Detecting Student Performance Index

Prepared By: Erum Akmal

1. **Problem Statement for Project:**
   In this project, I have predicted the student performance index based on Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours and Sample Question Papers Practiced. Linear Regression has been used for Multivariate Dataset.

2. **Data Sources and Description:**
   For data sources I used Kaggle.

   For this project, I worked on a student performance index project with linear regression for multivariate data set. My data set includes performance index (dependent variable) and Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours and Sample Question Papers Practiced as independent variables.

3. **Data Mining Operations:**
   -Data wrangling:
   In preprocessing of data, YES/No were converted into numbers. Handling of data was not required since there was no missing data.

   After that data was normalized using min max scaler. In the next step, splitting of data was done. Data was partitioned into 80% training and 20% testing.

   Linear regression was performed and model evaluations were done; followed by the generation of graphs.

   -Libraries Used: I have used multiple libraries in this project including pandas, NumPy, matplotlib, seaborn, scikit-learn.

   -Data Modeling: I have performed the analysis below in my project:

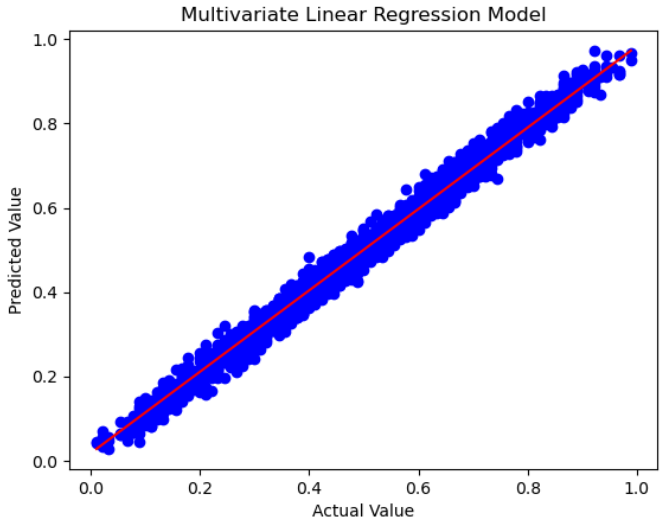
   - Preprocessing of data
   - Correlation metrics
   - Min Max Scaler
   - Splitting dataset into 80-20
   - Modelling with Linear regression
   - Evaluation: MSE, MAE, r2_score
   - Scatter plot graph
   - Correlation heatmap
   - Histogram of diff or target and prediction
   - Box plot for dataset

4. **Model Outputs:**
   Below is my evaluation of the model performance

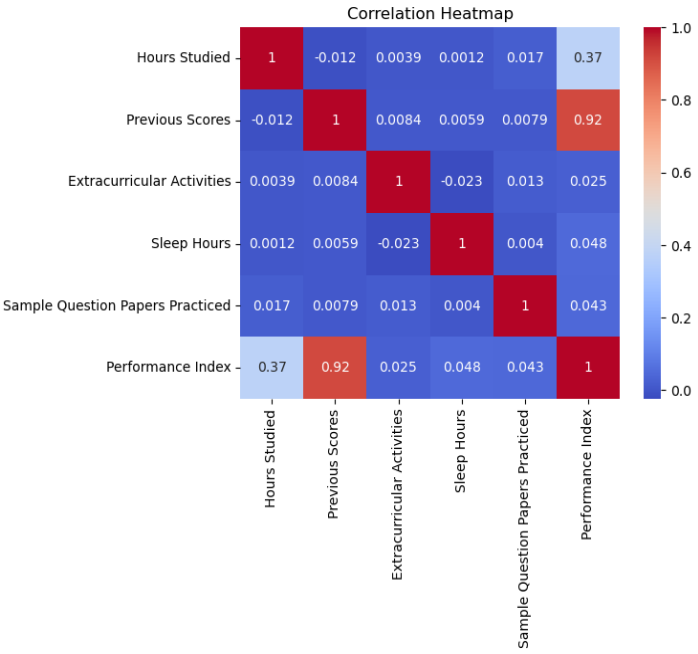
   - mean squared error: 0.0005
   - mean absolute error: 0.0179
   - r2score: 0.988

After the analysis, I can observe a positive correlation between performance index and other dependent variables. The highest correlation was between performance index and previous scores, the second highest correlation was observed for hours studies. The lowest correlation was observed between performance index and extra-curricular activities. A 98% variance score in performance index is explained here.

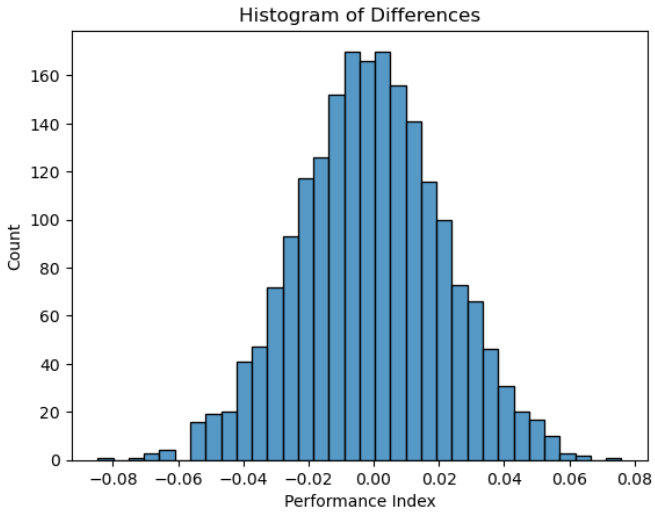
Scatter plot was generated with predicted value on y-axis and actual value of x-axis.



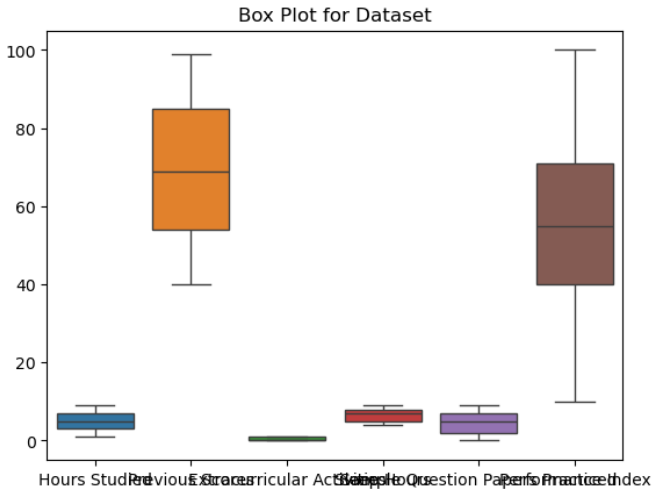
The correlation heatmap was generated next for performance index and all other independent variables. For features, previous scores and hours of studies are affecting the performance index the most. Extracurricular activities have least amount of affect on performance index.



Histogram of differences for student performance index was plotted next.



At the end, box plot was generated.



5. Limitations and Challenges:

Some of the limitations of this study are that I have only used 5 variables. There could be multiple other factors that can potentially affect performance index. This model worked well but there might be multiple other reasons that can affect student performance which were not covered in the data set.

After looking at r2 score, it can be observed that there is 98% variance in student performance index.

6. Conclusion:

I was able to solve the problem partially. By adding more data and variables in this study, a better model could have been predicted for performance index. By adding more number of features and rows, we can test with multiple factors that can play a role in student performance.

Linear regression on multi variate data set has been performed in this study. A linear positive correlation is shown in graphs above which means that with increase in independent variables; student performance index is likely to be increased.