DATA ANALYSIS

**TOPIC:**

***Student Performance Analaysis***

**SECTION:**

**BS INFOTECH – 3B**

**MEMBERS:**

Cabaliza, Mark Nathaniel

Dosado, John Michael

Gementiza, Errol

Romulo, Randy

Logero, 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**

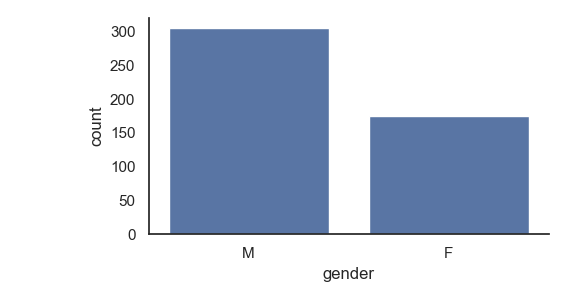
The key factors considered for this analysis include gender, relation, topic, section, grade, nationality, class, stage ID, semester, parents' participation in surveys, parents' satisfaction with the school, and students' absence days. These elements were chosen to examine their influence on various outcomes and provide a deeper understanding of the patterns and trends within the data.

**Screenshots of Reports.**

**Figure 1. Gender Feature Visualization**

**A close-up of a computer code

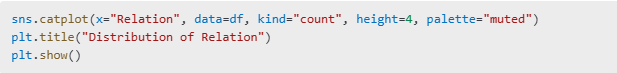
Description automatically generated**

**A graph of a person and person

Description automatically generated**

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

**Figure 2. Relation Feature Visualization**

****

**A graph of a number of people

Description automatically generated with medium confidence**

**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

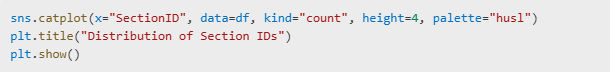
Description automatically generatedFigure 3. Educational Topic Visualization**

**A graph with blue rectangles

Description automatically generated**

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

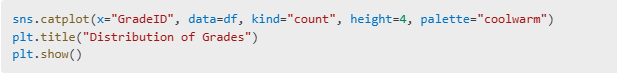
**Figure 4. Section Feature Visualization**

****

A graph of a bar graph

Description automatically generated with medium confidence

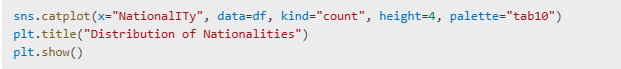
**Visualization**: The count plot for SectionID displays the distribution of students across different sections, labeled as A, B, and C.

**Figure 5. Grade Feature Visualization** 

A graph of a number of bars

Description automatically generated with medium confidence

**Visualization**: The count plot for GradeID shows the frequency of students in various grade levels, such as G1, G2, and G3.

**Figure. 6 Nationality Feature Visualization **

A graph of a bar graph

Description automatically generated with medium confidence

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

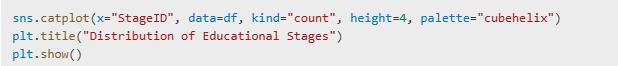
**Figure 7. Class Feature Visualization**



**A graph of a bar

Description automatically generated with medium confidence**

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

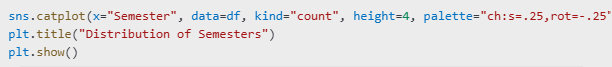
** Figure 8. Educational Stage Feature Visualization**

A graph of different levels of performance

Description automatically generated with medium confidence

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

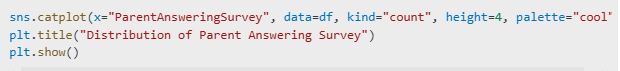
**Figure 9. Semester Feature Visualization**

****

**A blue rectangular object with white text

Description automatically generated**

**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 bar

Description automatically generated with medium confidence**

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

**Figure 11. Parent Satisfaction Feature Visualization**



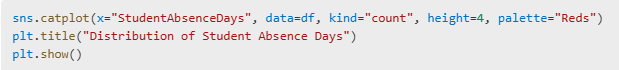
**A graph of a bar graph

Description automatically generated with medium confidence**

**Figure 11. Parent Satisfaction Feature Visualization**

**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 bar graph

Description automatically generated with medium confidence**

**Visualization:** The count plot for StudentAbsenceDays 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.

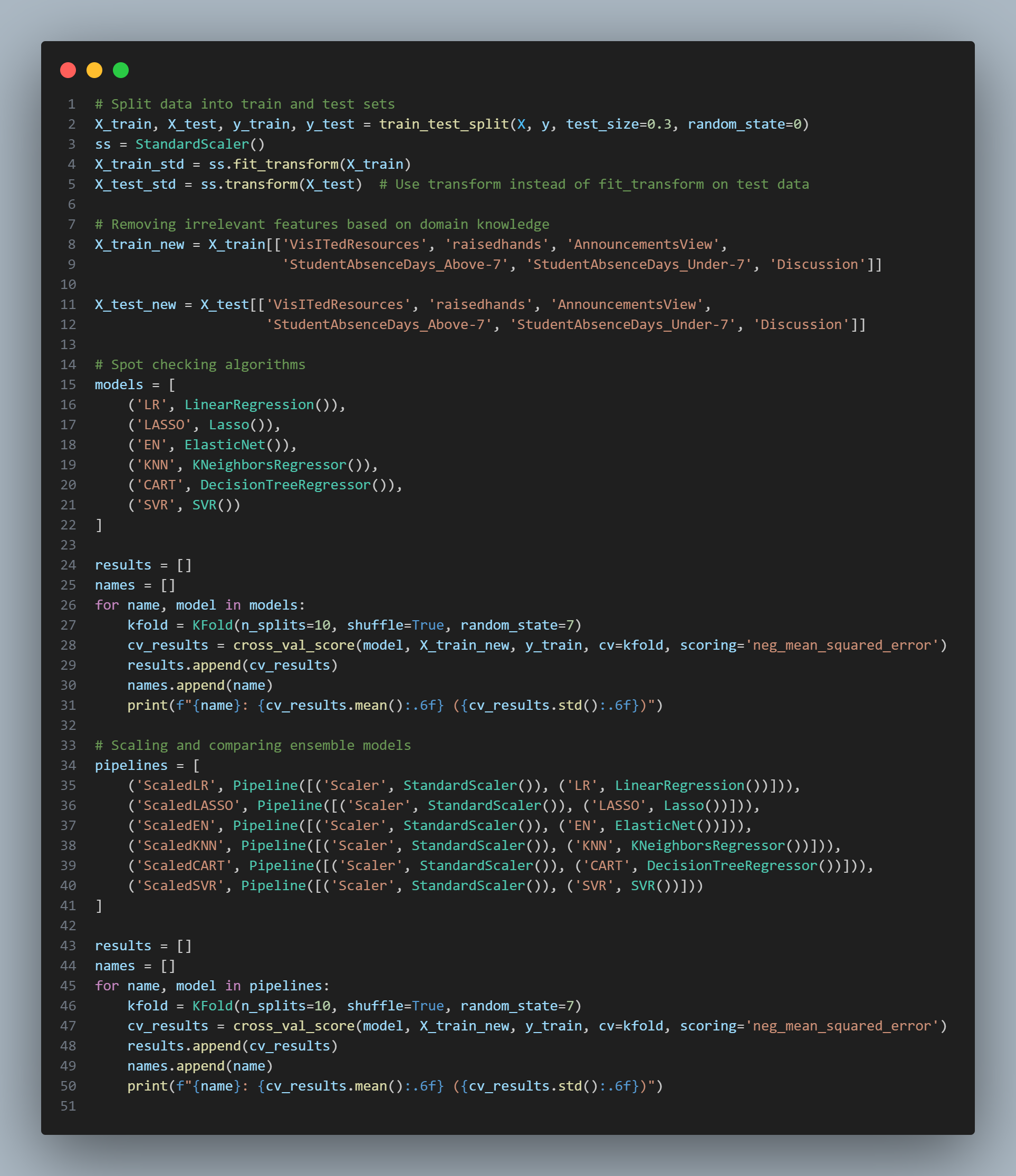
**Data Cleaning**: We conducted a thorough check for null values and inconsistencies within the dataset. Python code was written to identify and handle any missing data, ensuring the integrity of the dataset.

**Feature Selection**: We employed information gain techniques to identify the most significant features affecting student performance. Behavioral features, such as engagement metrics, were ranked highest in their predictive power.

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

**CODE SCRIPT:**

****

****

**A screen shot of a computer program

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

**REFERENCE**

1. Here is the citation in APA format for the given Kaggle dataset:

Aljarah, I. (2020). *xAPI Edu Data* [Data set]. Kaggle. <https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data>

1. Here is the citation in APA format for the given GitHub repository:

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