My Project Report: Finding the Best Model to Predict Student Performance

Part 1: Getting the Data Ready

Before I could build any models, I had to prepare the data. This was a crucial first step. What I did: First, I loaded the Student_Performance.csv file. I checked for any missing values (there were none) and removed all the duplicate rows. A key step was converting the 'Extracurricular Activities' column from text ('Yes'/'No') into numbers (1/0). I also converted the Performance Index column to a float data type to make sure it was ready for calculations. Finally, I renamed the columns, replacing spaces with underscores to make them easier to work with in my code. My reasoning: I knew that starting with high-quality data was essential. By cleaning and standardizing it, I made sure my results would be accurate and reliable. After all the cleaning, I saved the final, preprocessed data into a new file called preprocessed_data.csv. I also created a correlation heatmap to explore the relationships in the data. From the heatmap, I saw that Previous_Scores had the strongest correlation with Performance_Index, so I decided to use that as the starting point for my predictions.

Part 2: Building and Testing My Models

I built four different models to compare their performance.

Model 1: Simple Linear Regression

What I did: I built a model to predict Performance_Index using only Previous_Scores. What I found: The R^2 score was 0.84, which was pretty good. It meant that 84% of a student's performance could be explained by their previous scores. However, the MSE was 59.96, which seemed a bit high, so I knew I could probably do better.

Model 2: Multiple Linear Regression

What I did: I added Hours_Studied as a second predictor variable. What I found: The results were amazing! The R^2 score jumped to 0.99, and the MSE dropped all the way down to 5.57. This told me that I had found a much more accurate model. Adding Hours_Studied was the key.

Model 3: Polynomial Regression

What I did: I tested a polynomial model to see if it could fit the data better than a straight line. What I found: The R^2 score was 0.82, which was actually a little worse than my simple

linear model. This experiment proved that the relationship was linear after all, and making the model more complex wasn't helpful here.

Model 4: Logistic Regression

What I did: I tried to predict whether a student did extracurricular activities based on their performance index. What I found: The model's accuracy was only 0.51, which is basically a coin toss. This showed me that a student's performance score isn't a good way to predict if they participate in extracurriculars.

My Final Conclusion

After all my tests, it was obvious that the Multiple Linear Regression model was the best one. The biggest lesson I learned from this project was how important feature selection is. Just by adding one more relevant variable (Hours_Studied), I was able to make my model incredibly accurate.