STATS 101C Final Project

Predictive Analysis of Obesity Status

Kevin Ngo, Andrew Darwin, Alia Shibly, Cassia Ramelb,
Briana Nguyen

1 Abstract

The objective of this study is to develop and evaluate predictive models for classifying individuals' obesity status based on 29 predictors related to physical characteristics. The dataset, consisting of 32,014 observations, required employing various data imputation methods to address missing values. This report extensively describes the modeling process including exploratory data analysis, data preprocessing, feature selection, model construction, and evaluation. The random forest model, optimized with ntree = 4 and mtry = 3, achieved the desired balance of accuracy and simplicity with a kaggle score of 0.94537. The final model identified daily water intake, height, age, physical activity frequency, and time using technology devices as the most important predictors.

2 Introduction

Obesity is a global health issue and is one of the leading risk factors for various chronic diseases, such as diabetes, cardiovascular conditions, and some cancers. The prevalence of obesity has been steadily increasing worldwide, with about 43% of adults classified as overweight and 16% as obese. In the United States this issue is even more disastrous, with about 40% of adults suffering from obesity as of 2023. The complexity of addressing obesity lies in the multitude of factors that contribute to its development. Access to healthy food, opportunities for physical activity, industrialization of food production, cultural attitudes toward health, economic status, genetic predisposition, and ethnicity are all examples of these interwoven factors. Together, they create challenges in treating and preventing obesity, which carries significant health and economic burdens.

Because obesity is associated with such significant health risks and economic costs, early identification and prevention is crucial for improving public health. Effective prediction of obesity status based on physical and lifestyle factors can play a vital role in tailoring interventions and promoting healthier behaviors. This project focuses on developing predictive models that balance accuracy with simplicity and interpretability. Creating a model that accurately predicts obesity is crucial, but its practical application also must be clear and accessible for widespread use. Thus, our aim is to construct a model that effectively classifies individuals as obese or not obese but also can offer meaningful, actionable insights to address obesity.

3 Methods

3.1 Data Preprocessing

The Obesity dataset consists of both training and testing subsets. The training set has 32,014 observations and 30 variables, including 29 predictors and the response variable, ObStatus. The testing set has 10,672 observations and 29 predictors. The predictors are divided into 12 numeric variables, such as Age, Height, and avg_glucose_level, and 17 categorical variables, including Gender, FAVC, and CAEC. The response variable, ObStatus, indicates whether an individual is classified as Obese or Not Obese. The class proportions are relatively imbalanced, with 61.01% of individuals in the Not Obese category and 38.99% in the Obese category.

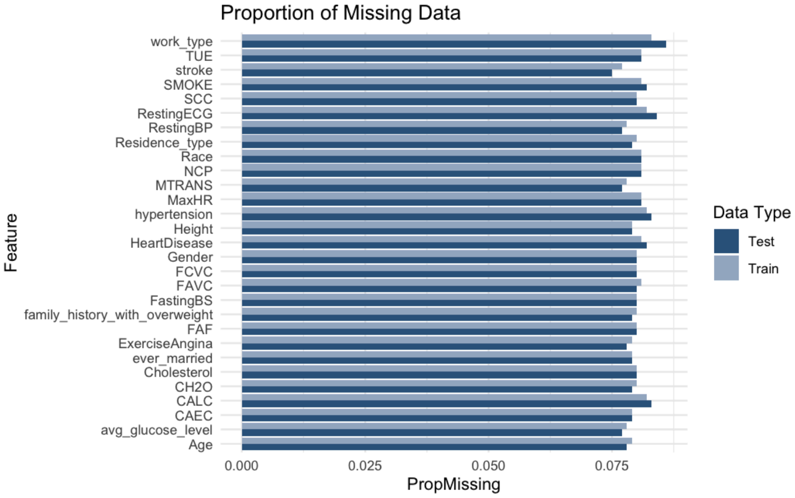


Figure 3.1: Proportion of Missing Values for Each Feature

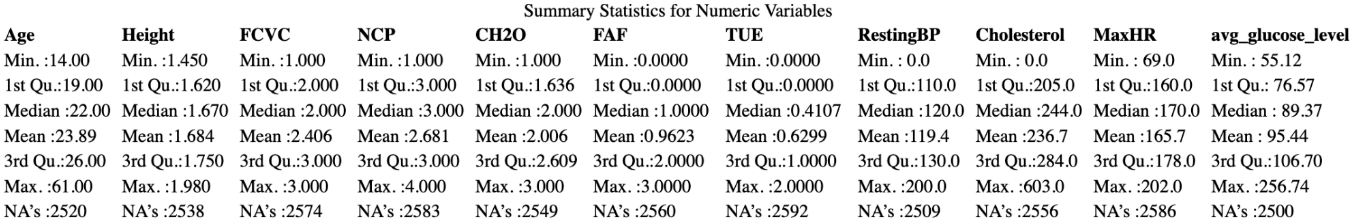
After analyzing its structure, we tackled the missing values that were apparent in our dataset. We found that all given predictors in each subset had about 8% of missing values with significant missingness apparent in variables such as NCP, FAF, and TUE. Handling the missing data during the preprocessing stage was crucial to our analysis, as it would have otherwise reduced the accuracy of our model and limited the machine learning algorithms available to us.

For numeric features, we used multivariate imputation by chained equations (MICE) because it uses the relationship between variables to impute missing values based on the natural patterns in the dataset. This approach preserves the variability of the dataset and reduces bias, compared to simpler imputation methods. For categorical features, missing values were imputed using class proportions. We used this approach because it ensures that the distribution of the imputed values align with the observed proportions of categories in the dataset. For example, if a categorical variable like MTRANS had missing entries, the imputed values would be assigned proportionally to the observed frequency of its categories ("Automobile", "Walking", or "Public_Transportation"). Our data cleaning did not reduce the dimensions of the data.

3.2 Data Exploration

3.2.1 Summary Statistics

Table 3.2.1: Summary Statistics for Numeric Variables



Summary Statistics for Numeric Variables

Age	Height	FCVC	NCP	CH2O	FAF	TUE	RestingBP	Cholesterol	MaxHR	avg_glucose_level
Min. :14.00	Min. :1.450	Min. :1.000	Min. :1.000	Min. :1.000	Min. :0.0000	Min. :0.0000	Min. : 0.0	Min. : 0.0	Min. : 69.0	Min. : 55.12
1st Qu.:19.00	1st Qu.:1.620	1st Qu.:2.000	1st Qu.:3.000	1st Qu.:1.636	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:110.0	1st Qu.:205.0	1st Qu.:160.0	1st Qu.: 76.57
Median :22.00	Median :1.670	Median :2.000	Median :3.000	Median :2.000	Median :1.0000	Median :0.4107	Median :120.0	Median :244.0	Median :170.0	Median : 89.37
Mean :23.89	Mean :1.684	Mean :2.406	Mean :2.681	Mean :2.006	Mean :0.9623	Mean :0.6299	Mean :119.4	Mean :236.7	Mean :165.7	Mean : 95.44
3rd Qu.:26.00	3rd Qu.:1.750	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:2.609	3rd Qu.:2.0000	3rd Qu.:1.0000	3rd Qu.:130.0	3rd Qu.:284.0	3rd Qu.:178.0	3rd Qu.:106.70
Max. :61.00	Max. :1.980	Max. :3.000	Max. :4.000	Max. :3.000	Max. :3.0000	Max. :2.0000	Max. :200.0	Max. :603.0	Max. :202.0	Max. :256.74
NA's :2520	NA's :2538	NA's :2574	NA's :2583	NA's :2549	NA's :2560	NA's :2592	NA's :2509	NA's :2556	NA's :2586	NA's :2500

The table summarizes the numeric variables, showing key statistics like the minimum, median, mean, and maximum values, as well as the count of missing data. Key observations include wide ranges for RestingBP and Cholesterol, while MaxHR shows less variability. Missing values are significant in NCP, FAF, and TUE, which were addressed during analysis.

3.2.2 Correlation Heatmap

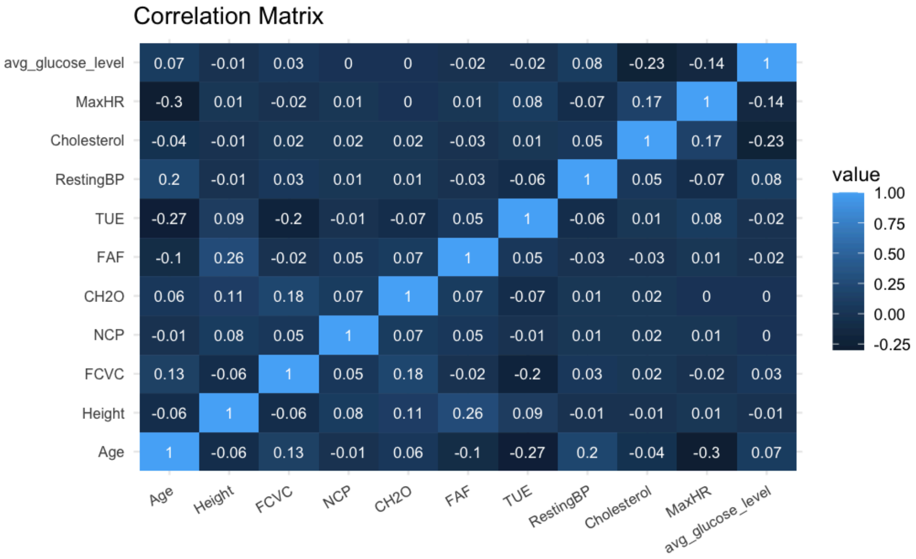


Figure 3.2.2: Correlation Matrix for Numeric Variables

The correlation heatmap shows relationships between numeric features, with most correlations being weak to moderate, indicating minimal multicollinearity. Age and Height have low correlations with other variables, while FAF and TUE show slight negative relationships with Age.

However, variables like FAF and CH2O display slight positive relationships, suggesting they may be related. Overall, the heatmap helps identify potential feature relationships and guides feature selection for modeling.

3.2.3 Density Plots and Boxplots for Numeric Variables

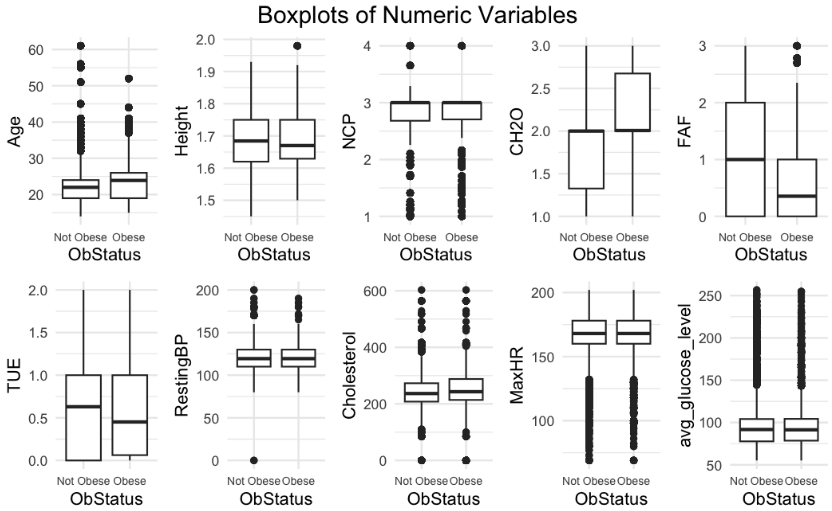


Figure 3.2.3.1: Boxplots for Numeric Variables by Obesity Status

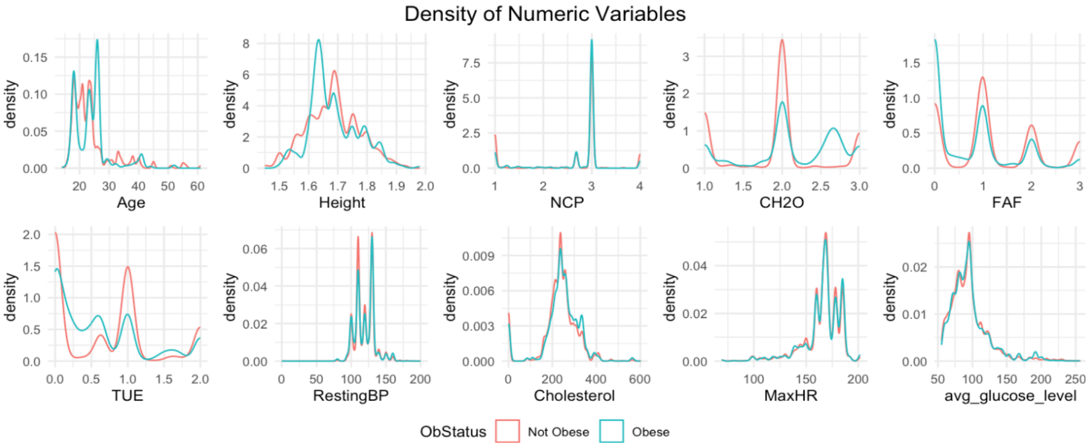


Figure 3.2.3.2: Density Plots for Numeric Variables by Obesity Status

The density plots compare numeric variables for “Obese” and “Not Obese” individuals. Age and Height show clear differences, with “Obese” individuals being older and shorter. Peaks in CH2O suggest common patterns across both groups, while FAF and TUE indicate “Not Obese” individuals tend to have higher physical activity. These trends highlight key variables for distinguishing between the two groups and predicting obesity status. These plots also show the distribution of the numeric data. It can be seen that Age, Height, and RestingBP have relatively normal distributions, while Cholesterol and avg_glucose_level are skewed. Discrete variables like FCVC, NCP, and FAF have clear peaks, suggesting common values. These patterns reveal data variability and potential outliers that may need further analysis.

3.2.4 Bar Charts for Categorical Variables

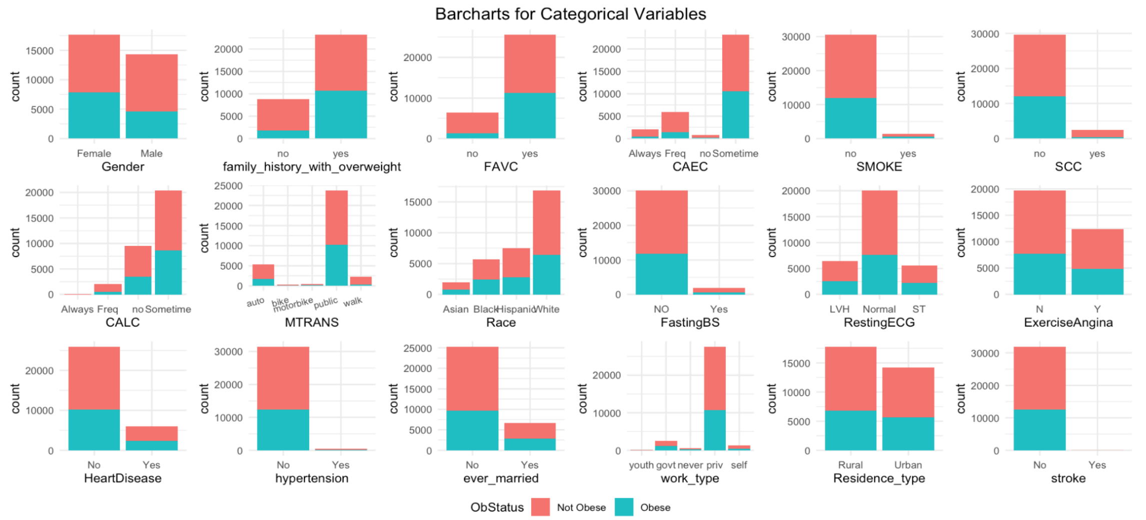


Figure 3.2.4.1: Bar Charts of Counts for Categorical Variables by Obesity Status

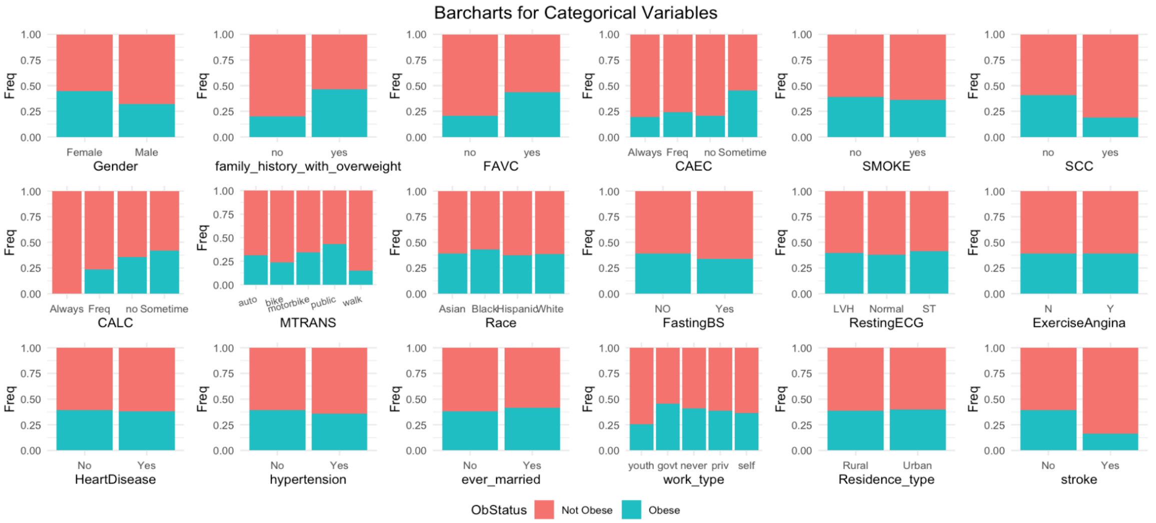


Figure 3.2.4.2: Bar Charts of Proportion for Categorical Variables by Obesity Status

The bar charts show the distribution of categorical variables by obesity status. Features like family_history_with_overweight and FAVC have strong associations with obesity, with more “Obese” individuals in the “Yes” category. Trends in CALC and MTRANS also highlight group differences. Smaller features like SMOKE and SCC still provide useful insights for identifying key factors related to obesity.

3.3 Data Modeling

We tested several models to classify Obesity Status ranging from simple to complex. We then chose the model that had a good balance between high accuracy on the testing data and low complexity.

3.3.1 Logistic Regression

Logistic regression was used to create our baseline models, since it is a simplistic and efficient way to solve binary classification problems. A full model was first created using all of the 29 predictors and then a reduced model of 17 predictors was obtained using only statistically significant predictors from the full model summary. Additionally, stepwise (forward and backward) regression was used to create reduced models of 16 predictors based on the criteria of minimizing AIC or BIC.

Lastly, regularization was applied to address any multicollinearity and analyze feature importance. Cross-validation was applied to both lasso and ridge regularization to obtain the best penalty parameter of 0.01 for ridge and 0.0005 for lasso. From the coefficients of our lasso regression, we determined the top significant predictors in the model, which include Height, CAEC, family_history, stroke, FAVC, and MTRANS.

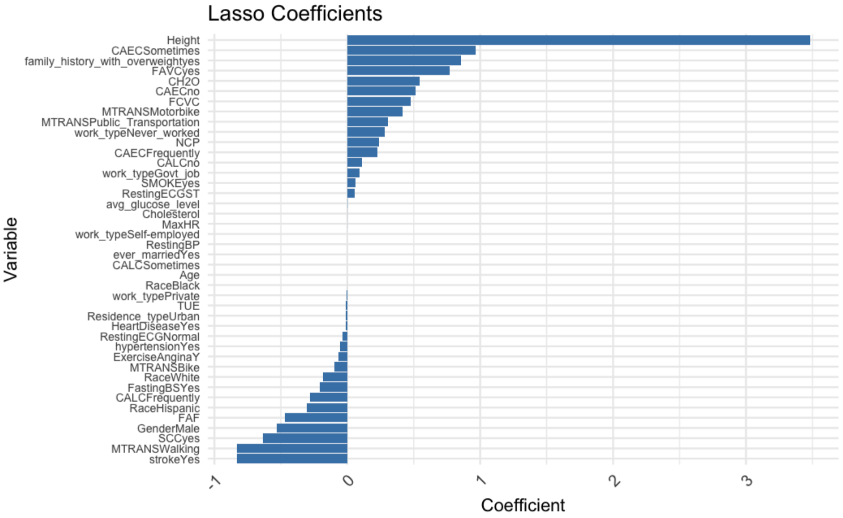


Figure 3.3.1: Feature Importance Using Lasso Regression

3.3.2 Decision Trees

We tested decision trees because of their ability to learn nonlinear relationships well. To find the optimal tree, we tuned the complexity (cp) and the max_depth parameters via a grid search that performed 5-fold cross validation on every possible parameter combination in the grid.

We also tried pruning and reducing the number of features used for our decision tree model, but the best performing model was a tree with cp = 0.001 and max_depth = 10. Because of the complexity of the model, a limitation is that it can overfit to the data, meaning that it might not generalize well to unseen patient data.

3.3.3 Random Forest (Best Model)

Because the decision tree was relatively complex, but still was not the best in terms of accuracy, we tested a random forest model. Using bagging and training each tree on a random subset of features, random forests are made of multiple slightly different decision trees. The final decision made by the model is a majority vote of the classifications of the individual trees. It used the following 16 features: CH2O, FAF, Age, SMOKE, Race, FCVC, NCP, TUE, FAVC, Height, SCC, Gender, CALC, CAEC, MTRANS, and family_history_with_overweight. Because multiple trees are involved, the random forest model is more complex than a decision tree, but is less likely to overfit to the data and should generalize better due to the averaging effect of multiple trees.

The random forest was tuned via a grid search that tested the mtry and ntree parameters using 5-fold cross-validation. We also used the feature importances generated by the random forest to reduce the number of features used in our model. We tested different random forests using the top 5, 10, and 20 most importance features, and compared their results. However while tuning, we noticed that the optimal model outputted by the grid search was very complex, and looking at the results of the less optimal models, they were simpler while sacrificing a negligible amount of accuracy. So we graphed the results in an elbow plot to determine the best model that performed accurately but also did not sacrifice simplicity for only marginal increases in accuracy.

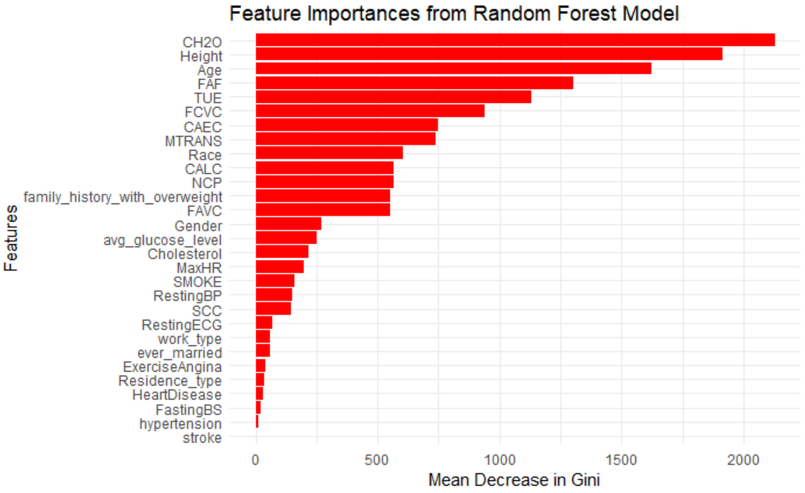


Figure 3.3.3.1: Feature Importances of Random Forest

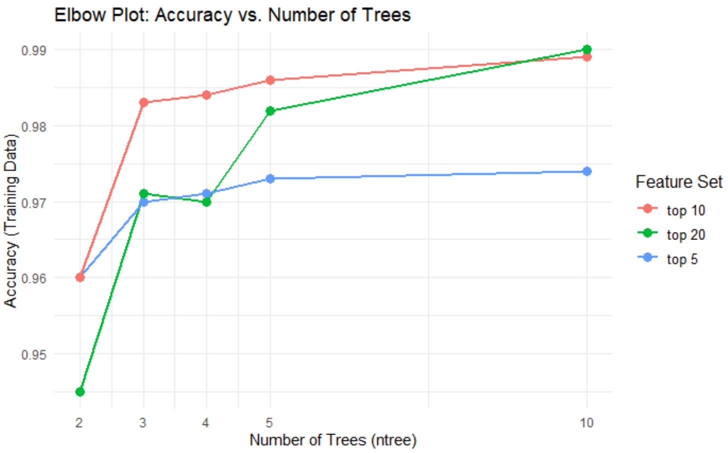


Figure 3.3.3.2: Elbow plot of Accuracy by Number of Tree split by feature subsets

Based on the elbow plot, we determined that the most appropriate random forest was one that was trained on the top 5 features, mtry = 3, and ntree = 4. The 5 features used were the following: CH2O, Height, Age, FAF, and TUE. While it doesn't appear to be the most accurate model in the elbow plot, it was simple and performed the best compared to the other hyperparameter combinations.

3.3.4 Support Vector Machine (SVM)

Since our data contains nonlinear decision boundaries and is high-dimensional with 29 predictor variables, we also utilized support vector machines (SVM) with various kernels. We used a linear kernel, a polynomial kernel and a radial kernel. For the SVM with the polynomial kernel, we performed a 5-fold cross-validation with parameter tuning using degrees from 2 to 5 in order to obtain the optimal degree of 3.

4 Results and Limitations

The results (training confusion matrices and training/testing accuracies) of the various models can be found below.

4.1 Logistic Regression

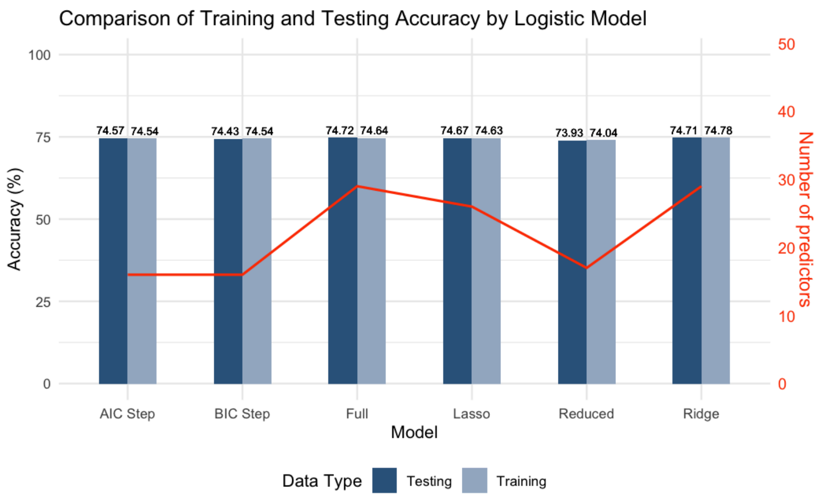


Figure 4.1: Performance of Different Logistic Models on Training and Testing Data

Table 4.1.1: Confusion Matrix for Best Logistic Model (AIC Step) on Training Data

	Predicted		
		Yes	No
Actual	Yes	16518	3013
	No	5138	7345

Table 4.1.2: Training and Testing Accuracies for Best Logistic Model (AIC Step)

Training Accuracy	**Testing Accuracy**
74.54%	74.57%

4.2 Decision Tree

Table 4.2.1: Confusion Matrix for Best Decision Tree on Training Data

	Predicted		
		Yes	No
Actual	Yes	19145	1667
	No	386	10816

Table 4.2.2: Training and Testing Accuracies for Best Decision Tree (cp = 0.001, max_depth = 10)

Training Accuracy	**Testing Accuracy**
93.42%	91.7%

4.3 Random Forest (Best Model)

Table 4.3.1: Confusion Matrix for Best Random Forest on Training Data

	Predicted		
		Yes	No
Actual	Yes	19424	107
	No	613	11870

Table 4.3.2: Training and Testing Accuracies for Best Random Forest (mtry = 3, ntree = 4)

Training Accuracy	**Testing Accuracy**
97.02%	94.5%

4.4 Support Vector Machine (SVM)

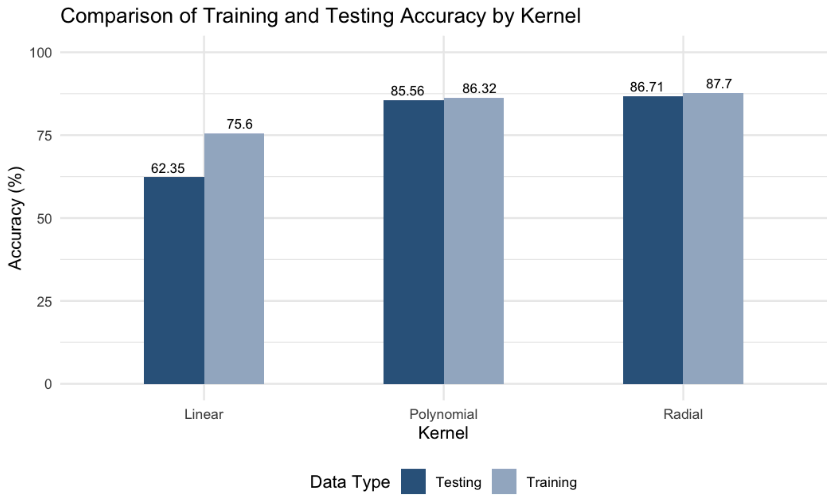


Figure 4.4: Training and Testing Accuracies by Kernel

Table 4.4.1: Confusion Matrix for Best SVM (Radial) on Training Data

	Predicted		
Actual		Yes	No
	Yes	18320	1211
	No	2729	9754

Table 4.4.2: Training and Testing Accuracies for Best SVM (Radial)

Training Accuracy	**Testing Accuracy**
87.70%	86.71%

4.5 Limitations

The first limitation was that we could have tried different ways to preprocess the data before inputting the data into the model. We tried data imputation with various combinations of multivariate imputation by chained equations (mice), amelia, mean, median, mode, and randomly based on existing class proportions, and found that the dataset that imputed numeric features with mice and categorical features with existing class proportions performed best. However, if given more time, we would have tried performing PCA and feeding the principal components into the model and experimenting with various transformations such as log to see if that improves the results of our models.

Another limitation of our modeling was the method we used to reduce the number of features of the random forest. We determined arbitrary cutoffs (top 5, 10, and 20 predictors) and tested different random forests models using these subsets. However, this method has limitations because arbitrarily cutting off the rest of the features fails to account for potential interaction between the different features. To improve this study, in the future, we would try to implement recursive feature elimination, which is a method of variable selection for random forests. This method is comparable to backward stepwise selection, which considers the combined effects of features and their contribution to the model. This allows for a more comprehensive and robust selection of features.

5 Discussion and Recommendations

Each of these previously discussed models offered various balances of accuracy, complexity, and interpretability for predicting obesity status. Logistic regression offered a baseline as the simplest, most interpretable model that we reduced to 17 predictors. Logistic regression provides extremely efficient computations and insight into predictor importance, however the testing accuracy of about 75% was not suitable for the goal of our study, likely indicating a nonlinear relationship between obesity status and the predictors. Decision trees address this issue of nonlinearity, offering improved performance while remaining relatively interpretable. However, they are prone to overfitting, and even with the best performing tree limited to a maximum depth of 10, misclassification on testing data was around 9%. The support vector machine (SVM) model is more computationally expensive than the other models but captures much more complex patterns in the data. SVM with a radial kernel is also very difficult to interpret as a result of the intricate decision boundaries and performed worse than the decision tree. Finally, the most complex random forest models we tried resulted in nearly 99% testing accuracy, but they also suffered from the complexity and difficulty to interpret that SVM does. To combat this we decided on a model with reduced parameters of ntree = 4 and mtry = 3, resulting in a more practical and interpretable solution while maintaining predictive power.

Several strategies could be explored to further enhance the performance and applicability of all models and to deepen the investigation of this topic. Feature interaction terms and transformations could possibly provide improved results, especially in models like logistic regression. Investigating alternative models, such as K-Nearest Neighbors, could offer valuable information about the data, especially because of its simplicity and ability to capture grouping patterns. Similarly, we could also implement KNN based imputation strategies to possibly further increase data quality and model robustness. As discussed previously, the response classes are imbalanced so addressing that with oversampling, undersampling, or some other augmentation of the data could also potentially improve outcomes. By incorporating these strategies, we can generate a broader range of models and metrics to test, improving our ability to understand and address the effects of the obesity epidemic.

6 Acknowledgements

World Health Organization. “Obesity and Overweight.” *World Health Organization*, WHO, 1 Mar. 2024, www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

CDC. “Obesity and Severe Obesity Prevalence in Adults: United States, August 2021–August 2023.” *CDC.gov*, 2024, www.cdc.gov/nchs/products/databriefs/db508.htm.