Predicting Student Performance Analysis

- using Linear Regression

Data set : Students data

Source of the data : KAGGLE

Libraries : numpy, Pandas, Sciket-learn, pandas, seaborn

Model : Linear Regression, Supervised learning.

https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression/code

Objective :

The objective of the study is Predicting the Student Performance index.

The Student Performance Index contains, which is a real number. So the predicted variable is continuous. We will predict the values by using Linear Regression algorithm under Supervised Learning method.

It typically involves finding the line of best fit for a given set of data points, estimating how much variation can be explained by that model, and predicting the value of one variable (known as the dependent variable) based on the values of others (known as independent variables).

Introduction:

Definition:
Multiple Linear Regression is a statistical technique used to predict an outcome based on multiple independent variables. It extends Simple Linear Regression by considering more than one predictor to improve accuracy.

1.. Steps to Implement MLR:

- Feature Extraction
  Formulate the problem, which we wanted to solve, what we want to predict.
- Data Preprocessing
  - EDA
  - Missing Value Treatment
  - Duplicate rows detection
  - Outliers detection & Treatment
  - Univariate analysis
  - Bivariate analysis
- Feature Engineering
- Encoding
- Scaling
- Feature Selection
- Splitting the data into training & testing
- Model Building
  - Model
  - Fit the model

- Make predictions on training data set
- Fit the model to testing data set
- Make predictions on testing data set

- Model Evaluation
- Model Deployment

1. Feature Extraction:

Extracted data from Kaggle, which is Student performance data.

2. Data Preprocessing:

- Which, is crucial step in my model.
- The data set size – [10000,6]
  The data set contains 10000 rows and 6 columns
  The column names are
  Hours studied,
  Previous Scores,
  Sleep Hours,
  Sample Question Papers Practiced,
  Performance Index,
  Extra-Curricular Activities
- Checking data set columns
  Here, most of the column names contains spaces, I replaced them with underscore(_).
- Performed basic descriptive analysis
  This helped me about the data, which contains
  Min, Max, Count, standard deviation, 25%, 50% 75 %
- Checking Discrepancies
  There is no discrepancies
- Missing values
  The dataset has no missing values
- Checking for duplicates
  The date set contains 127 duplicated records. I removed those 127 duplicated records
  The data set contains 9873 rows and 6 columns after removing the duplicated records
- Univariate Analysis
  Plotted Histogram for to understand the count of the values,
  Plotted Boxplot for identifying the outliers
  There is no outliers present in the data set.
- Bivariate Analysis
  Constructed Heatmap to find the relationship between the numerical values

3. Encoding

In the dataset, Extra Curricular Activities column contain, categorical values, which are Yes, No
For Machine Learning Model Building, we need to convert those text related items or attributes into numerical ones.
For this column I chosen, LabelEncoder

4. Scaling:

Which is used to arrange the values into one scale, to improve the model performance,
Here I used, StandardScaler() method from sklearn.preprocessing.

5.Feature Selection:

Features selection helps to identify the most relevant features for improving model performance.
There are different techniques available in feature selection

1**. SelectKBest:**

It uses, statistical test like Chi-Squared test, ANOVA, F-test to score and rank the features based on their relationship with the target variable.
It selects the K best features with the highest scores to be included in the final feature subset.

2. RFE (Recursive Feature Elimination):

It is a feature selection technique that iteratively removes the least important features to identify the most relevant ones.

6. Checking Multicollinearity

Multicollinearity refers to the collinearity exists among the independent features, which leads our model to overfitting.
Technique: VIF – Variance Inflation Factor.
There is no Multicollinearity exists in the dataset

7. Split the data

Splitting the data into training & testing to build the model and checking the model performance.
I allocated 20 percent of the data for training & 80% of the data for testing.

8. Model Building:

Constructed the Linear Model, using sciket learn library.

- Import the model
- Fit the model to train data set
- Make predictions
- Evaluate the model performance using evaluation metrics
- Transform the model to testing dataset
- Make predictions on test dataset
- Evaluate the performance for test dataset
-

9. Checking for Assumption:

Our model, residual points follow normality. This will indicate our model goodness.

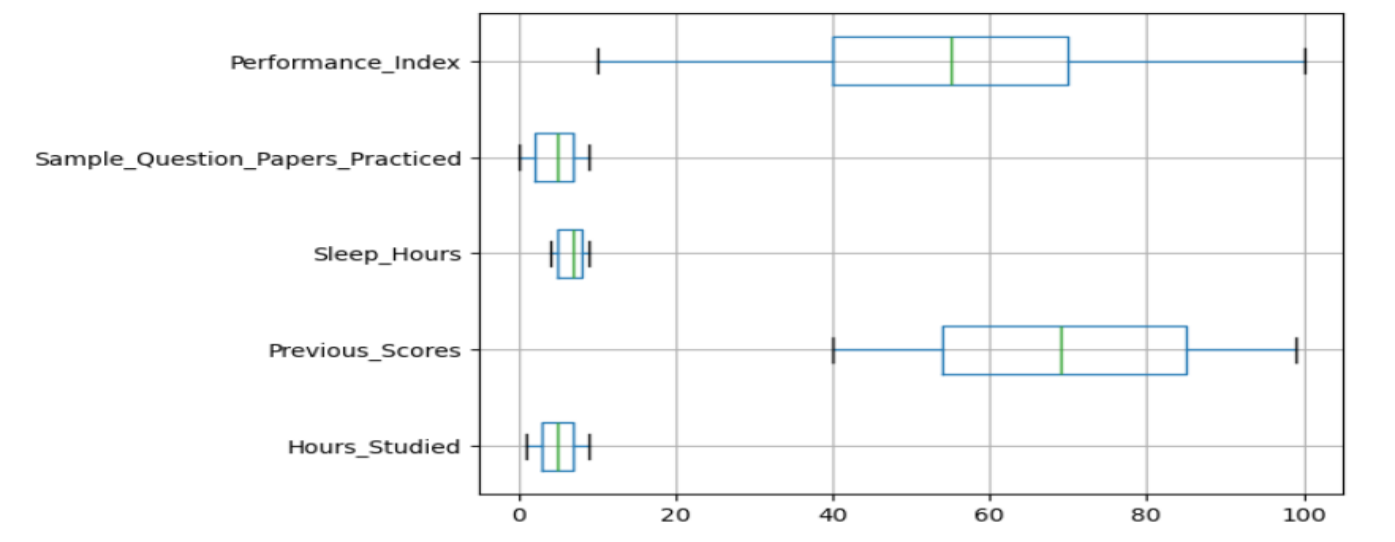
10. Conclusion:

R2_score for training data : 0.98872 which means 98%
MSE for train data : 4.1646
RMSE for train data: 2.0407
R2_score for test data : 0.9885 which means 98%
MSE for test data : 4.2137
RMSE for test data: 2.0527

Our model explains 98% of the variance, meaning it captures key patterns in the data.

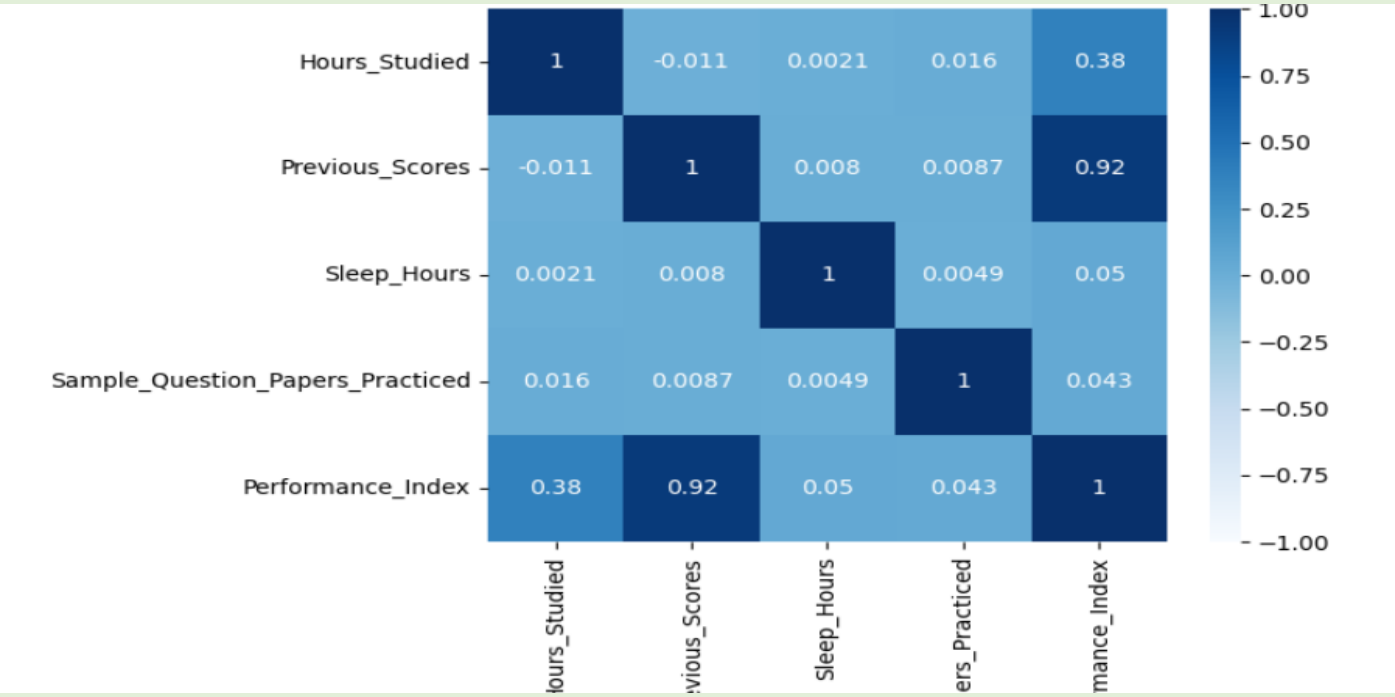
Visualizatons:

The below Boxplot explains there is no outliers in the dataset for each column.



HeatMap :

The below heatmap explain the correlation among variables/features. We observe, which variables are highly correlated and low correlated.



Feature Selection based on SelectKBest :

	Feature	score
0	Hours_Studied	1618.585346
1	Previous_Scores	50863.230088
2	Extracurricular_Activities	6.715704
3	Sleep_Hours	25.090267
4	Sample_Question_Papers_Practiced	18.658435

From the above tables, which is drawn from SelectKBest feature selection technique, which tells, which features has got high score, we only include those features in model building. The Hours_studied, Previous_Scores, Sleep_hours are having high scores.

Feature Selection Based on RFE:

	Feature	Ranking
0	Hours_Studied	1
1	Previous_Scores	1
2	Extracurricular_Activities	2
3	Sleep_Hours	1
4	Sample_Question_Papers_Practiced	1

The table is drawn from the RFE(Recursive Feature Elimination method). Which will give rankings based on the importance of the features. Here we only keep low rankings, almost all features got low rankings, only one feature got 2nd rank. We will build model including all features.

Checking Multi-Collinearity among independent variables:

```
Hours_Studied VIF =  1.0
Previous_Scores VIF =  1.0
Extracurricular_Activities VIF =  1.0
Sleep_Hours VIF =  1.0
Sample_Question_Papers_Practiced VIF =  1.0
```

Making Predictions

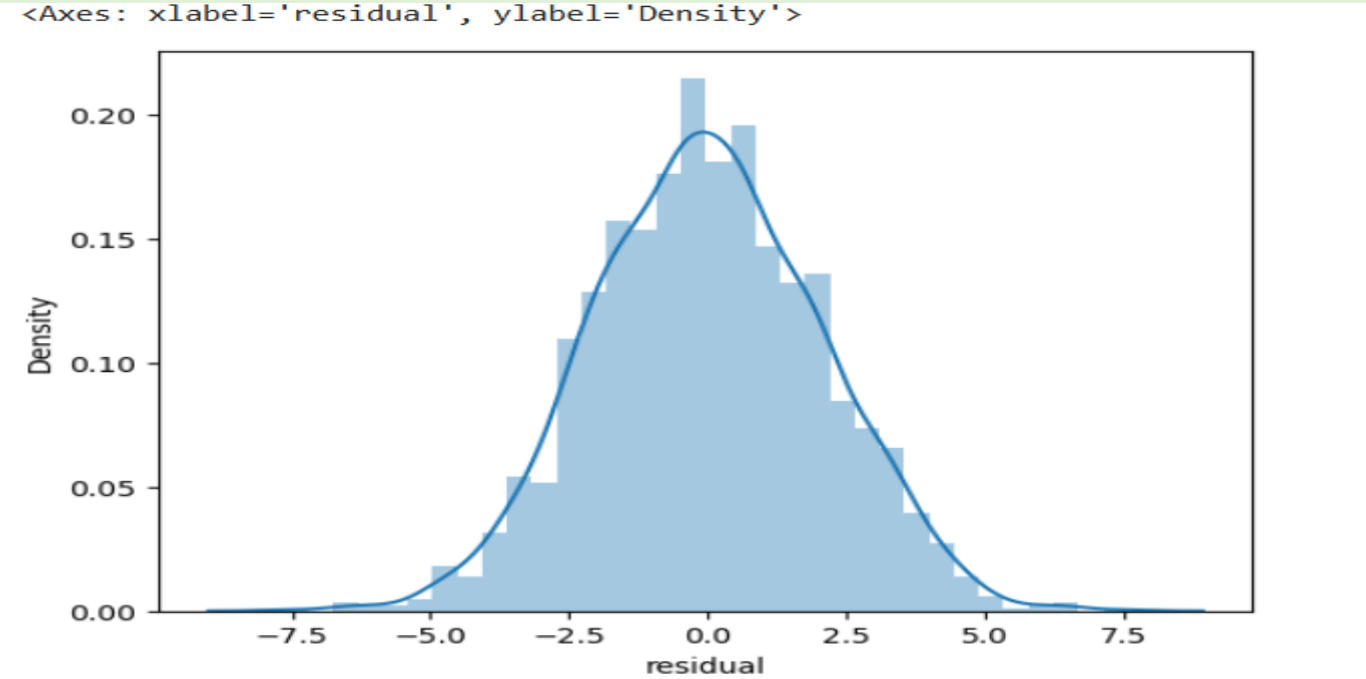
Predictions on Training Dataset **Predictions on Testing Dataset**

	actual	prediction
961	77.0	80.228480
4570	86.0	82.588988
9625	79.0	83.200300
8066	57.0	57.270900
5775	54.0	56.310515
...	...	...
6248	20.0	17.347553
501	82.0	78.216872
6841	67.0	66.685465
4469	16.0	18.939353
8670	82.0	82.589396

	actual	prediction	residual
3195	65.0	64.335915	0.664085
2543	60.0	63.049620	-3.049620
4101	91.0	92.719301	-1.719301
1222	51.0	49.554673	1.445327
6249	59.0	55.140917	3.859083
...	...	...	...
3364	70.0	69.807205	0.192795
1638	75.0	72.743444	2.256556
9540	54.0	50.693129	3.306871
2125	81.0	80.825609	0.174391
7513	37.0	40.224051	-3.224051

Assumption Check:

The residual points should follow Normal Distribuiton;
The below plot showing bell curve, which indicates, the residual points following Nomral distribution.



The Below plot will show normality

The blue points are perfectly, aligned closely to the straight line. The straight line, which is the best fit line which came from the model.

```
Text(0.5, 1.0, 'Q-Q Plot for Residual Normality Check')
```