Student Performance Index Predictor Model

Introduction

In the educational sector, predicting student performance is a critical task that can help identify areas where students may need additional support. This project involves developing a linear regression machine learning model to predict the performance index of students based on various factors such as study hours, sleep hours, extracurricular activities, previous score, and the number of sample question papers practiced. This report details the methodology, analysis, and results of the project.

Problem Definition and Algorithm Description

Problem Definition

The primary goal of this project is to create a predictive model that can accurately forecast a student's performance index. The performance index is influenced by multiple factors, including:

- **Study hours**
- **Sleep hours**
- **Extracurricular activities**
- **Previous scores**
- **Number of sample question papers practiced**

Algorithm Description

Linear regression is employed as the predictive algorithm for this project. Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more

independent variables by fitting a linear equation to observed data. The model's performance is evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R^2) score.

Methodology

Data Description

The dataset contains 10,000 records with the following features:

- **Study_hours: Number of hours spent studying**
- **Sleep_hours: Number of hours spent sleeping**
- **Extracurricular_activities: Participation in extracurricular activities (binary variable)**
- **Previous_score: Previous academic score**
- **Sample_papers_practiced: Number of sample question papers practiced**
- **Performance_index: The target variable representing the student's performance index**

Libraries Used

The following Python libraries were used in this project:

- **pandas: For data manipulation and analysis**
- **numpy: For numerical operations**
- **matplotlib.pyplot: For data visualization**
- **seaborn: For advanced data visualization**
- **sklearn.model_selection: For splitting the data into training and testing sets**
- **sklearn.linear_model: For building the linear regression model**

- sklearn.preprocessing: For normalizing the data

- sklearn.metrics: For evaluating the model's performance

Data Preprocessing

1. Boxplot Visualization for Outliers:

Boxplots were used to identify outliers in the dataset.

2. Visualizing Categorical Data:

Categorical data was visualized to understand its distribution.

3. Correlation Analysis:

The correlation among different factors was calculated and visualized using a heatmap.

4. Label Encoding:

LabelEncoder was used to convert categorical data into numerical data.

5. Defining Response and Predictor Variables:

The response variable (Performance_index) and predictor variables were defined.

6. Train-Test Split:

The dataset was split into training and testing sets, with 25% of the data allocated for testing.

7. Standardizing the Data:

Standardization was performed using StandardScaler.

Model Training

A linear regression model was trained using the scaled training data.

Model Evaluation

The model's performance was evaluated using Mean Squared Error (MSE) and R-squared (R^2) score.

Results:

- Mean Squared Error (MSE): 4.637
- R^2 Score: 0.9874

Visualization

1. Actual vs Predicted Values:
2. Residuals vs Predicted Values:

Summary Table

A summary table was created including the intercept, coefficients, MSE, and R^2 score.

Feature Coefficients Plot

The feature coefficient values were plotted.

Conclusion

The linear regression model successfully predicted the performance index of students with a high degree of accuracy, as indicated by an R^2 score of 0.9874. The Mean Squared Error was found to be 4.637. The model's performance was visualized through various plots, providing insights into the relationship between the predictors and the target variable.

This project demonstrates the effective use of data preprocessing, visualization, and machine learning techniques to solve a real-world problem in the educational sector. The methodology and results can be extended to similar predictive modeling tasks in other domains.