NATIONAL TEXTILE UNIVERSITY

Department of Computer Science

Student Performance Analysis and Prediction Project Report

Submitted to:

Dr. Muhammad Adeel

Submitted by:

Noorhan Yaseen (20-NTU-CS-1175)

Date of Submission:

January 08, 2024

Session:

BSIT (2020-2024)

Introduction

In contemporary educational paradigms, assessing student performance has transcended conventional methods. Utilizing comprehensive datasets encompassing demographic nuances, academic history, and multifaceted factors, educators are increasingly adopting advanced analytics and machine learning models. These tools not only personalize learning experiences but also identify areas for improvement and streamline resource allocation efficiently.

This project undertakes the crucial task of examining the correlation between student test scores and various determinants, such as Gender, Ethnicity, Parental Education, Lunch preferences, and Test Preparation. By employing data mining techniques and classification algorithms, it aims to predict student performance, unearth discrepancies in teaching models, and discern anomalies in assessment patterns.

Understanding the Problem Statement

This project aims to analyze how students' test scores relate to variables like Gender, Ethnicity, Parental Education, Lunch, and Test Preparation. It seeks to uncover insights using data mining techniques to predict student performance, identify issues in teaching models, detect anomalies in assessments, and predict enrollment patterns. Leveraging classification methods like decision trees and probabilistic classification, it assesses student capabilities across subjects, facilitating the identification of at-risk students and aiding educators in providing targeted support and counseling.

Dataset Information

- gender: students -> (Male/female)
- race/ethnicity: ethnicity of students -> (Group A, B, C, D, E)
- parental level of education: parents' final education ->(bachelor's degree, some college, master's degree, associate's degree)
- lunch: having lunch before test (standard or free/reduced)
- test preparation course: complete or not complete before test
- math score
- reading score
- writing score

Exploring Data (Visualization)

The process of EDA encompassed thorough visualization techniques and statistical exploration of the dataset:

- **Visualizations:**

Utilizing Histograms, Kernel Density Estimation (KDE), and combined plots to visualize the distribution of scores across Math, Reading, and Writing.

- **Insights Gained:**

Balanced Gender Distribution: With 48% female and 52% male students.

- **Ethnic Diversity:**

Predominance of students from Groups C and D, while Group A had the lowest representation.

- **Parental Education Levels:**

Majority of parents having completed college-level education.

- **Bivariate Analysis:**

Students with parents holding master's and bachelor's degrees exhibited higher scores.

Methodology

1. Data Collection and Understanding:

- The dataset acquisition process involved gathering information on student demographics, parental education, lunch preferences, test preparation, and academic performance across multiple subjects.
- An initial assessment ensured the dataset's completeness, addressing potential issues like missing values, duplicate entries, or inconsistencies.
- Understanding the structure and contents of the dataset was crucial to proceeding with subsequent analyses effectively.

2. Data Preprocessing:

- Data cleaning procedures were implemented to handle missing values and ensure uniformity in the dataset.

- Techniques like imputation or removal of missing values were employed based on the nature and context of the missing data.
- Categorical variables were transformed into numerical formats via encoding methods like one-hot encoding or label encoding to prepare the data for analysis.

3. Exploratory Data Analysis (EDA):

- Visualization techniques, including histograms, scatter plots, box plots, and heatmaps, were utilized to unveil patterns, distributions, and correlations within the dataset.
- Statistical measures such as mean, median, variance, and correlation coefficients were calculated to gain deeper insights into relationships between variables.
- EDA aimed to discover trends, outliers, and potential relationships between features and target variables.

4. Feature Selection and Engineering:

- Features were meticulously examined to select the most influential ones in predicting student performance.
- Techniques such as correlation analysis, feature importance scores from models, or domain expertise were used to identify relevant features.
- New features might have been engineered by combining or transforming existing features to enhance the model's predictive capabilities.

5. Model Selection and Training:

- The dataset was split into training and validation sets to facilitate model training and evaluation.
- Various regression models including Linear Regression, Lasso Regression, Decision Trees, Random Forests, XGBoost, CatBoost, and AdaBoost were instantiated and trained using the training dataset.

6. Hyper parameter Tuning:

- Techniques like GridSearchCV and RandomizedSearchCV were applied to fine-tune the hyperparameters of the models.
- This step aimed to optimize the models by identifying the best set of hyperparameters, enhancing their performance and generalization ability.

7. Model Evaluation and Comparison:

- The performance of each model was evaluated using established regression metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 Score.
- Comparative analysis was conducted to determine the model that exhibited the most accurate predictions and robust performance.

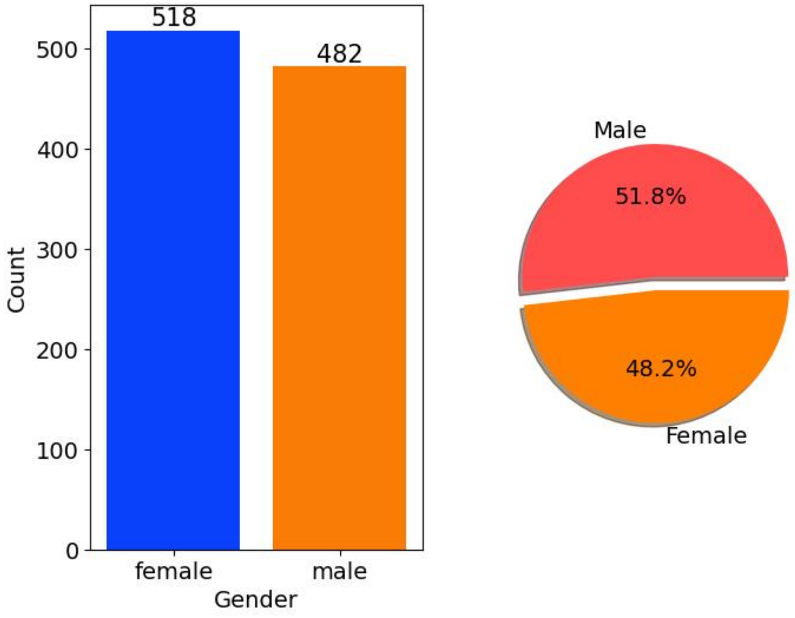
8. Selection of Best-performing Model:

Based on comprehensive evaluation and comparison, the Linear Regression model emerged as the most effective in predicting student performance, demonstrating an accuracy rate of 88.03%.

Exploring Data (Visualization)

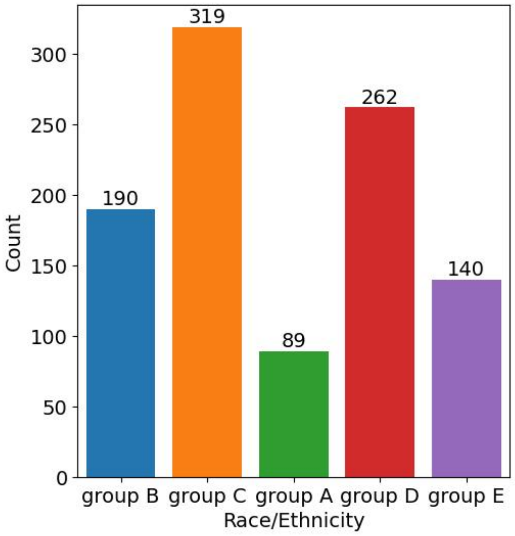
Visualize Average Score Distribution

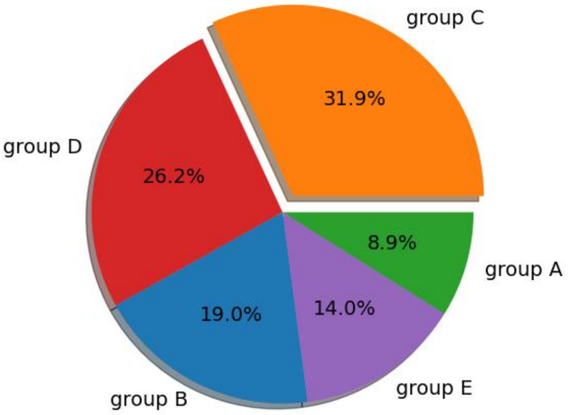
Gender Column



The dataset displays balanced gender representation, with female students accounting for 518 (48%) and male students for 482 (52%). This gender balance within the dataset ensures a fair representation and minimizes biases that might skew the analysis or model predictions based on gender-specific trends or patterns.

Race/Ethnicity Column



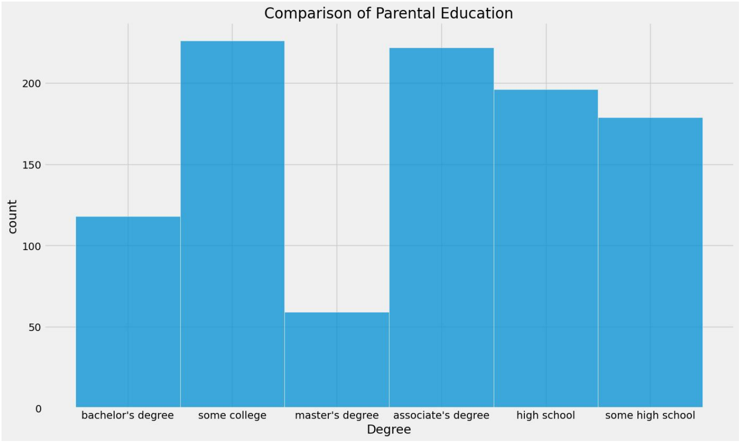


Insights from the Race/Ethnicity column highlight that a substantial number of students belong to groups C and D, while the lowest representation is observed in Group A. This distribution might reflect certain demographic characteristics within the dataset, indicating potential disparities in academic representation across ethnic groups, which could impact performance analysis.

id = Insights>Insights

- Most of the student belonging from group C /group D.
- Lowest number of students belong to group A.

Parental Level of Education Column

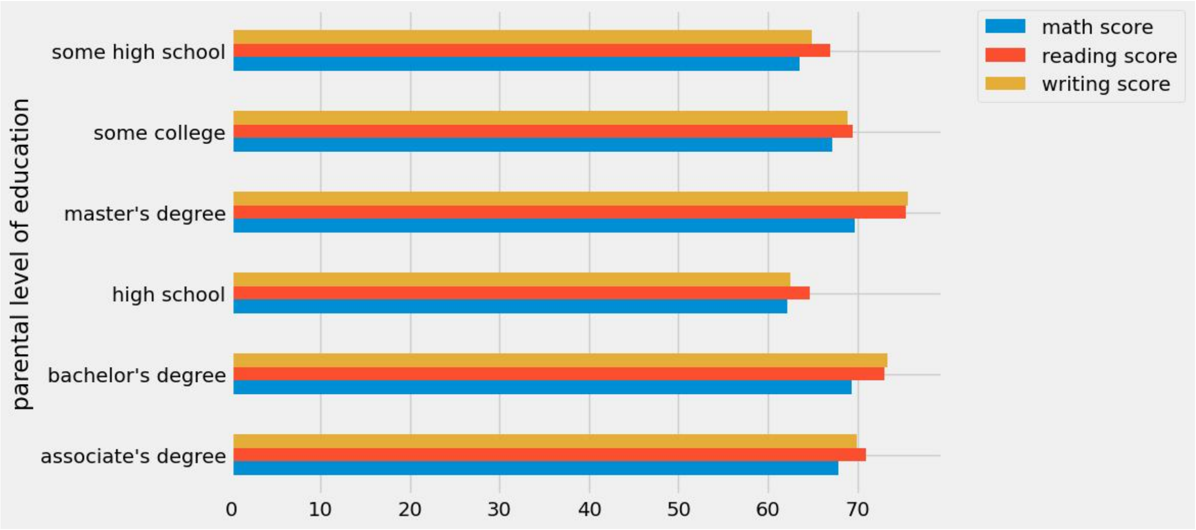


Observing the parental education levels, the dataset portrays a notable presence of parents with a college education, indicating that a significant proportion of students may come from families with higher educational backgrounds. This aspect often influences student performance, as parental education levels can correlate with the academic support and guidance students receive at home.

id = Insights>Insights

- Largest number of parents are from college.

Bivariate Analysis

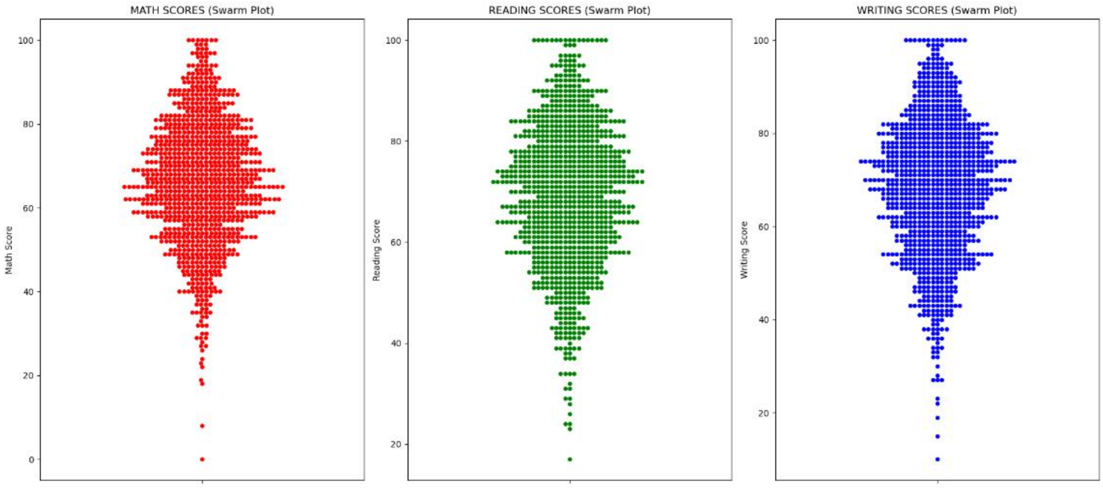


The bivariate analysis indicates a positive correlation between parental education levels and student scores. Specifically, students with parents possessing master's and bachelor's degrees tend to achieve higher scores. This finding underscores the impact of parental education on student academic outcomes, emphasizing the importance of a conducive educational environment at home.

id = Insights>Insights

- The score of student whose parents possess master and bachelor level education are higher than others.

Maximum Score of Students in All Three Subjects

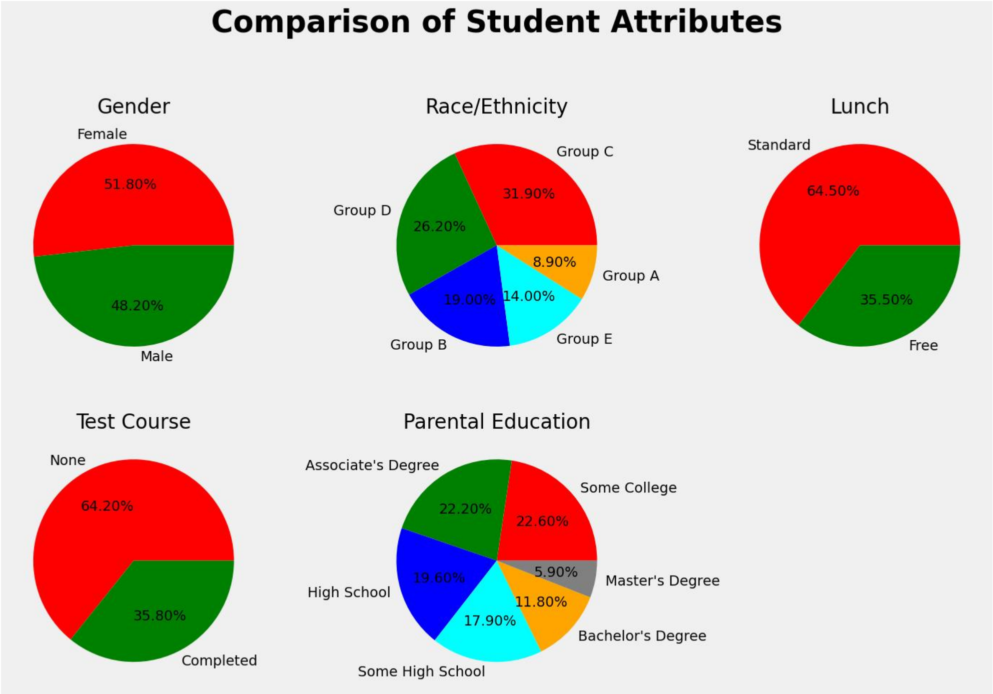


The analysis of maximum scores across subjects reveals the concentration of student scores within specific ranges. Most students score between 60-80 in Math, while in Reading and Writing, the majority perform within the 50-80 range. This distribution sheds light on the predominant performance levels across subjects, allowing for a targeted focus on areas where students might excel or struggle.

id = Insights>Insights

From the above three plots its clearly visible that most of the students score in between 60-80 in Maths whereas in reading and writing most of them score from 50-80

Insights from Visualization and Student Characteristics



id = Insights>Insights

- The number of Male and Female students is almost equal.
- The number of students is higher in Group C.
- The number of students who have standard lunch is greater.
- The number of students who have not enrolled in any test preparation course is greater.
- The number of students whose parental education is “Some College” is greater followed closely by “Associate’s Degree”.

From the above plot, it is clear that all the scores increase linearly with each other.Student’s Performance is related to lunch, race, and parental level education.

- Females lead in pass percentage and also are top-scorers.
- Student Performance is not much related to test preparation course.
- The finishing preparation course is beneficial.

Model Training and Evaluation

Preprocessing:

- The initial step involved separating the features (independent variables) from the target variable (dependent variable, i.e., student scores).
- The ColumnTransformer was utilized to apply specific transformations, such as encoding categorical variables or scaling numerical features, ensuring the data was appropriately formatted for modeling.

Model Selection:

Multiple regression models were evaluated to ascertain the most suitable approach for predicting student performance. Models considered included:

- Linear Regression
- Lasso Regression
- K-Neighbors Regressor
- Decision Tree
- Random Forest Regressor
- XGBRegressor
- CatBoosting Regressor

- AdaBoost Regressor

Hyper parameter

- GridSearchCV or RandomizedSearchCV techniques were employed to fine-tune the hyperparameters of each regression model.
- Hyperparameters significantly impact model performance, and optimizing them ensures improved accuracy and robustness

Tuning:

Optimized model hyperparameters using GridSearchCV or RandomizedSearchCV.

Model Evaluation:

The performance of each model was rigorously assessed using various regression metrics, including:

- Root Mean Squared Error (RMSE)
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R2 Score (Coefficient of Determination)

Model Selection and Performance

- After extensive evaluation, Linear Regression emerged as the best-performing model, exhibiting the highest accuracy rate of 88.03%.
- Detailed comparative analysis of the models allowed for a comprehensive understanding of their respective strengths and weaknesses.
- The examination of differences between actual and predicted values offered insights into the predictive capabilities and performance characteristics of the chosen Linear Regression model.

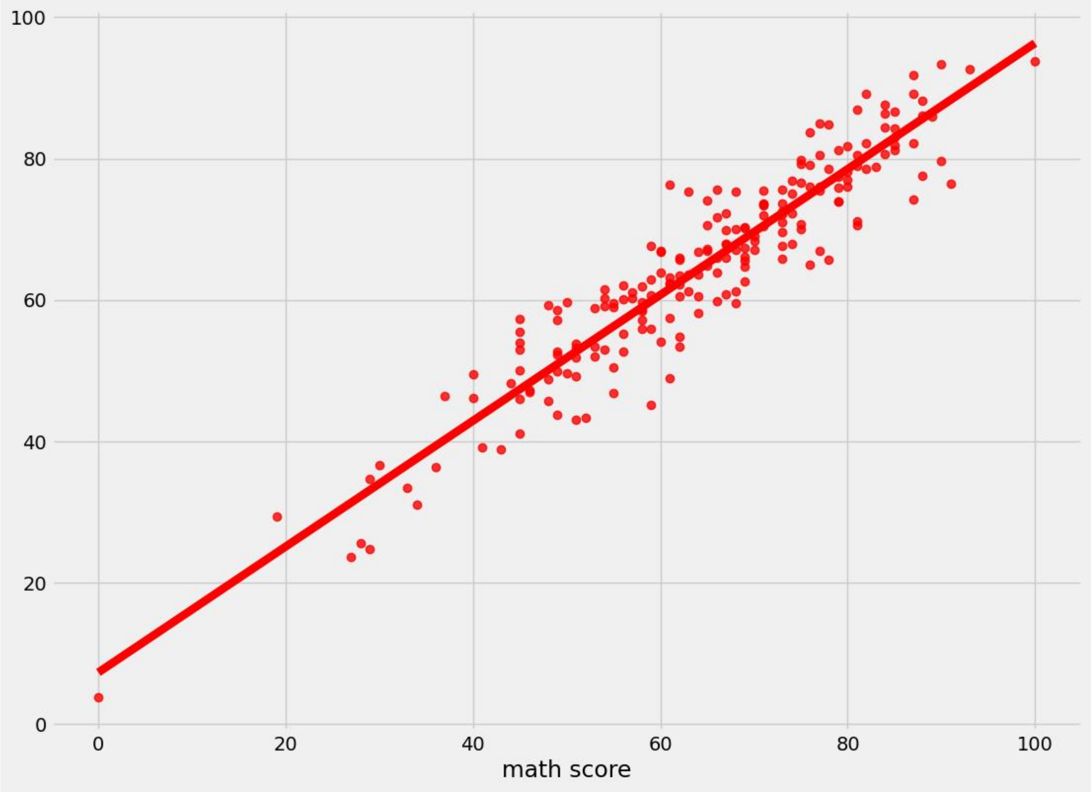
Key Insights:

- **Significance of Student Result Analysis and Prediction:**

➢ The ability to predict student performance is pivotal for educational institutions, guiding interventions and support mechanisms effectively.

- **Superiority of Linear Regression:**

- Among the regression models evaluated, Linear Regression stood out, showcasing superior accuracy compared to other models.
- Its effectiveness in providing reliable predictions, with an accuracy of 88%, establishes its potential utility in student performance forecasting.



Conclusion:

This project focused on leveraging regression algorithms to predict student performance. Through an extensive evaluation encompassing various regression models, Linear Regression emerged as the optimal solution, demonstrating an impressive accuracy rate of 88%. The rigorous assessment of models and the subsequent selection of Linear Regression highlight its superior predictive capabilities, emphasizing its potential significance in guiding educational strategies and interventions for improved student outcomes.

Key Insights:

- Student performance prediction is crucial for educational institutions.
- Linear Regression outperforms other regression models in accuracy.
- With an accuracy of 88%, Linear Regression provides the most reliable predictions.