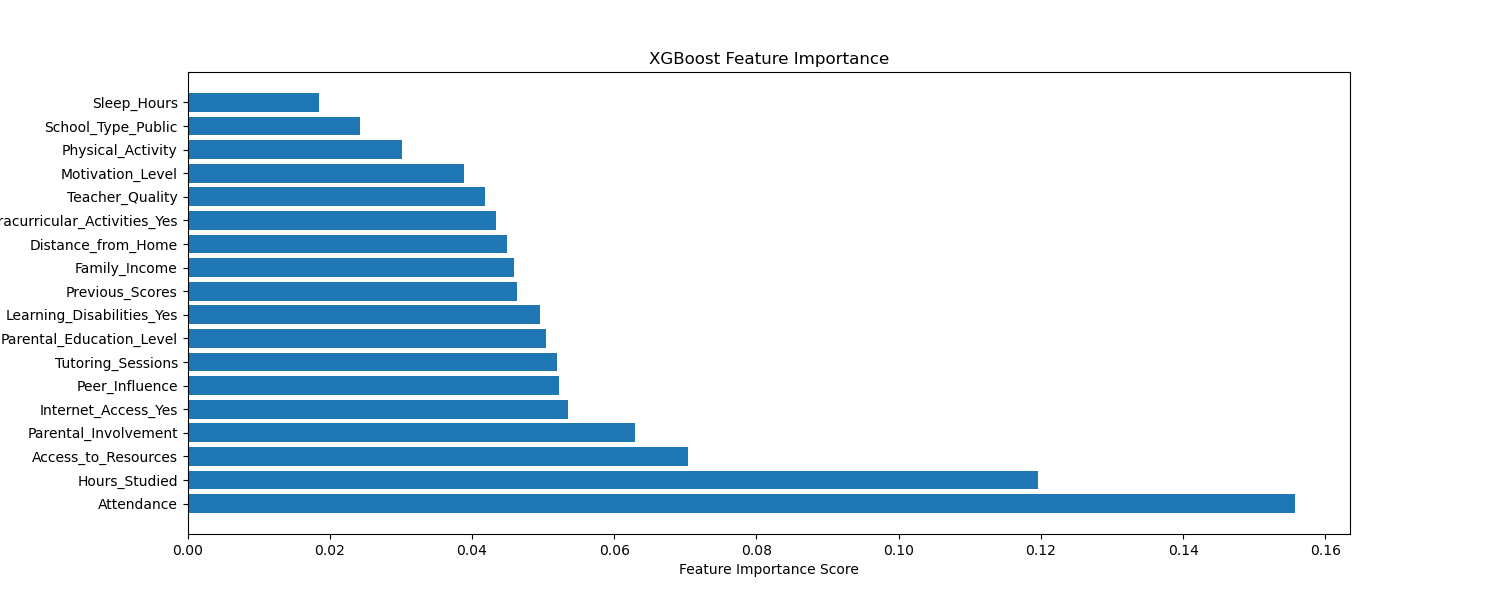
False Positive: Higher-grade students are classified into lower grades, which is acceptable as they can receive guidance.

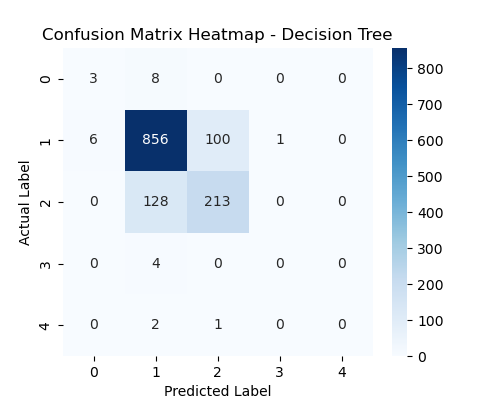
False Negative: Lower-grade students are classified into higher grades, which is problematic since they might miss needed guidance.

* We must rank models based on metrics that minimize False Negatives while balancing False Positives and Negatives.
* XGBoost appears to be the best model for this issue.
* XGBoost excels in Recall and F1 Score, essential for our imbalanced dataset.
* Attendance, Hours\_Studied, Access\_to\_Resources, Parental\_Involvement, Internet Access seems to be top 5 Important features.
* From personal experience as a student, the important features given by the model seems obvious in improving better academic performance.

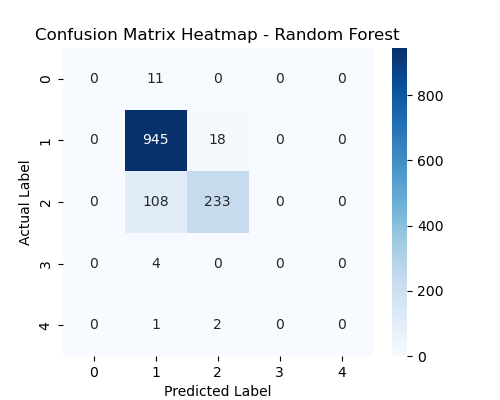
A black text with black letters

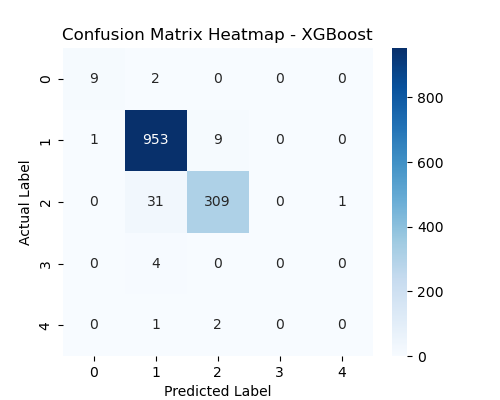
AI-generated content may be incorrect.



A graph of a heatmap

AI-generated content may be incorrect.A graph of a graph with numbers and a number

AI-generated content may be incorrect.A graph of a heatmap

AI-generated content may be incorrect.

**Insights**

* **Takeaways**

1. Attendance at school and study hours are significant factors in achieving good grades.
2. Access to appropriate resources is important for academic success.
3. Positive peer influence, parental involvement, and educational background have a beneficial effect on student performance.
4. There are several indirect factors that can contribute to a student's personal growth, such as physical activity and motivation levels.

* **Recommendations**

*Schools:*

1. Encourage students to maintain consistent attendance.
2. Design awareness programs to increase study hours outside of school.
3. Facilitate student participation in physical and extracurricular activities for personal development.

*Policy makers:*

1. Provide necessary resources for all students to enhance their studies.
2. Develop assistance and relief programs to offer internet services to students from low-income families.

*Teachers & Educators:*

1. Identify students with declining grade predictions compared to previous scores and schedule interactive sessions.
2. Proactively engage with students and provide guidance to improve their motivation for academic success.

* **Future Scope**

1. Information related to social interactions can yield deeper insights.
2. Features reflecting student interests can enhance understanding and guide them in specific areas.
3. Plans for college education and career interests can facilitate mentorship and guidance accordingly.