**Lab 3 – Quantitative Metrics**

**Data Preprocessing and Initial Challenges**

I successfully addressed two key data preprocessing challenges in my analysis. Firstly, I resolved the issue of missing values and tackled data type inconsistencies. Following this, I preserved the current state of the dataset by saving it. Subsequently, I reloaded the data and performed data standardization using the 'LabelEncoder' from the Scikit-Learn library.

My analysis extended to feature selection, where I employed various techniques to identify the most relevant columns. After rigorous evaluation, I identified a set of 30 features that exhibited the strongest associations with the target variable. As a result, I pruned the remaining columns, focusing on these 30 selected features for subsequent analysis.

**Application of Linear Regression Model**

I then applied a Linear Regression algorithm to the refined dataset and assessed its performance. Initially, when evaluating the model on the training data, I observed a remarkable accuracy score of 1. However, this seemingly perfect accuracy raised concerns of overfitting, prompting me to verify the model's performance on unseen data.

**Testing Data Evaluation and Visualization**

To evaluate the model's generalization capability, I employed a scatter plot to visualize the disparities between actual and predicted outputs. To my satisfaction, the plot revealed a close alignment between the two, suggesting that the model's predictions were highly accurate on the testing data.

**Quantitative Model Assessment**

To quantitatively gauge the model's performance, I calculated three common metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). All three metrics consistently indicated that the model's error was nearly zero, reinforcing the conclusion that the model was performing exceptionally well.

**High Accuracy Justification**

This exceptional accuracy can be attributed to the high correlation observed between various features and the target variable, with some correlations surpassing the 0.85 threshold. Moreover, the meticulous and iterative feature selection process, which distilled the initial 55 columns down to the optimal 30, further contributed to the model's outstanding performance.