Camden Alpert, Qasim Bhutta

Anurag Nagar

CS 4375.001

**Assignment 1 Report**

**Log**

**Run 1 (All Features Used - Adaptive Learning Rate with Initial Learning Rate of 0.001):**

R Squared Score (Training Set): 0.28

R Squared Score (Test Set): 0.27

Mean Squared Error (Training Set): 2.74

Mean Squared Error (Test Set): 2.78

**Run 2 (Removed: "Gender", "Height", "FAVC", "FCVC", "NCP", "SMOKE", "CH2O", "SCC", "TUE", "MTRANS"):**

R Squared Score (Training Set): 0.24

R Squared Score (Test Set): 0.25

Mean Squared Error (Training Set): 2.87

Mean Squared Error (Test Set): 2.85

**Run 3 (Removed: "Gender", "Height", "FAVC", "FCVC", "SMOKE", "MTRANS"):**

R Squared Score (Training Set): 0.25

R Squared Score (Test Set): 0.27

Mean Squared Error (Training Set): 2.83

Mean Squared Error (Test Set): 2.76

**Run 4 (Features Used: 'Age', 'Weight', 'family\_history\_with\_overweight', 'CAEC')**

R Squared Score (Training Set): 0.23

**R Squared Score (Test Set): 0.32**

Mean Squared Error (Training Set): 2.92

**Mean Squared Error (Test Set): 2.73**

\*(Run 4 had best test results)

**Questions**

**Are you satisfied that you have found the best solution? Explain.**

We are not fully satisfied that we have found the best solution overall, as the results from both the training and test sets were underwhelming. However, we believe we have found the best solution within the constraints of a linear regression model, having undertaken various steps such as feature selection, hyperparameter tuning, data preprocessing, and outlier removal to optimize performance. Despite these efforts, the final R-squared score of 0.32 and MSE of 2.73 on the test set fell short of our goal (R-squared ≥ 0.5, with a lower MSE). This suggests that while we have maximized the linear regression approach, a non-linear model might better capture the complexities of predicting obesity, as the relationship does not appear to be linear. Additionally, the low correlation values, with a maximum of approximately 0.38, likely contributed to the model’s limited performance.

**Plots**

**A graph with a line

Description automatically generated**

**A graph with blue dots

Description automatically generated**