**Student Performance Analysis Report**

**Introduction**

This report presents an analysis of a dataset focused on predicting student performance based on various factors such as hours studied, previous scores, extracurricular activities, sleep hours, and sample question papers practiced. The main objective of this analysis is to identify the key factors that influence student performance and build a regression model to predict the performance index of students.

**Objective**

The primary goal of this project is to develop a predictive model using regression techniques that can accurately estimate a student's performance index based on various input features. This analysis aims to assist educational institutions in identifying critical areas to improve student performance.

**Dataset Description**

The dataset used in this analysis is titled **"Student\_Performance.csv"**, containing 10,000 records with six columns:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Hours Studied | Number of hours a student studies daily |
| Previous Scores | Previous academic scores of the student |
| Extracurricular Activities | Participation in extracurricular activities (Yes/No) |
| Sleep Hours | Average hours of sleep a student gets daily |
| Sample Question Papers Practiced | Number of sample question papers practiced |
| Performance Index | Target variable representing student performance |

The target variable is **Performance Index**, which ranges from **10** to **100**.

**Exploratory Data Analysis (EDA)**

**1. Dataset Overview**

* The dataset contains **10,000 records** and **6 columns**.
* There are **no missing values** in the dataset.

**2. Summary Statistics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Mean** | **Std. Dev** | **Min** | **Max** |
| Hours Studied | 4.99 | 2.59 | 1 | 9 |
| Previous Scores | 69.45 | 17.34 | 40 | 99 |
| Sleep Hours | 6.53 | 1.70 | 2 | 9 |
| Sample Question Papers Practiced | 4.58 | 2.87 | 0 | 9 |
| Performance Index | 55.22 | 19.21 | 10 | 100 |

**3. Correlation Analysis**

The correlation matrix reveals key relationships between features:

* **Previous Scores** has the highest correlation with **Performance Index** (0.92).
* **Hours Studied** also shows a moderate positive correlation with **Performance Index** (0.37).
* **Sleep Hours** and **Sample Question Papers Practiced** show weak correlations with **Performance Index**.

The heatmap visualization of the correlation matrix confirmed these relationships.

**Model Selection and Training**

**Models Trained**

Several regression models were trained to predict the **Performance Index**:

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Absolute Error (MAE)** | **Root Mean Squared Error (RMSE)** |
| Linear Regression | **1.7453** | **2.1949** |
| Ridge Regression | 1.7453 | 2.1953 |
| LASSO Regression | 16.2039 | 19.2105 |
| Support Vector Machine | 1.7456 | 2.1951 |
| Decision Tree Regressor | 2.9900 | 3.7055 |
| Random Forest Regressor | 2.0295 | 2.5583 |
| CatBoost Regressor | 1.7976 | 2.2621 |

**Best Model: Linear Regression**

The **Linear Regression** model was selected as the best model based on its lowest **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** scores.

**Model Evaluation Metrics**

* **R² Score**: 0.9871
* **Adjusted R² Score**: 0.9871

These metrics indicate that the Linear Regression model explains approximately **98.71%** of the variance in the target variable and provides an accuracy of **96.84%**.

**Key Findings and Insights**

1. **Previous Scores** is the most significant predictor of student performance, with a correlation of **0.92**.
2. **Hours Studied** also plays an important role, showing a moderate correlation of **0.37** with the performance index.
3. **Sleep Hours** and **Sample Question Papers Practiced** have minimal impact on performance.
4. The **Linear Regression model** was the best-performing model, achieving an **R² score of 0.9871** and an **Adjusted R² score of 0.9871**.

**Suggestions for Improvement**

1. **Collect additional data** on other factors such as class participation, teacher feedback, and learning environment, which may further improve the model’s performance.
2. **Apply hyperparameter tuning** on other models such as Random Forest or CatBoost to see if their performance can be improved.
3. **Experiment with interaction terms** in the Linear Regression model to capture more complex relationships between features.

**Conclusion**

The analysis shows that **Previous Scores** and **Hours Studied** are the most important factors in predicting student performance. The **Linear Regression model** was the most accurate model, explaining **98.71% of the variance** in the performance index. The insights gained from this project can help educational institutions better understand the factors affecting student performance and take appropriate measures to improve it.