MINI PROJECT

PREDICTING STUDENT PERFORMANCE USING MULTIPLE REGRESSION ANALYSIS

Submitted by

KAMANA PHANINDRA REDDY(BU22CSEN0101454)

SET-A

SUBJECT: Artificial Intelligence And Applications

Table of Contents:

1. Abstract

This project addresses the growing need for predictive models in educational environments to assess student performance. Predicting student outcomes based on various academic and lifestyle factors can help educational institutions tailor their interventions to maximize student success.

This project utilizes a machine learning approach to forecast a student's **Performance Index**—a holistic measure of academic achievement—based on features such as hours studied, previous exam scores, participation in extracurricular activities, sleep patterns, and the number of sample question papers practiced.

The primary objective is to build a machine learning model capable of accurately predicting performance while identifying key contributors to student success. A **Linear Regression** model was chosen due to its effectiveness in handling continuous target variables and its interpretability.The project includes the full pipeline of data preparation, model training, and evaluation using standard metrics such as the **R^2 score**, **Mean Squared Error (MSE)**, and **Mean Absolute Error (MAE)**. The results demonstrate the model's strong predictive power, with an R^2 score of 0.988, indicating that the model explains 98.8% of the variance in student performance.

This paper discusses the problem statement, preprocessing steps applied to clean and standardize the data, and the development of the model. We also highlight the practical implications of predicting student outcomes based on lifestyle and academic factors.

The findings suggest that certain behaviors, like studying more hours and completing more sample question papers, are strong predictors of academic success. Future research can explore other machine learning algorithms and expand the dataset for improved accuracy.

2. Introduction

Machine learning has revolutionized various industries, offering tools for data-driven decision-making. In education, predictive models are gaining popularity for their ability to provide insights into student behavior and performance.

These models help educators identify at-risk students early, personalize learning experiences, and optimize teaching methods. Predicting student academic performance is one such application where machine learning models analyze diverse factors to estimate future success.

In this project, we focus on creating a machine learning model to predict a student's **Performance Index**.

This score is a composite indicator that takes into account academic achievement, lifestyle habits, and preparation strategies. The objective is to develop a predictive system that uses

readily available student data, such as study hours, prior academic performance, and sleep patterns, to forecast outcomes.

This approach can help institutions by providing them with actionable insights into the factors that most significantly impact student success.

Understanding these factors can allow for personalized interventions, where struggling students receive more targeted support and attention.

Additionally, high-performing students can be encouraged to maintain or improve their habits to maximize their potential.

This paper explores the data used, preprocessing techniques applied, and the step-by-step construction of the predictive model. The evaluation of the model's performance based on a testing set ensures that the results are not only accurate but also generalizable to unseen data.

3. Problem Definition

The academic success of students is a multifaceted issue influenced by a variety of factors, both within and outside the classroom. Educational institutions continually strive to identify those factors that most strongly correlate with student performance in order to implement targeted interventions that can improve outcomes.

This project aims to address the challenge of predicting a student's academic performance, which we represent as the **Performance Index**. The Performance Index is a numerical score that summarizes a student's overall achievement based on their grades, study habits, and personal lifestyle choices.

The central question this project seeks to answer is: **Can we predict a student's academic performance based on data related to their study habits, extracurricular participation, and lifestyle choices?** More specifically, we aim to determine whether factors like the number of hours studied per day, previous academic performance, the number of sample question papers practiced, and hours of sleep per night can serve as reliable predictors for overall academic achievement.

Key objectives of this project include:

1. Developing a machine learning model that can accurately predict the Performance Index based on input features.
2. Identifying which factors have the most significant impact on academic performance.

3. Evaluating the model's performance using standard metrics to ensure robustness and reliability.

4. Data Source

The success of any machine learning model largely depends on the quality and relevance of the data it is trained on. For this project, we collected a dataset containing information about students, including their study habits, previous academic performance, participation in extracurricular activities, and lifestyle choices such as sleep hours. Each row in the dataset represents a student, and the columns correspond to different variables that could potentially influence their performance.

Data Set Link: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiplelinear-regression

The dataset contains the following key features:

1. **Hours Studied**: This feature captures the average number of hours a student spends studying per day. It is a critical indicator of how much time and effort the student devotes to their academic work.
2. **Previous Scores**: This feature records the student's academic performance in the previous semester. It serves as a benchmark to assess whether past success is a strong predictor of future performance.
3. **Extracurricular Activities**: This binary feature indicates whether the student participates in extracurricular activities, such as sports, arts, or clubs.
4. **Sleep Hours**: This feature captures the average number of hours of sleep the student gets per night.
5. **Sample Question Papers Practiced**: This feature measures the number of sample question papers the student has solved in preparation for their exams.
6. **Performance Index**: This is the target variable, representing the overall academic performance of the student.

5. Data Preprocessing

Data preprocessing is a crucial step to ensure that the machine learning model is trained on clean and well-structured data. The raw dataset contained some missing values and categorical features that needed to be converted into numerical representations. Below are the key preprocessing steps applied:

1. **Handling Missing Values**: Missing values in the dataset were handled using imputation techniques. For numerical features like hours studied or previous scores, the mean of the column was used to fill in missing values.

```
[47]: df.isna().sum()
```

```
[47]: Hours Studied                      0
      Previous Scores                   0
      Extracurricular Activities        0
      Sleep Hours                       0
      Sample Question Papers Practiced  0
      Performance Index                 0
      dtype: int64
```

2. **Encoding Categorical Features**: The binary categorical feature, extracurricular activities, was encoded into numerical values (0 for no participation, 1 for participation).

```
[31]: #Machine learning
      label_encoder = LabelEncoder()

      df['Extracurricular Activities'] = label_encoder.fit_transform(df['Extracurricular Activities'])
```

3. **Feature Scaling**: Since features like study hours and sleep hours have different units, we scaled the data using standardization to ensure that all features contributed equally to the model.

The final preprocessed dataset was ready for model training.

```
[36]: scaler = StandardScaler()
      X_scaled = scaler.fit_transform(X.drop(["Extracurricular Activities"] ,axis=1))
      scaler_y = StandardScaler()
      y_scaled = scaler_y.fit_transform(y.values.reshape(-1, 1))
```

6. Model Creation and Training

We selected **Linear Regression** as the algorithm for this project due to its simplicity and effectiveness in handling continuous target variables. The model was implemented using the following steps:

Code Example - Linear Regression from

sklearn.model_selection import train_test_split from

sklearn.linear_model import LinearRegression

Splitting the data into training and testing sets

X = dataset[['Hours_Studied', 'Previous_Scores', 'Extracurricular', 'Sleep_Hours', 'Sample_Papers']] y =

dataset['Performance_Index']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

Creating the Linear Regression model

model = LinearRegression() model.fit(X_train,

y_train)

Making predictions on the test set y_pred

= model.predict(X_test)

```
[38]: X_train, X_test, y_train, y_test = train_test_split(X_combined, y_scaled, test_size=0.2, random_state=42)

[39]: model = LinearRegression()
      model.fit(X_train, y_train)

[39]: LinearRegression
      LinearRegression()

[40]: y_train_pred = model.predict(X_train)
      y_pred = model.predict(X_test)
```

7. Evaluation Metrics

The performance of the Linear Regression model was evaluated using three metrics:

1. **R^2 Score**: Measures how well the regression model captures the variability in the data.
2. **Mean Squared Error (MSE)**: Indicates the average squared difference between predicted and actual values.
3. **Mean Absolute Error (MAE)**: Represents the average absolute difference between predicted and actual values.

from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

Evaluation metrics r2 = r2_score(y_test,

y_pred) mse = mean_squared_error(y_test,

y_pred) mae = mean_absolute_error(y_test,

y_pred)

print(f'R² Score: {r2}')

print(f'MSE: {mse}') print(f'MAE:

{mae}')

```
[41]: r2 = r2_score(y_test, y_pred)
      print(f"R² Score: {r2}")

      R² Score: 0.9884301209927054

[44]: mse = mean_squared_error(y_test, y_pred)
      mae = mean_absolute_error(y_test, y_pred)
      print(f"Mean Absolute Error (MAE): {mae}")
      print(f"Mean Squared Error(MSE): {mse}")

      Mean Absolute Error (MAE): 0.08574577950768571
      Mean Squared Error(MSE): 0.011671265615520255
```

8. Results and Analysis

The Linear Regression model achieved an **R^2 score of 0.988**, indicating that 98.8% of the variance in student performance was captured by the model. The MSE and MAE values were low, demonstrating the accuracy of the predictions.

Example:

Hours_Studied=7,Previous_Scores=99, Extracurricular=0, Sleep_Hours=9, Sample_Papers=1

```
[47]: input_data = np.array([[7, 99, 0, 9, 1]])

      y_pred = model.predict(input_data)

      print("Predicted Performance Index:", y_pred[0])

      Predicted Performance Index: [93.89895519]
```

9. Conclusion

In this project, we built a machine learning model to predict student performance based on several important factors, such as hours spent studying, previous academic scores, involvement in extracurricular activities, sleep patterns, and the number of practice papers completed. By using linear regression, we aimed to find how these factors relate to the performance index.

After splitting the data into training and testing sets, we trained the model and evaluated it using metrics like R^2 score, mean squared error (MSE), and mean absolute error (MAE). The results indicated that the model performs well in predicting student performance, but there is room for improvement. More advanced algorithms or incorporating additional features could lead to better accuracy.

This project highlights how machine learning can provide meaningful insights into education, helping us understand the key factors that influence student outcomes. With further refinement and exploration, models like this could be useful in designing more personalized learning strategies and support systems for students, ultimately enhancing their academic success.