DATA ANALYSIS

**TOPIC:**

***Student Performance Analysis***

**SECTION:**

**BS INFOTECH – 3B**

**MEMBERS:**

Cabaliza, Mark Nathaniel

Dosado, John Michael

Gementiza, Errol

Romulo, Randy

Lojero, Viosel

Monternel, Jodelyn

Orquilla, Dona lyn

Tarrayo, Bea

**SUBMITTED TO:**

**Ms. Myka Pilpa**

**Instructor**

INTRODUCTION 1

DATA EXPLORATION 2

METHODS 3

MODELLING AND RESULTS 4

CONCLUSION 5

REFERENCES 6

**INTRODUCTION**

In this project, we focused on understanding and predicting student performance in academic settings. We examined various factors that could influence students' grades, such as attendance, study habits, and behavioral patterns. The dataset we utilized includes information on gender, relationship status, educational topics, section, grade, nationality, class, stage ID, semester, parental involvement in surveys, parental satisfaction with the school, and student absence days. By analyzing these factors, we aimed to identify patterns that could help predict whether a student is likely to excel or struggle academically. Understanding these patterns is crucial for educators and policymakers to implement effective interventions and support systems.

**DATA EXPLORATION**

**How The Data Was Collected**

The dataset used in this analysis was collected from a GitHub repository. It contains information about student behavior and academic performance. The dataset was downloaded to a local folder and then imported into Visual Studio for further analysis.

**Features Identified for Analysis**

The key factors considered for this analysis include gender, Nationality, Place of birth, Stage ID, Grade ID, Section ID, Topic, semester, Relation, parents' participation in surveys, parents' satisfaction with the school, and students' absence days. These elements were chosen to examine their influence on various outcomes and provide a deeper understanding of the patterns and trends within the data.

**A screenshot of a computer

Description automatically generated**

**Screenshots of Reports.**

**Figure 1. Gender Feature Visualization**

**A close-up of a computer code

Description automatically generated**

**A graph of a person and person

Description automatically generatedA graph of a number of blue squares

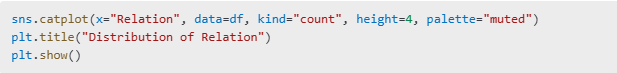
Description automatically generated with medium confidenceA graph of a person and person

Description automatically generated**

**Insights:** The number of males is higher because it could reflect a gender imbalance in school enrollment or cultural factors where boys are more likely to attend school**.**

**Visualization:** The count plot for gender shows the distribution of male and female students in the dataset.

**Figure 2. Relation Feature Visualization**

****

**A blue rectangular bars with white text

Description automatically generated**

**Insights**: Father are more involved in school activities, as they may traditionally take on roles related to their children's education, such as attending meetings or surveys.

**Visualization:** The count plot for Relation highlights the distribution of students' relationships with their parents, categorized as either "Father" or "Mum".

**A close-up of a computer code

Description automatically generatedFigure 3. Educational Topic Visualization**

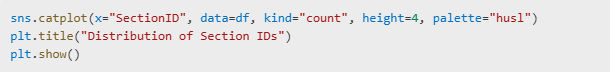
**A graph of a number of people

Description automatically generated with medium confidence**

**Insights:** IT is likely a core subject with mandatory participation, making it appear more frequently compared to electives like French or Arabic.

**Visualization:** The count plot for Topic illustrates the frequency of different educational topics, such as Math, Science, and English, among students.

**Figure 4. Section Feature Visualization**

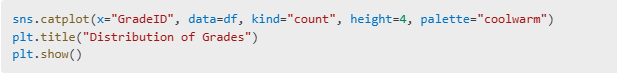
****

A graph of a bar graph

Description automatically generated with medium confidence

**Insights**: Section A might represent the most populated class due to scheduling priorities or a higher intake of students.

**Visualization**: The count plot for Section ID displays the distribution of students across different sections, labeled as A, B, and C.

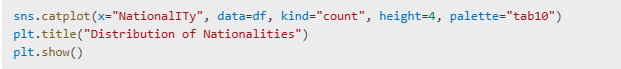
**Figure 5. Grade Feature Visualization**

A graph of a number of numbers and a number of numbers

Description automatically generated with medium confidence

**Insights:** If some Grade IDs have higher counts, it likely indicates a focus on specific grade levels during data collection, such as mid-level grades where academic interventions are frequent.

**Visualization**: The count plot for Grade ID shows the frequency of students in various grade levels, such as G2, G4, and G5.

**Figure. 6 Nationality Feature Visualization **

A graph of a bar graph

Description automatically generated with medium confidence

**Insights:** The KW have the higher frequency in nationality because maybe in the datasets the most nationality in the class is KW

**Visualization**: The count plot for NationalITy reveals the diversity of nationalities in the dataset, highlighting the most common nationalities among students

**Figure 7. Class Feature Visualization**

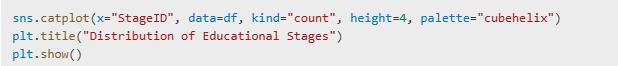


**A graph of a bar

Description automatically generated with medium confidence**

**Insights:** Class" distribution reflects performance categories. If one class has a higher frequency (e.g., "High" performance), it suggests that many students excel, possibly due to better teaching quality, academic focus, or external resources.

**Visualization**: The count plot for Class illustrates the distribution of students across performance categories, such as Low, Medium, and High.

** Figure 8. Educational Stage Feature Visualization**

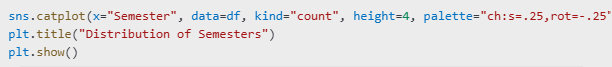
A graph of a number of blue rectangular objects

Description automatically generated with medium confidence

**Insights:** Middle school may have the largest student population in the dataset because this is the age group with the highest school participation.

**Visualization**: The count plot for StageID highlights the spread of students across educational stages, such as Elementary, Middle, and High School.

**Figure 9. Semester Feature Visualization**

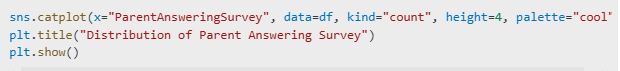
****

**A graph showing a number of classes

Description automatically generated with medium confidence**

**Insights**: higher count in one semester suggests that assessments, surveys, or data collection were more concentrated during that semester, making the data unbalanced.

**Visualization:** The count plot for Semester compares the number of students in the first and second semesters.

**Figure 10. Survey Feature Visualization**

**A graph of a parent and a parent satisfaction

Description automatically generated with medium confidence**

**Insights:** Yes is dominant, it shows strong parental involvementis.

**Visualization:** The count plot for Parent Answering Survey depicts the participation of parents in answering surveys, categorized as Yes or No.

**Figure 11. Parent Satisfaction Feature Visualization**



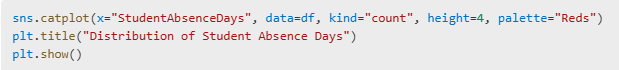
**A graph of a parent and a bad

Description automatically generated with medium confidence**

**Insights:** Good is higher due to the possibility that the parents is satisfied to the school teaching and environment.

**Visualization:** The count plot for ParentschoolSatisfaction shows parent satisfaction levels, categorized as Good or Bad.

**Figure 12. Student Absence Feature Visualization**

****

**A graph of a student absence days

Description automatically generated with medium confidence**

**Insights**: Under-7 is higher, it signifies better attendance and discipline among students

**Visualization:** The count plot for Student Absence Days compares the number of students with absence days under 7 and above 7.

**METHODS**

**Pre-Processing Techniques Used**

The following pre-processing techniques were implemented to prepare the dataset for analysis:

**Data Importation**: The dataset was imported using the **pandas** library, ensuring that all necessary libraries (Pandas, NumPy, Seaborn, Matplotlib) were installed for data analysis and visualization. The dataset was loaded using the **read\_csv** function from Pandas, which is essential for handling CSV files.

A white background with black dots

Description automatically generated

**Data Cleaning**: We conducted a thorough check for null values and inconsistencies within the dataset. Python code was written to identify and handle any missing data, ensuring the integrity of the dataset.



Data cleaning is performed here to handle missing values by filling numerical columns with their mean, remove duplicate rows to avoid redundancy, and drop irrelevant columns that do not contribute to the analysis. Additionally, it ensures consistency in categorical variables by checking for unique values, helping to identify and resolve potential issues. These steps are essential to ensure the dataset is accurate, consistent, and ready for reliable analysis.

A screenshot of a computer program

Description automatically generated**Data Normalization**: To ensure that all features were on a comparable scale, we applied scaling techniques, which is crucial for many machine learning algorithms.

We choose normalization in our code to ensure that all features contribute equally to the model, especially for algorithms sensitive to the scale of input data, such as K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Lasso Regression. These models can be disproportionately influenced by features with larger ranges, which can lead to biased results or slower convergence. Normalization (scaling to have a mean of 0 and standard deviation of 1) ensures that all features are on the same scale, preventing any one feature from dominating the model. This improves algorithm performance, particularly in models that rely on distance calculations or optimization techniques.

**REFERENCE**

Watty12. (n.d.). *Student Performance Analysis*. GitHub. <https://github.com/watty12/Student-Performance-Analysis>

Dammonoit, A. (2018). *Student performance analysis using Big Data* [Repository]. GitHub. <https://github.com/Dammonoit/Student-performance-analysis-using-Big-data/tree/master>