Executive Summary:

Problem:

The COVID-19 outbreak introduced mask-wearing as a key prevention measure, but adherence and effectiveness vary. This project predicts mask-wearing frequency among individuals with non-household members. Questions assess various aspects of mask use, including perceived risks, benefits, social norms, and personal choice.

Data Description/Overview:

The dataset provides insights into A&M students attitudes towards mask-wearing during the early stages of the pandemic. It consists of 154 input variables (Questions) representing (asking) various aspects of mask-wearing, such as perceived benefits, risks, and social norms. The output variable corresponds to the frequency of mask usage around non-household members, with a scale from 1 (Never) to 5 (Always). The dataset includes responses from 479 individuals for all the 154 questions posed.

Data Clean-up:

We began by properly handling raw data, focusing on missing values for effective model development. The training data was checked, pre-processed, and imputed using medians to address missing values. All columns were transformed into integers, and a new data frame was created with processed data. We recorded the final dataset's dimensions, rows, and predictor columns to ensure the correct direction for future analyses.

Model Explanations:

We have used three models: Forward Subset Selection, Random Forests (Bagging), and LDA. Forward Selection and Random Forests regression models cannot be directly compared to a classification model such as LDA. So, a single parameter, test MSE is calculated for LDA also and compared to the remaining Regression models.

Test Results:

The performance of three different models was assessed to determine the best approach for our dataset. Among these models, the Forward Stepwise Selection showed moderate results, while the Random Forest model demonstrated significantly better performance, making it the preferred choice. Additionally, a classification model called LDA was evaluated, but it did not perform as well as the Random Forest model. In summary, the Random Forest model is the most effective option for our dataset, delivering the highest level of accuracy and reliability for decision-making purposes.

The best-performing model, Random Forest, has a Test Mean Squared Error (MSE) of 0.79, an Adjusted R-squared value for the training model is 0.86. These results indicate that the model is highly accurate and reliable in predicting outcomes, making it the optimal choice for our dataset of all the models.

Upon examining the results of the analyses, it was determined that the survey questions alone do not sufficiently capture the necessary information to predict whether an individual will or will not wear a mask around non-household members.

Although the survey may not capture all the necessary information, the most significant questions address an individual's understanding of the importance of mask-wearing in public, public perception of the COVID threat, public attitudes towards wearing masks, and their confidence and well-being.