Analysis of Student Activities and Performance

This presentation explores the relationship between student activities and academic performance, analyzing how various extracurricular engagements influence educational outcomes through multiple regression analysis

Group Members

- Muneeb Ashraf
- Saad Ahmed
- M. Danial Qamar
- Irfan Ahmad
- Mohsin Raza

Project Overview

This project analyze how student activities affect their academic performance using a dataset from Kaggle. We use multiple regression analysis to see how different extracurricular activities and other factors impact students' grades. We prepare the data by cleaning it up, converting categories into numbers, and using statistical modeling. Our goal is to find important relationships and make predictions about student performance. We also check how accurate our model is by using metrics like R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE), and we create visualizations to help us understand the results better.

Methodology

Data Collection

The dataset was obtained from Kaggle.

Preprocessing

Checked for null values in the dataset and Converted categorical data into numerical format suitable for regression.

Model Selection

A multiple regression model was fitted to the training data.

FLOWCHART

START

IMPORT LIBRARIES (NUMPY, PANDAS, SEABORN etc).

IMPORT DATASET (Student_Data.csv)

Data Cleaning (Drop unnecessary columns)

Data Exploration (info(), isnull().sum(), corr())

Label Encoding (Extracurricular_Activities column)

MODELING (MULTIPLE REGRESSION)

PREDICTION

VISUALIZATION (SCATTER PLOT, LINE PLOT)

EXECUTE

Data Collection and Preprocessing

Source:

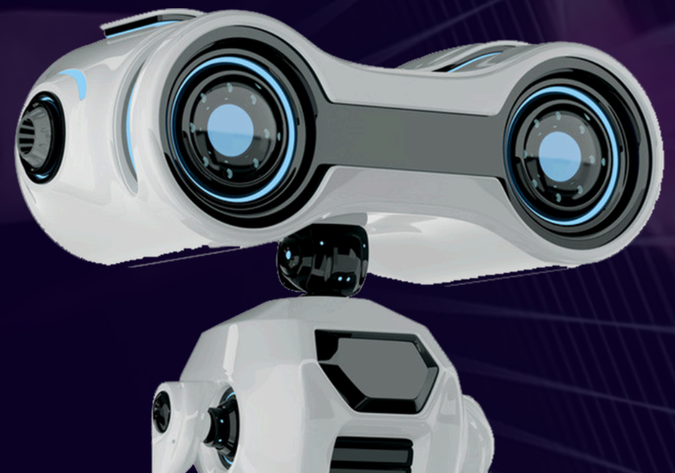
- The dataset was obtained from Kaggle.

Content:

- It includes features such as Studying Hours, Sleep Hours, Sample Question Papers Practiced, Extracurricular Activities, Previous Scores, Math Score, Reading Score, Writing Score. While Target variable is Performance Index.

Size:

- The dataset consists of 1000 entries and 8 features.



Data Preprocessing

Handling Missing Values:

- Checked for null values in the dataset using isnull().sum() and addressed them as needed.

Label Encoding:

- Purpose: Converted categorical data into numerical format suitable for regression.
- Implementation: The 'Extracurricular_Activities' column was encoded using LabelEncoder from sklearn.
- Outcome: The categorical variables were successfully transformed, enabling their use in regression analysis.

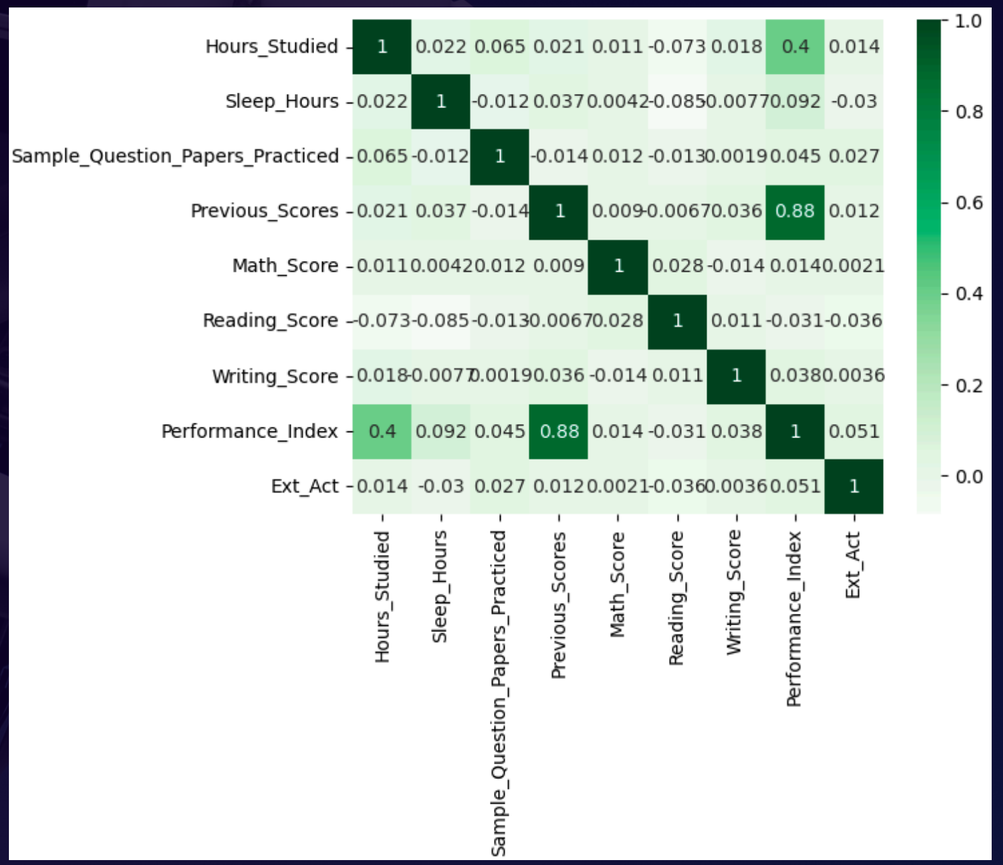
Data Analysis

Dataset Information:

- The info() method was used to understand the structure of the dataset, including data types and non-null counts.

Correlation Analysis:

- To identify relationships between variables and select predictors for the regression model.
- The correlation matrix highlighted significant correlations that were used to inform the regression model.



09

Multiple Regression Modeling

- **y=mx+b**
- **In multiple regression, we have multible values of x**
- **m is slope and b is y-intercept**

Objective:

- To predict academic performance based on multiple predictors.

Steps:

- The data was split into training and testing sets.
- A multiple regression model was fitted to the training data.

Evaluation Metrics

R-squared:

- Indicates the proportion of variance in the dependent variable explained by the independent variables.

MAE (Mean Absolute Error):

- Measures the average of the absolute differences between the actual and predicted values.

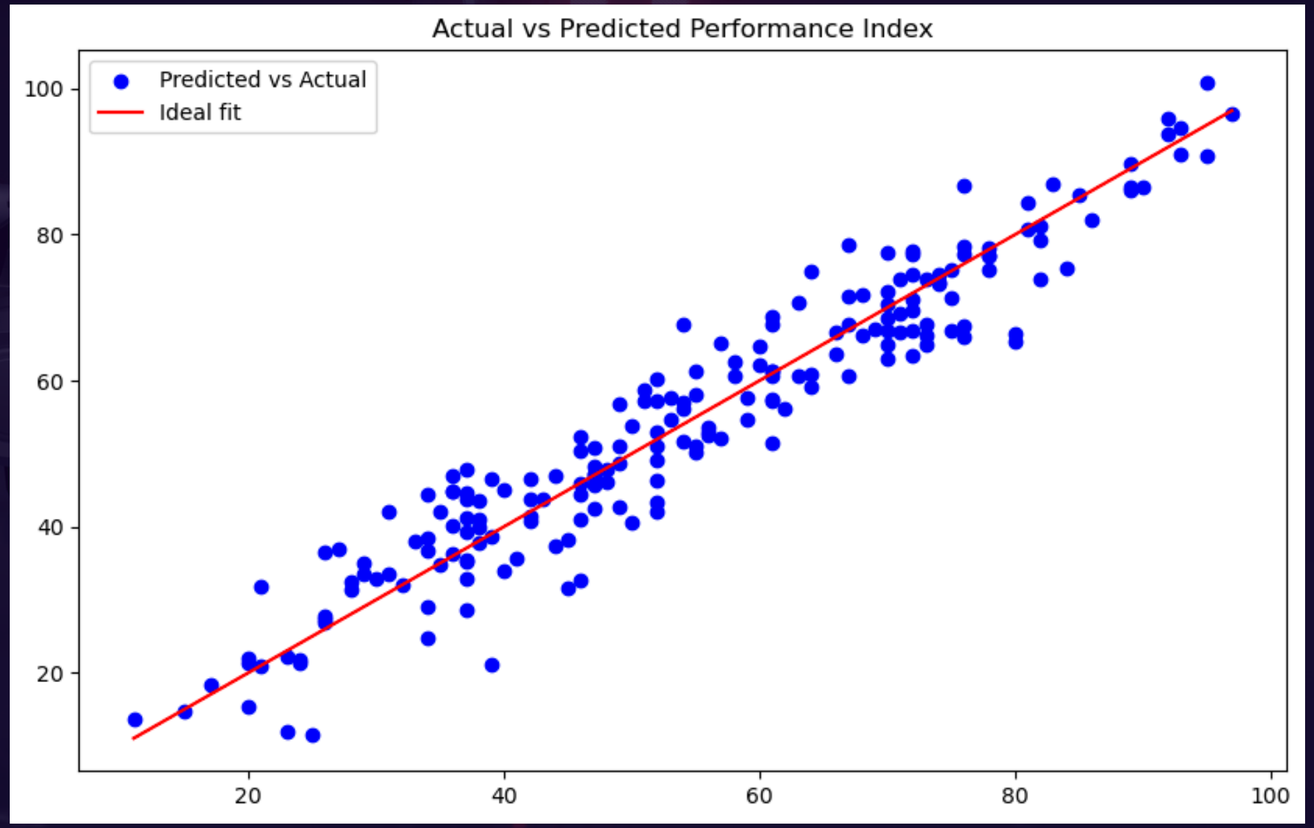
MSE (Mean Squared Error):

- Measures the average of the squares of the errors—that is, the difference between the actual and predicted values.

Visualization of Actual vs. Predicted Data

The purpose of this plot is to visually compare the actual values of the target variable with the predicted values from the regression model:

- **Scatter Points:** If the model performs well, the scatter points will be close to the red ideal fit line (y = x).
- **Evaluation:** This visualization helps in assessing the model's performance. A dense clustering of points around the ideal fit line indicates good predictive power, while a scatter away from the line suggests inaccuracies in predictions.



Results

The project successfully applied multiple regression to predict student performance, with key metrics like R-squared, MAE, and MSE used to evaluate the model.

- The multiple regression model provided an R-squared value of 0.916612981794711 7. An R^2 value close to 1 indicates that the model explains a large portion of the variance in the dependent variable, whereas an R^2 close to 0 indicates that the model does not explain much of the variance.
- The MAE is 4.378. MAE directly measures the average error without squaring it, so the units of MAE are the same as the original data. It treats all errors equally, whether they are large or small
- MSE value is 5.645. MSE makes sure that larger errors are penalized more, meaning that it gives more weight to larger mistakes.

Thank you for your attention...

We appreciate your interest in our project.....

We welcome any questions or feedback you may have..!