Machine Learning-Based Assessment of Obesity: An Investigation of Model Performance and Feature Selection

Aslanpour, Dareh

UNIVERSITY OF CALIFORNIA
Los Angeles

Machine Learning-Based Assessment of Obesity: An Investigation of Model Performance and Feature Selection

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Dareh Aslanpour

2023

ABSTRACT OF THE THESIS

Machine Learning-Based Assessment of Obesity: An Investigation of Model Performance and Feature Selection

by

Dareh Aslanpour
Master of Applied Statistics and Data Science
University of California, Los Angeles, 2023
Professor Yingnian Wu, Chair

The objective of this paper is to employ various machine learning algorithms to investigate the assessment of obesity levels based on eating habits and physical conditions. The study will utilize the obesity level estimation data provided by UCI Machine Learning Repository. The performance of different model candidates will be evaluated and compared in order to select the most robust model for obesity estimation or prediction. Moreover, this research aims to identify the crucial features used in the best predictive model to enhance the accuracy of obesity prediction. This study intends to contribute to the ongoing research in the field of machine learning and healthcare by providing insights into the prediction of obesity.

The thesis of Dareh Aslanpour is approved.

Maryam Mahtash Esfandiari

Frederic R Paik Schoenberg

Yingnian Wu, Committee Chair

University of California, Los Angeles

2023

To my parents and my brother…

who continually believe in me far more than I believe in myself

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Obesity has established itself as one of the most significant and ongoing epidemics in the United States, affecting nearly half of the population and linked to leading causes of preventable death such as heart disease and stroke [CDC22b]. Obesity is identified through a person's body mass index or BMI, which is calculated by dividing a person's weight in kilograms (kg) by their height in meters squared (m^2) where a BMI of 30 kg/m^2 or more is recognized as obese [CDC22a]. On the surface level, the concept of avoiding obesity may seem to be as simple as keeping your weight low relative to your height, but the issue of obesity is multifaceted as 50% of the variance in weight is explained by genetics and the other half by their environment [BLM15]. Therefore, it is of paramount importance to gain understanding of the environment in relation to obesity given its amenability to modification in comparison to genetics.

This study develops a classification model with the primary objective of accurately predicting obesity in individuals based on a multitude of environmental factors, including but not limited to their smoking habits and even mode of transportation to deepen the understanding of environmental impacts on obesity. By creating a myriad of machine learning models, our investigation will not only successfully predict obesity, but it will also allow us to achieve an understanding of the key drivers of obesity prediction and further our understanding on environmental factors that play a pivotal role in predicting whether or not a person is classified as obese. Conversely, the models will also aid in the understanding of factors that do not contribute to obesity.

CHAPTER 2

Exploratory Data Analysis

2.1 Data Set

The data set used is from the UCI Machine Learning Repository. Titled "Estimation of obesity levels based on eating habits and physical condition Data Set," it includes data for estimating the obesity levels of people from Mexico, Peru, and Colombia based on several habits and conditions. The data set contains 17 attributes and a total of 2111 observations where 23% of the data is collected and the other 77% is synthetic. The goal of implementing synthetic data into the data set was to not only increase the number of observations in general, but to also balance the number of observations that are obese and not obese. The variable we will be predicting is "NObesity" or obesity level which is split into seven levels: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III. For the purpose of our study, We will be converting this variable to a binary variable where 0 signifies not obese and 1 signifies obese. In addition, will be omitting the height and weight variables from our model due to its correlation to BMI as a BMI over 30 classifies obesity and is calculated by the formula $BMI = \frac{kg}{m^2}$. for reference, a test logistic regression model was implemented that classified obesity using solely the variables weight and height which predicts obesity at a 95% accuracy with the 5% incorrect being likely due to incorrect labeling of synthetic data. As a result, we will be converting height and weight variables to BMI and using it to supplement our understanding of BMI with respect to our categorical variables and obesity classification.

2.2 Attributes of Data Set

Gender: Female or Male (0 = Female, 1 = Male)

Age: Age in years

Height: Height in meters

Weight: Weight in kilograms

BMI: Body mass index of person ($\frac{Weight(kg)}{Height(m^2)}$)

family_history_with_overweight: Yes/No does the person have a family member who has suffered from being overweight (0 = No, 1 = Yes)

FAVC: Does the person frequently consume high caloric food (0 = No, 1 = Yes)

FCVC: How frequently does the person consume vegetables (0 = Never, 1 = Sometimes, 2 = Always)

NCP: How many main meals does the person consume a day (0 = 1 , 1 = 2-3, 2 = 3+)

CAEC: How often does the person consume food between their meals (0 = Never, 1 = Sometimes, 2 = Frequently, 3= Always)

SMOKE: Is the person a smoker (0 = No, 1 = Yes)

CH2O: How much water does the person drink per day (0 = ¡ 1 Liter, 1 = 1 -2 Liters, 2 = 2+ Liters)

SCC: Does the person monitor their daily calorie consumption (0 = No, 1 = Yes)

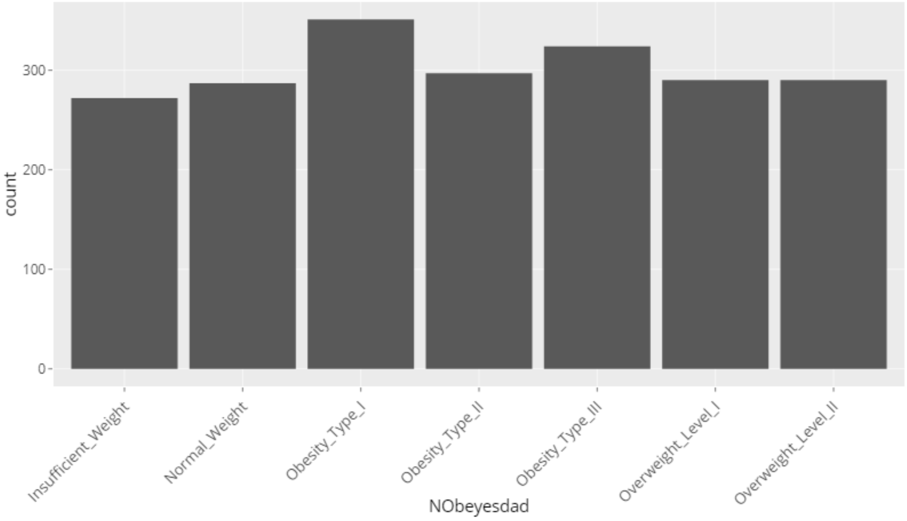
FAF: How many times per week is the person physically active (0 = None, 1 = 1 or 2 Days, 2 = 3 or 4 Days, 3 = 5+ Days)
TUE: How much time in hours does the person spend using technology (0 = 0-2 Hours, 1 = 3 to 5 Hours, 2 = 5+ Hours)
CALC: How frequent does the person consume alcohol (0 = No, 1 = Sometimes, 2 = Frequently, 3 = Always)
MTRANS: What transportation does the person usually take (0 = Walk or Bike, 1 = Automobile, 2 = Public Transportation)
NObeyesdad: Obesity level (Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III) Re-labeled to binary (0 = Not Obese, 1 = Obese)

As mentioned previously, Height and Weight will be omitted from all models as both are used to derive BMI which has direct relationship to obesity level. However, the BMI variable derived from the Height and Weight given in the data set will be used in exploratory data analysis and influence variable selection as it portrays obesity levels at a numeric level.

2.3 Dependent Variables

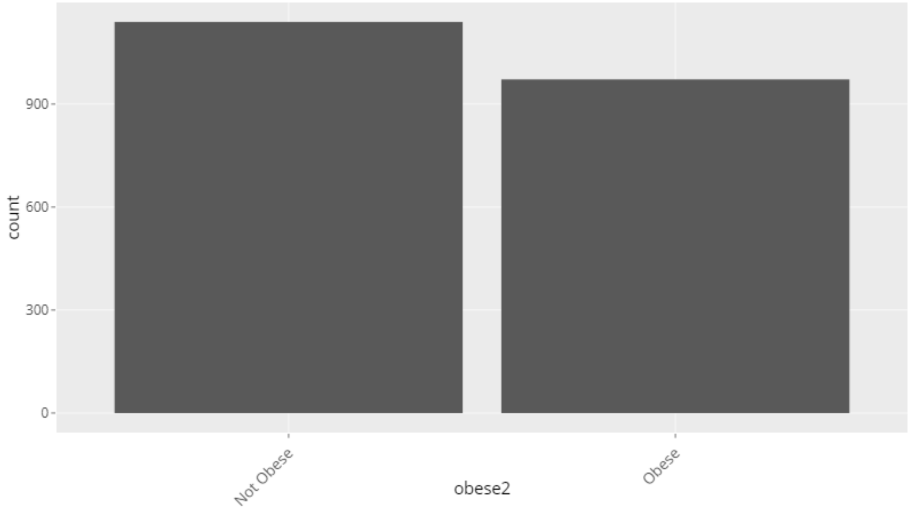
In the data set, the dependent variable is a categorical variable with seven levels: Insufficient Weight (272 occurrences), Normal Weight (287 occurrences), Overweight Level I (290 occurrences), Overweight Level II (290 occurrences), Obesity Type I (351 occurrences), Obesity Type II (297 occurrences), Obesity Type III (324 occurrences) which is far too many levels for our purpose of comparing obesity and non-obesity groups. From the barplot below we see that the distribution of the frequency of BMI classifications is roughly uniform although having seven levels.

Figure 2.1: Frequency of BMI Classifications



We opt to dissolve the seven levels into two levels which are “Not Obese” and “Obese” as the primary objective is based around a broad view of obesity rather than a holistic one that deep dives into every single stage of BMI Classification.

Figure 2.2: Frequency of Not Obese and Obese



As a result we now have 1,139 observations of people who are not obese and 972 observations of people who are obese as evident in Figure 2.2.

Lastly, for the purpose of exploratory data analysis, a secondary dependent variable, BMI, was created to further explore and understand how certain attributes might impact obesity. BMI is directly proportional to the classification of obesity since a BMI of 30 or more signifies obesity. Due to the presence of Height (in meters) and Weight (in kilograms) we are able to manually derive the BMI of each observation using the formula $\frac{Weight(kg)}{Height(m^2)}$. With the addition of BMI, we add another layer to our analysis as we can observe not only what variables are drivers in obesity prediction, but also in BMI. Our analysis is now observable at a numeric and categorical level. With this in mind and the fact that approximately 77% of the data set is synthetic, we remove observations from the data set that are classified as obese, but have a BMI less than 30, and observations that are classified as not obese, but have a BMI of 30 or greater. Due to the synthetic nature of the data, it is possible that the data created could be misclassified since it may not have been explicitly classified with the definition of obesity with respect to BMI in mind. As a result, our original data set of 2,111 observations drops to 2,105 after the removal of the misclassified observations. Furthermore, there are now 1,135 and 970 observations of non-obese and obese people, respectively, in the data set in comparison to the original 1,139 and 972. Lastly, we limit the data to observations of people who are age 18 or higher since due to personal intuition, environmental factors are likely to be far less controllable by someone who is under 18. Our data set is now limited to 1,992 observations with 1,030 being classified as not obese and 962 being obese.

Figure 2.3: Frequency of Not Obese and Obese After Misclassified Observation and Under 18 Removal

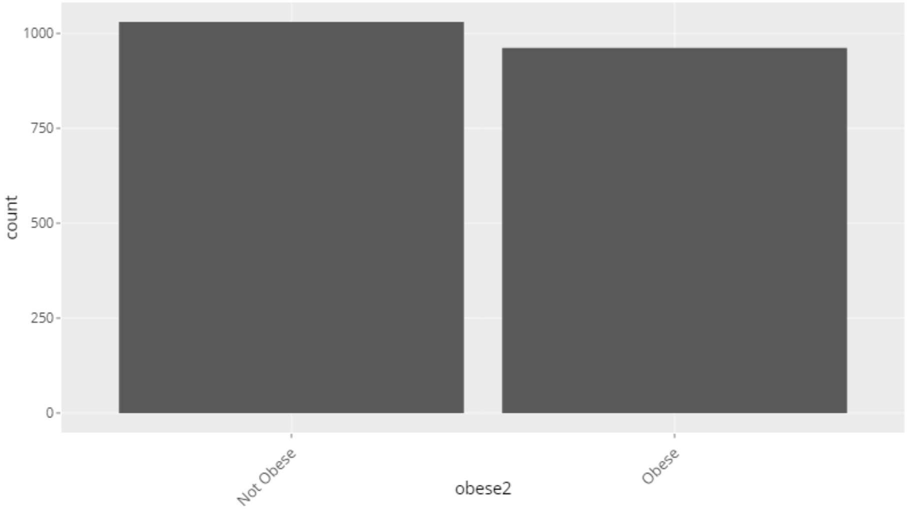


Figure 2.4: Boxplots of BMI with Respect to Obesity

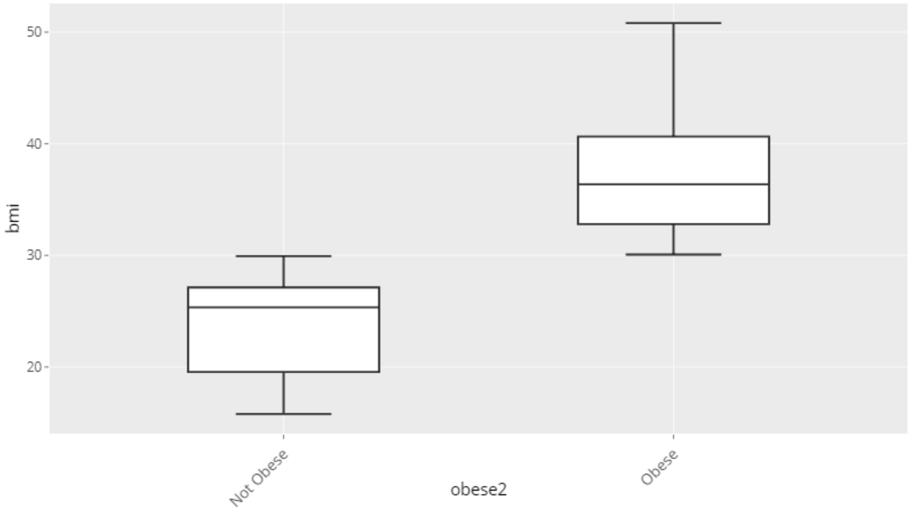
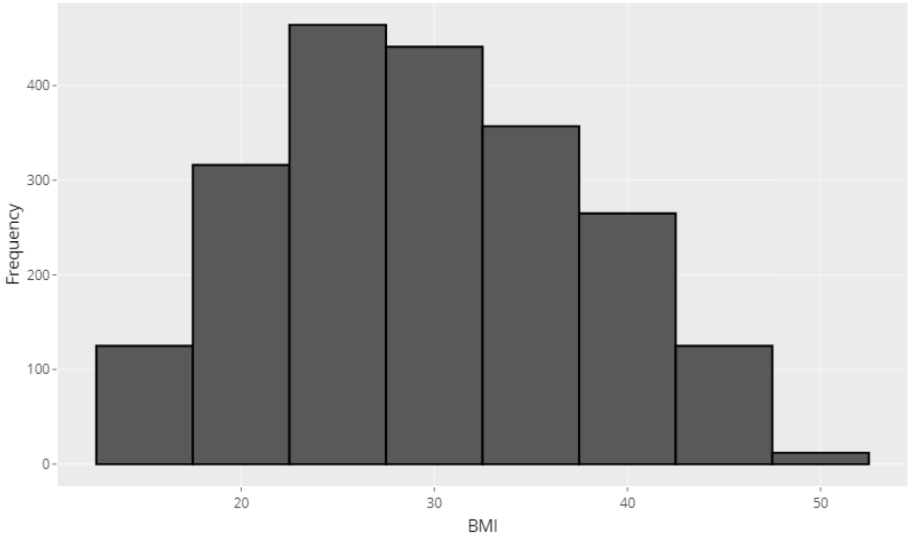


Figure 2.5: Histogram for the Distribution of BMI



In Figure 2.4 above, we can see that BMI with respect to Obesity is now structured the way we need it to be as “Not Obese” ends right before a BMI of 30 and “Obese” picks up right after. In addition, in Figure 2.5 it is also evident that the new variable “BMI” is also roughly normally distributed. Although it will not be explicitly used as a dependent variable in our model, it will be used as reference in variable exploration and selection, so is treated similarly to how a response variable in linear regression would normally be addressed.

2.4 Variable Exploration Methodology

In the data set there are three numeric variables and two of which are used in combination to calculate BMI which is directly correlated to the classification of obesity and are ultimately omitted from the analysis. This leaves Age as the only numeric predictor in the data set and was explored by creating histograms of the variable and various transformations of it to check for normality.

The rest of the variables in the data set are all categorical variables with the most important variable being gender due to its potential to be used in interaction effects or provide cause to create two different models, one for males and another for females, in the event that there is significant variance between the genders and their relation to obesity prediction. Every categorical variable will be plotted in the form of a table against obesity to see the overall distribution of the variable and if there looks to be any potential predictability power in hindsight. Moreover, boxplots will be created for BMI with respect to every level of each categorical variable to add a second layer of understanding and unveil how BMI is distributed for each level and whether or not there looks to be any visual disparity between them. Lastly, the boxplots will be repeated two more times, one for each subset of genders as means of exploring potential interaction effects.

2.5 Key Findings

Figure 2.6: Barplot of Gender with Respect to Obesity

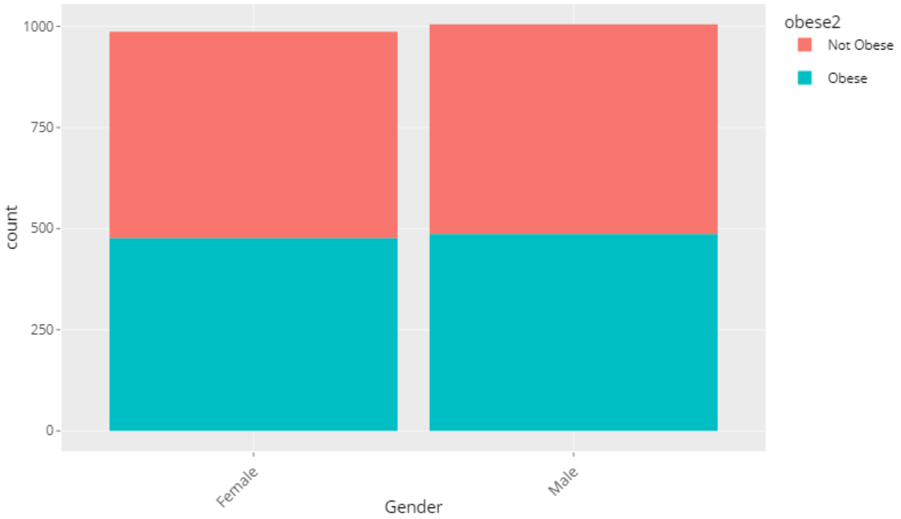


Figure 2.6 above suggests that in our data neither gender is more likely to be obese than the other as we have a similar amount of observations for each possible combination. This may be likely due to the synthetic nature of the data.

Table 2.1: Family History with Overweight vs Obesity Status

	Not Obese	Obese
no	336	8
yes	694	954

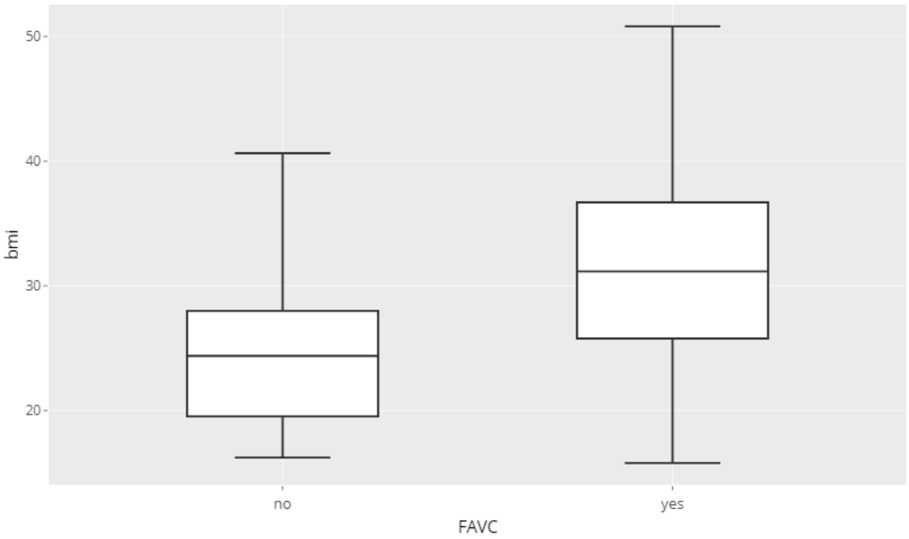
Table 2.1 contained one of the biggest findings within the analysis displaying that only 2.3% of people without family history of being overweight obese while 57.9% of people with a family history were obese. Although this has the potential to be a powerful finding in regard to obesity prediction it is important to note that there have been studies showing genetic factors to have a significant role in the pathogenesis of obesity [LB03]. Therefore, this variable will be omitted from any models as 50% of the variance in weight is explained by genetics and our focus is centered around controllable aspects [BLM15].

Table 2.2: Frequent Consumption of High Caloric Food vs Obesity Status

	Not Obese	Obese
no	211	18
yes	819	944

Table 2.2 provides similar information as 2.1, but is not a genetic related factor. From the table, we see that 53.5% of people who consume high caloric food frequently are almost as those with a family history.

Figure 2.7: Boxplots of BMI by Whether or Not Person Frequently Consumes High Caloric Food

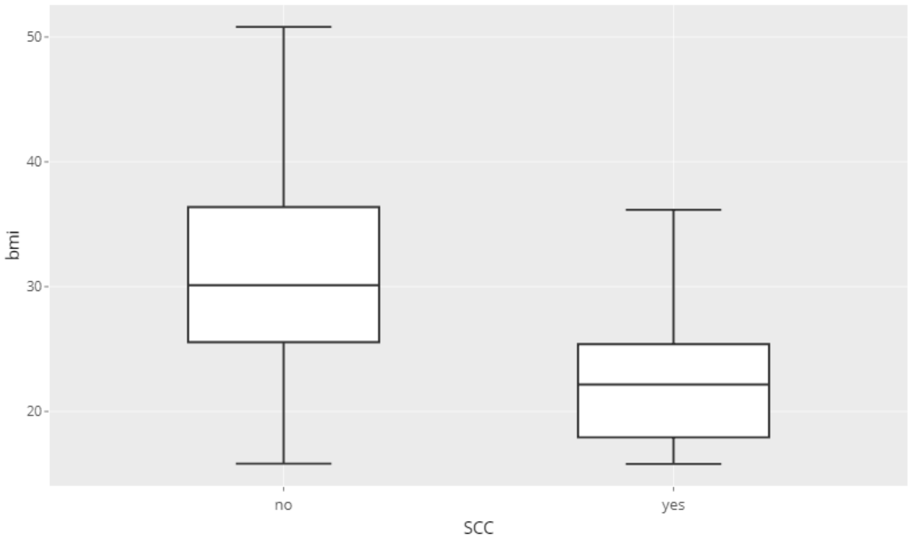


From Figure 2.7, those who frequently consume high caloric food have a median BMI of 31.15 in comparison to the 24.39 median BMI of those who do not. The results from the table and boxplot fall in line with the general intuition that high caloric foods would have an impact on obesity prediction and the difference in median between the BMIs of both groups suggest that frequent consumption of high caloric food could be a strong predictor of obesity.

Table 2.3: Daily Calorie Consumption Monitoring vs Obesity Status

	Not Obese	Obese
no	955	959
yes	75	3

Figure 2.8: Boxplots of BMI by Daily Calorie Consumption Monitoring



Although Table 2.3 demonstrates that there is a sparse number of observations in the data set of people who monitor their daily calorie consumption, it is still a major finding that 96.15% of the observations that monitored their calories were not obese. Generally speaking, tracking and monitoring calories is not something that is commonly done by the general population and the people that do monitor their calories tend to do so for specific reasons such as weight gain, weight loss, or bodybuilding. It is important to emphasize that correlation is not causation as tracking calories is likely linked to many other aspects in one's lifestyle that can influence obesity status. Nonetheless, from Figure 2.8, there is once again a clear difference in medians where those who monitored their calorie consumption had a median BMI of 22.15 in comparison to a 30.11 BMI for those who did not track BMI.

Figure 2.9: Variable Importance of Baseline Model by Mean Decrease Gini

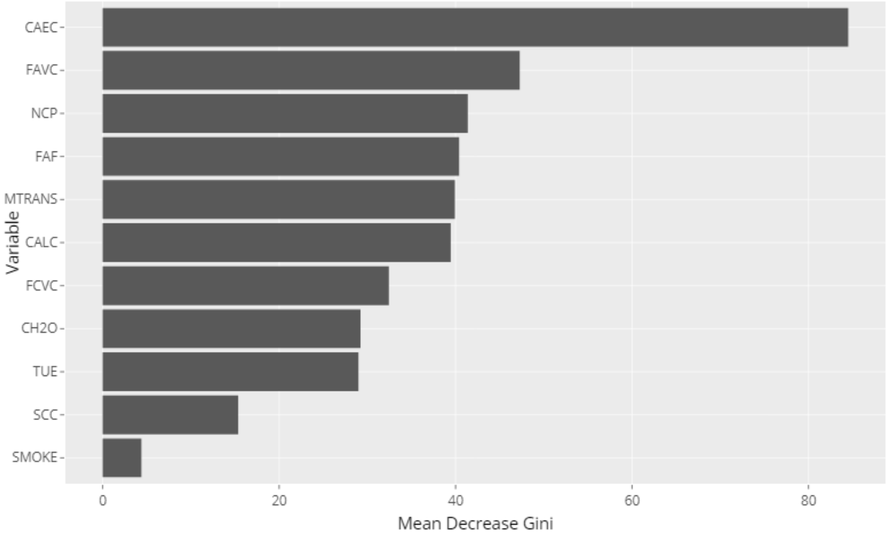
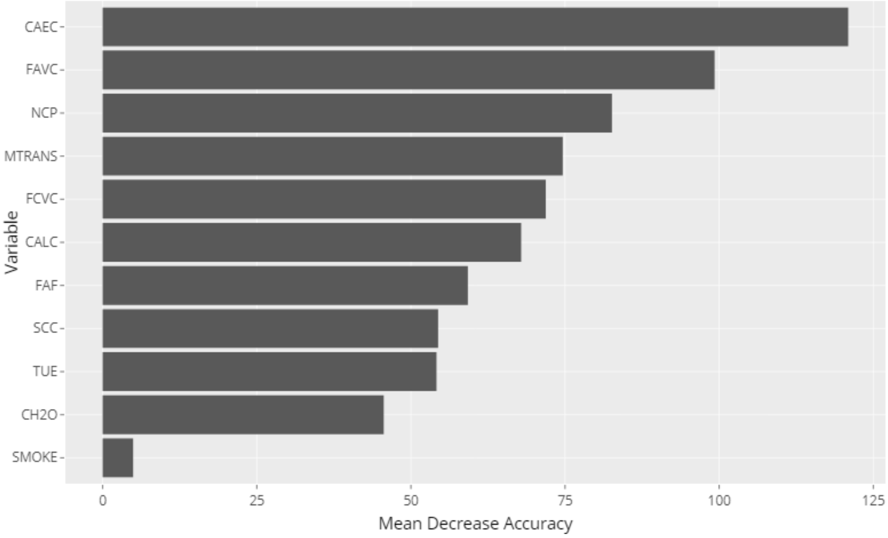


Figure 2.10: Variable Importance of Baseline Model by Mean Decrease Accuracy



To conclude the exploratory data analysis section, a baseline random forest model is developed to objectively determine the variable importance for obesity prediction, as shown in Figure 2.9. The variable importance scores or Mean Decrease Gini values align with the key takeaways from the earlier analysis. For example, the second most important variable with respect to Mean Decrease Gini is FAVC, indicating the frequency of consumption of high calorie foods has a major impact on obesity prediction. However, the variable CAEC, which represents how often a person eats between meals, surpasses all other variables in terms of importance in regard to the Mean Decrease Gini.

CHAPTER 3

Methodology

3.1 Logistic Regression

Logistic regression falls into the classification group of supervised machine learning, where in comparison to linear regression, logistic regression is used to predict a binary or categorical outcome. In the case of this research, logistic regression is used to predict whether or not a person is or is not obese. In comparison to logistic regression, if the objective of the study was to predict BMI, then multiple linear regression would be used instead since BMI is a continuous variable. Although logistic regression does technically predict the class of an observation it is important to note that it predicts the probability of the outcome which is furthermore what is used in predicting an outcome or in other words the likelihood. For instance, through logistic regression we can predict that the probability of an observation being obese is, for example, 55% and use that to classify an observation as obese depending on the threshold selection, typically being 50%.

The hypothesis for logistic regression can be defined with the formula:

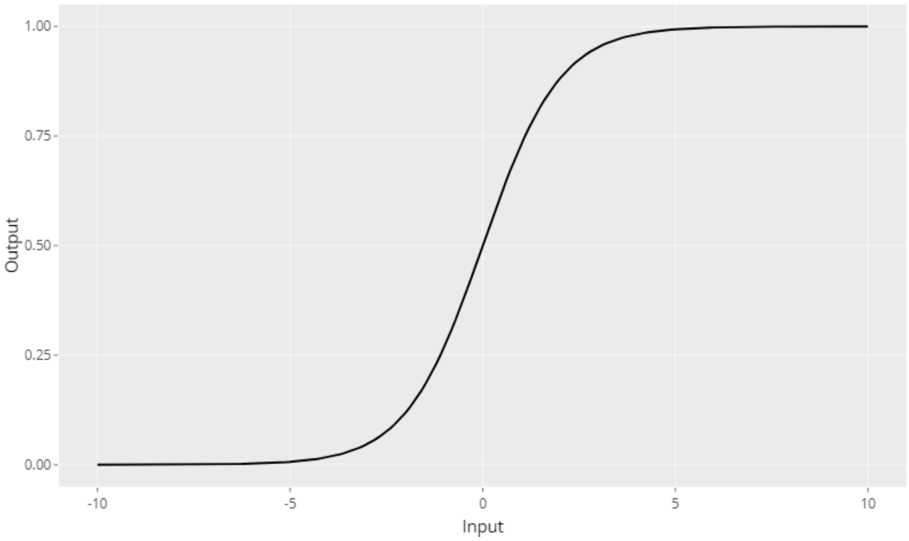
$$h_\theta(x) = g(\theta^T x) \tag{3.1}$$

And the function g is defined by the sigmoid function :

$$g(z) = \frac{1}{1 + e^{-z}} \tag{3.2}$$

Ultimately, the sigmoid function serves to calculate the probability of the outcome of the logistic regression and therefore results in values ranging from 0 to 1 as shown below in Figure 3.1 [Nas17].

Figure 3.1: Visual of Sigmoid Function Plotted



In the process of creating a logistic regression model, our study will begin with a baseline model that includes every single eligible variable similar to the approach in Figure 2.9.

After the baseline model is created, we will perform variable selection where the baseline random forest will be referenced. The mean decrease Gini and mean decrease accuracy will be referenced and each variable will be ranked based on importance from the random forest model. Once ranked, a backwards selection style approach will be used as we will iteratively go through the model removing variables based on importance and assessing model performance. For each iteration of each method, the training accuracy and testing accuracy will be logged for different levels of probability in regard to classification. For instance, by default a probability of 50% or higher will classify an observation as a person who is obese, but various probabilities will be tested ranging from 40% to 60% in case a certain threshold of probability is better than others in predicting obesity. Furthermore, the log odds at each probability threshold will be evaluated to see how the odds change. The most robust model will be selected and evaluated against the other top performing models in the selection of the best one.

3.2 Random Forest

In comparison to logistic regression, random forest is a supervised learning method that is used for both classification and regression where random forest is an ensemble learning method that uses a large number of decision trees leading to a reduction in variance when compared to using individual decision trees [CPB18]. The random forest implements bootstrapping on the training data as a means to enhance variation and furthermore randomizes a subset of the variables used within a tree to create less correlation amongst the trees [CPB18]. The final classification of the observation selected by the random forest is determined via a majority vote of the trees in the random forest.

The general methodology that will be used for selecting the best random forest model will be centered around the baseline model that was created in the exploratory data analysis earlier. From that baseline model, we have established the mean decrease Gini and the mean decrease accuracy which will be used in dropping off variables one by one until the model with the best training and testing accuracy is selected without the use of any hyperparameters. After the best subset of variables is selected, we move on and adjust the hyperparameters. Although "ntree" or the number of trees is not necessarily a hyperparameter, it will be adjusted to be large enough so the features that are randomly selected have adequate opportunities to be selected [CPB18]. The next hyperparameter is "mtry" which by default is set to $\sqrt{p}$ where p is the number of features in the data set. Since there are relatively few features in the data set, every value of p will be tested. Lastly, "nodesize" which signifies the minimum size of the terminal nodes will be adjusted depending on whether or not improvements occur when the baseline value 1 is changed [CPB18]. Similar to logistic regression, the most robust model will be selected and evaluated against the other models.

3.3 Gradient Boosting

Another one of the more popular and powerful machine learning methods within machine learning is known as Gradient Boosting, an ensemble method that utilizes the creations of weak models and furthermore combine them together in order to create an overall better performing model [Mas22]. Whereas models such as the random forest “rely on simple averaging of models,” the boosting family of models are based on the idea of adding new models to an ensemble sequentially where in every iteration “a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far” [NK13]. These weak learners in gradient boosting tend to be decision trees, which will be added one at a time and follow a “gradient descent procedure” in order to minimize the loss [Mas22]. The gradient boosting model essentially calculates residuals with respect to the classification previously found, then fits a weak learner, for example trees, to those residuals to which a new learner is now built via predicting the loss at every iteration [Mas22]. This iterative process will continue until the error falls below a specific pre-decided threshold.

3.4 XGBoost

The next and final model that will be assessed is XGBoost. XGBoost is similar to gradient boosting in the sense that it has “gradient boosting at its core” according to, however it is more regularized since it “uses advanced regularization (L1 & L2), which improves model generalization capabilities” [RUS16, Kha20]. Similar to the previous methods, XGBoost will also have a baseline model that will be hypertuned after variable selection is complete in order to maximize performance.

CHAPTER 4

Model Analysis

To evaluate our models, our data set will be split into 70% training and 30% testing at random meaning that of our original 1,992 observation data set will be broken out as 1,394 observations for training and 598 observations for testing. The models that will be tested correspond with the previous chapter and will be Logistic Regression, Random Forest, Gradient Boosting and XGBoost. The most robust model using each supervised learning method will be evaluated in its respective section.

In the following sections, a confusion matrix will be displayed for each model to provide a visualization in regard to the breakout or overall distribution of predictions for each model, various metrics in regard to the confusion matrix, and the AUC score. In the confusion matrix, we can see how many true positives, true negatives, false positives, and false negatives there are. Furthermore, with this any disproportionality will be evident as the confusion matrix will show us the precision and recall of our model and if any bias in classification is prevalent.

The "accuracy" of a model measures the number of correct predictions out of the total number of predictions.

$$Accuracy \;=\; \frac{Number\ of\ Correct\ Predictions}{Number\ of\ Predictions} \tag{4.1}$$

Although accuracy is in a sense the gold standard in regard to model success it is important to look one layer deeper to fully understand where our models succeed and where they fail. This is especially significant when dealing with imbalanced data such that the variable you are classifying is split, for example, 20% obese and 80% not obese. If the data were to be imbalanced to this degree, a model could classify every single observation to be not obese and it would have an accuracy of 80% which is high, but misleading. Although our data is balanced we will still look at both precision, recall, and F1 score in addition to accuracy.

$$Precision \ = \ \frac{Number \ of \ True \ Positives}{Number \ of \ True \ Positives \ + \ Number \ of \ False \ Positives} \tag{4.2}$$

$$Recall \ = \ \frac{Number \ of \ True \ Positives}{Number \ of \ True \ Positives \ + \ Number \ of \ False \ Negatives} \tag{4.3}$$

$$F1 \ Score \ = \ 2 \ * \ \frac{Precision \ * \ Recall}{Precision \ + \ Recall} \tag{4.4}$$

Furthermore, the models that reference the mean decrease Gini and mean decrease accuracy plots of our baseline model will be assessing the variables using the following hierarchy:

1. CAEC

2. FAVC

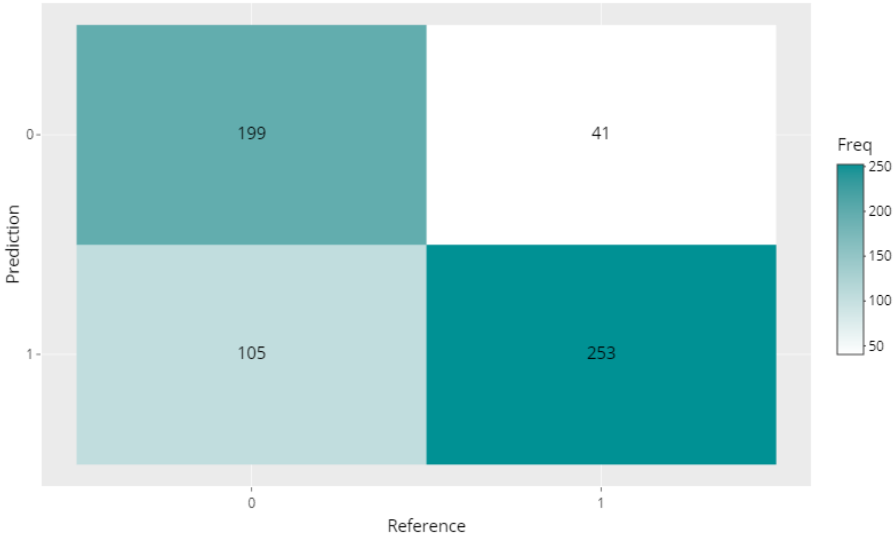
3. NCP

4. MTRANS
5. FCVC
6. CALC
7. FAF
8. SCC
9. TUE
10. CH2O
11. SMOKE

4.1 Logistic Regression

In constructing our logistic regression model we firstly begin by creating a baseline model that includes every single variable and then removing variables one by one with respect to the importance plots from our random forest baseline model that was part of the exploratory data analysis. Furthermore, the training accuracy and testing accuracy were logged for each iteration to determine the best model on the basis of accuracy. From this we found that the best performing model was the original one that included every single variable which hoists an accuracy of 75.6% for both the training and the test data. The close proximity in training and test accuracy signifies that overfitting is likely not an issue for our model. Once this model was selected, several thresholds were tested in regard to prediction classification. As mentioned earlier, logistic regression gives the probability that an observation belongs to a class and by default the threshold is set to 50%. We next re-ran our prediction for every whole number threshold from 20% to 80%. Predictions using values less than 50% would be more conservative in labeling observations as obese and predictions using values greater than 50% would be more aggressive in labeling observations as obese. This process confirmed that the standard 50% yielded the highest test accuracy while still having a high precision, recall, and the top F1 score with respect to the other thresholds.

Figure 4.1: Logistic Regression Confusion Matrix (0 = Not Obese, 1 = Obese)



The final results of our logistic regression model are:

- Training Accuracy: 75.6%
- Testing Accuracy: 75.6%
- Precision: 70.7%
- Recall: 86.1%
- F1 Score: 77.6%
- AUC Score: 0.7576

Table 4.1: Odds-Ratio of Significant Predictors (p-value ≤ 0.05)

	Variable	Odds-Ratio
1	FAVC1	10.61
2	NCP1	2.05
3	NCP2	0.25
4	CAEC1	14.29
5	CH2O2	1.96
6	SCC1	0.05
7	FAF3	0.24
8	TUE1	0.74
9	TUE2	0.43
10	CALC2	0.38
11	MTRANS1	5.59
12	MTRANS2	7.30

In Table 4.1 above, we see the Odds-Ratios of our significant predictors. Their interpretations are as followed:

- FAVC1 - Individuals that frequently consume high caloric food have 10.61 times the odds of being obese than those who do not
- NCP1 - Individuals that consume 2-3 meals per day have 2.05 times the odds of being obese than those who consume 1
- NCP2 - Individuals that consume 3+ meals per day have 4 times lower odds of being obese than those who consume 1
- CAEC1 - Individuals that sometimes consume food between their meals have 14.29 times the odds of being obese than those who do not
- CH2O2 - Individuals that drink over 2 liters of water a day have 1.96 times the odds of being obese than those who drink less than 1 liter
- SCC1 - Individuals that monitor their daily calorie consumption have 20 times lower odds of being obese than those who do not
- FAF3 - Individuals that are physically active 4-5 times per week have approximately 4.17 lower odds of being obese than those who are not physically active
- TUE1 - Individuals that spend 3 to 5 hours a day using technology devices have 1.35 lower odds of being obese than those who spend 0 to 2 hours
- TUE2 - Individuals that spend 5+ hours a day using technology devices have 2.33 lower odds of being obese than those who spend 0 to 2 hours

- CALC2 - Individuals that frequently consume alcohol have 2.63 lower odds of being obese than those who do not consume any
- MTRANS1 - Individuals that usually take an automobile as transportation have 5.59 times the odds of being obese than those who walk or bike
- MTRANS2 - Individuals that usually take public transportation have 7.30 times the odds of being obese than those who walk or bike

4.2 Random Forest

Similar to our logistic regression model, we begin fitting our random forest model by referencing the random forest baseline model created during our exploratory data analysis. Using the mean decrease Gini and mean decrease accuracy variables will be removed one by one and the training and testing accuracy will be logged for model comparison. In comparison to the logistic regression model, the random forest model showed improvement when removing the least important variable SMOKE from the model. Using a baseline of 1200 trees, once SMOKE was removed the training accuracy improved from 87.7% to 87.8% and the testing accuracy increased from 82.1% acccuracy to to 82.9% accuracy. Due to the 0.8% increase in testing accuracy and marginal increase in training accuracy as well as model simplification, we opt to omit SMOKE from the random forest model.

In the next step the parameters ntree, mtry, and nodesize were tuned. In the process of tuning these, for loops were created for ntree, mtry, and nodesize in that order where several for each parameter were run to find which produces the best combination of training and test accuracy. This process found the 800 trees that were used for our baseline model to provide the best combination of training and test accuracy. It found that the second best mtry value, 10, increased the training accuracy from 87.8% to 90.0%, but decreased the testing accuracy from 82.9% to 82.6% meaning it is likely to be a victim of overfitting due to the large increase in training accuracy and a roughly stagnant testing accuracy. As a result, we will stand with the default value for mtry set by the random forest function. Lastly, the process was repeated for nodesize to which the default nodesize was shown to be the most promising, leaving the training and testing accuracy unchanged. Our final random forest model includes all variables except SMOKE, uses an ntree of 1200, and the remaining tuning factors are untouched.

Figure 4.2: Mean Decrease Gini of Final Random Forest Model

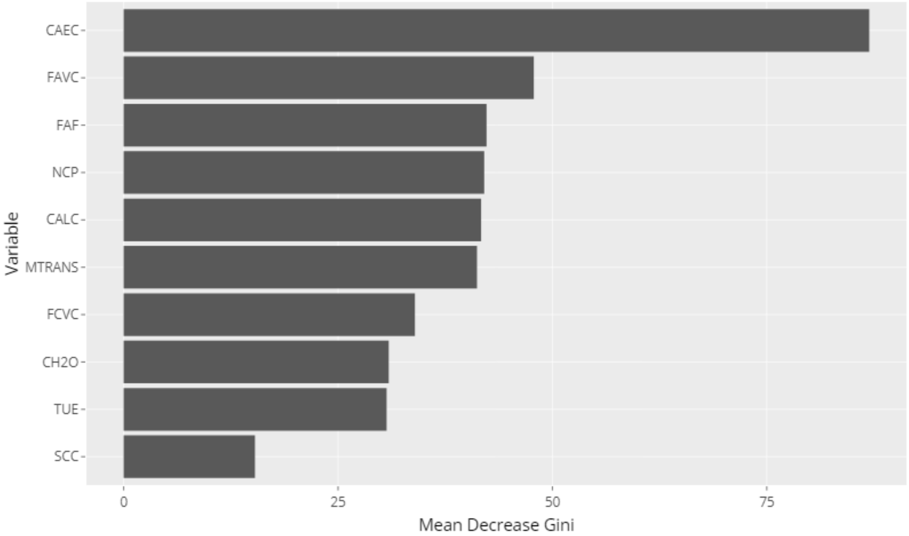


Figure 4.3: Mean Decrease Accuracy of Final Random Forest Mode

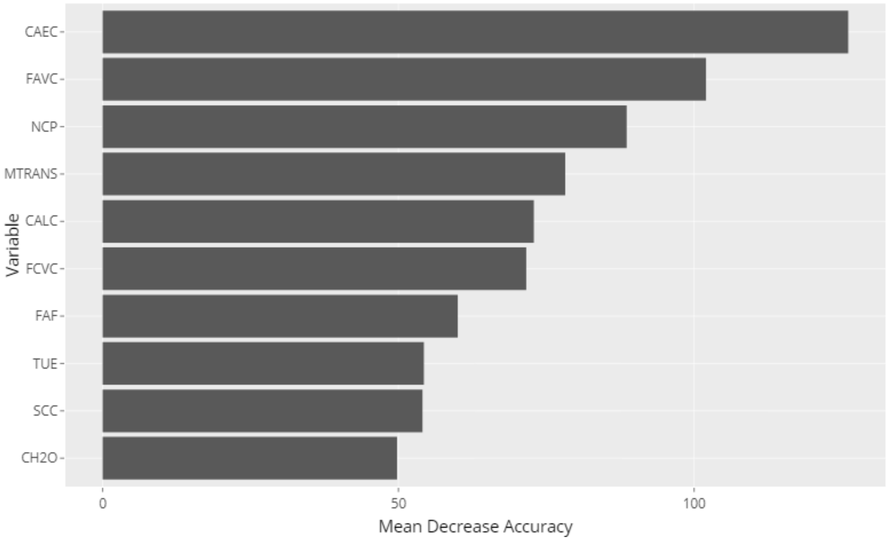
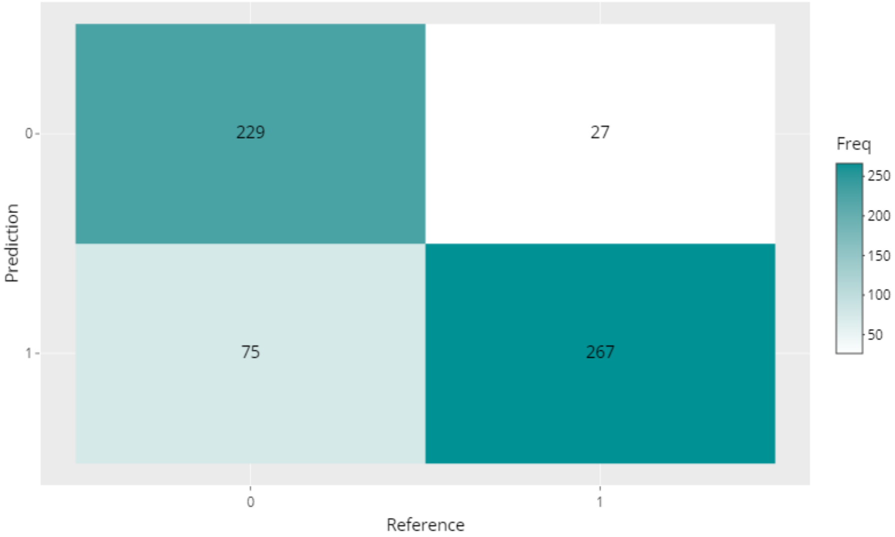


Figure 4.4: Random Forest Confusion Matrix (0 = Not Obese, 1 = Obese)



The final results of our random forest model are:

- Training Accuracy: 87.8%
- Testing Accuracy: 82.9%
- Precision: 78.1%
- Recall: 90.8%
- F1 Score: 84.0%
- AUC Score: 0.8307

4.3 Gradient Boosting

The first step in creating the Gradient Boosting model was to set up a well-performing baseline model including all variables in order to get an output of the importance plot to understand how the model uses the variables available. After that, like all the other models, variables were redacted from the model one by one and accuracies were logged in order to choose a viable variable subset for our model. From the importance plots it was evident that SCC and SMOKE had the least relative influence of all variables so those were the first investigated, but upon removal of variables the model only digressed and never showed improvement so we stuck with the original model. With the variables now selected, the next order of business was to tune the parameters: shrinkage, n.trees, and interaction.depth to which it was found that a shrinkage of 0.1, interaction.depth of 10, and n.trees of 1,500 yielded the best model. The final model has a training accuracy of 90.2% and a testing accuracy of 83.6%.

The final results of our gradient boosting model are:

Figure 4.5: Importance Plot of Final Gradient Boosting Model

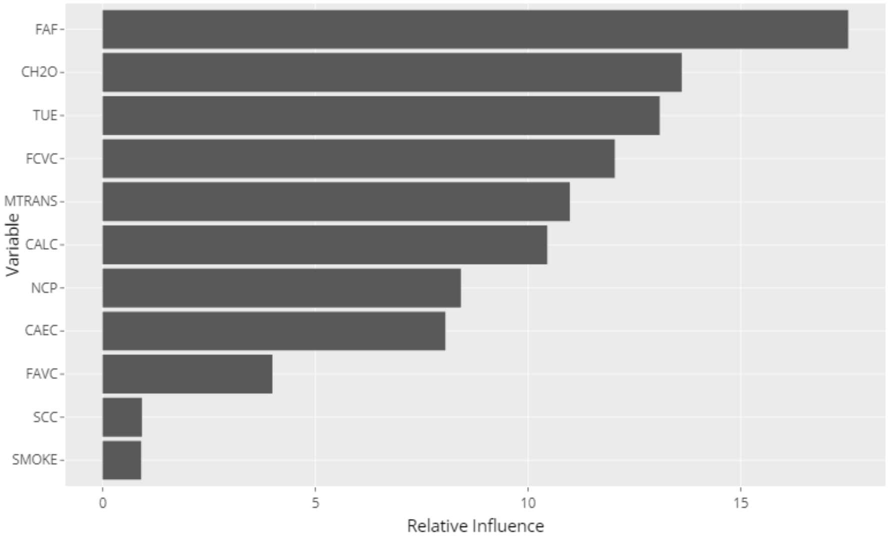
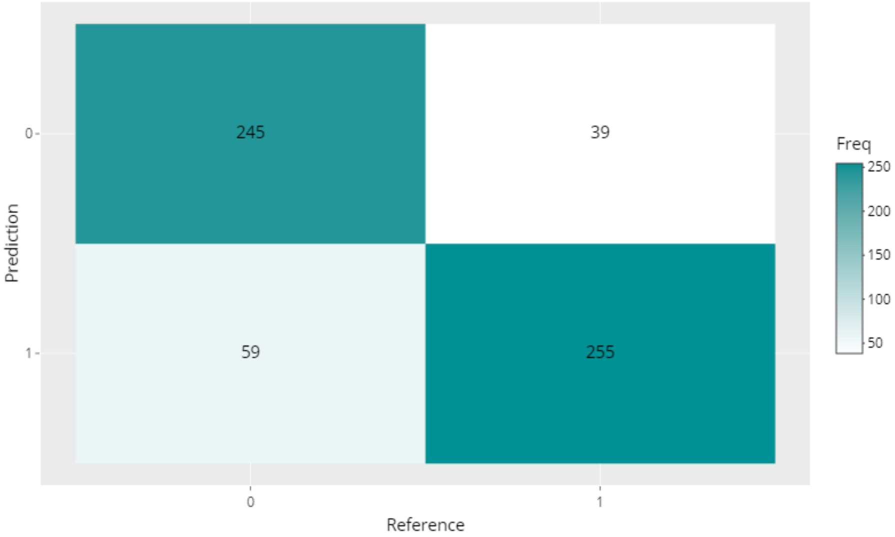


Figure 4.6: Gradient Boosting Confusion Matrix (0 = Not Obese, 1 = Obese)



- Training Accuracy: 90.2%
- Testing Accuracy: 83.6%
- Precision: 81.2%
- Recall: 86.7%
- F1 Score: 83.9%
- AUC Score: 0.8366

4.4 XGBoost

Our process of model selection for the XGBoost model falls in line with the process utilized for the random forest model except we created an importance plot from a baseline XGBoost model as a reference for variable selection rather than using the random forest baseline mean decrease plots. With this it was found that the model with every single variable was the best performer out of all subsets. After that was found, focus was then shifted toward tuning the hyperparameters eta, iterations, and maximum tree depth in order to find the absolute best model. In order to find the best subset of hyperparameters several models were created and tested. The final XGBoost model selected uses an eta of 0.11, max depth of 8, and 1,000 rounds.

Figure 4.7: Importance Plot of Final XGBoost Model

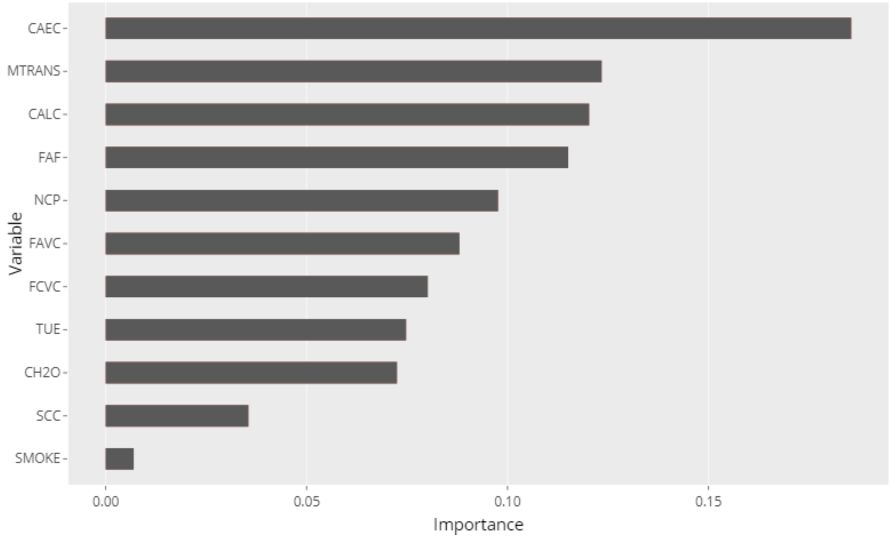
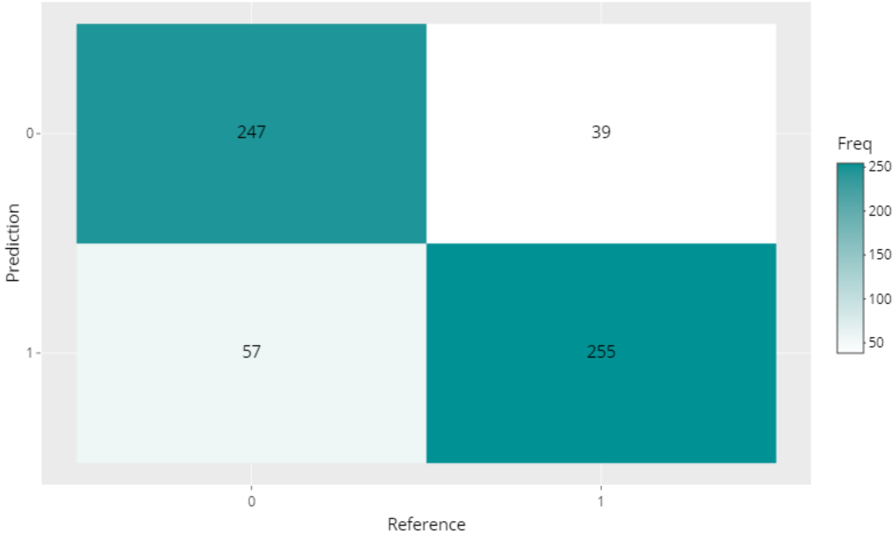


Figure 4.8: XGBoost Confusion Matrix (0 = Not Obese, 1 = Obese)



The final results of our XGBoost model are:

- Training Accuracy: 90.3%
- Testing Accuracy: 83.9%
- Precision: 81.7%
- Recall: 86.7%
- F1 Score: 84.2%
- AUC Score: 0.8399

4.5 Model Comparison

Table 4.2: Model Summary

Model	Training Acc.	Testing Acc.	Precision	Recall	F1 Score	AUC Score
Logistic Regression	75.6%	75.6%	70.7%	86.1%	77.7%	0.7576
Random Forest	87.8%	82.9%	78.1%	90.8%	84.0%	0.8307
Gradient Boosting	90.2%	83.6%	81.2%	86.7%	83.9%	0.8366
XGBoost	90.3%	83.9%	81.7%	86.7%	84.2%	0.8399

The models ranked by testing accuracy are XGBoost (83.9%), Gradient Boosting (83.6%), Random Forest (82.9%), and Logistic Regression (75.6%). In regard to using the models for actual prediction, it is evident that there is a clear disparity between logistic regression and the other models as logistic regression trails 7.3% behind the third best model random forest. On the other hand, the top three models are all separated from each other by 1% or less meaning they all performed vert close.

When it comes to precision, XGBoost ranks first at 81.7% ahead of gradient boosting, 81.2% and random forest 78.1%. However, random forest tanks first in recall at 90.8% with gradient boosting and XGBoost tied at 86.7%. When you take into account F1 score which is a form of combination of precision and recall XGBoost is at the top again with 84.2%, trailed by random forest at 84.0%, and gradient boosting at 83.9%. Needless to say all three of the models perform fairly close outside of recall. The major difference seen between the top three models would be the training accuracy which is what stands out for random forest. All three of the models perform roughly on par in regard to F1 score and testing accuracy, but random forest's training accuracy is 87.8% in comparison to gradient boosting's 90.2% and XGBoost's 90.3% allowing us to assume that although random forest performs marginally worse on the testing data, it is likely to be me less susceptible to overfitting due to the lower training accuracy percentage.

Although our logistic regression did not perform on par with the other models, it still provided plentiful information and context to our main objective of investigating predictors of obesity. For instance, from the odds-ratio of the model, it was found that individuals that frequently consume high caloric food have 10.61 times the odds of being obese than those who do not and individuals that sometimes consume food between their meals have 14.29 times the odds of being obese than those who do not. In contrast, what stood out the most was individuals that monitor their daily calories consumption have 20 times lower odds of being obese than those who do not. Ultimately, logistic regression did not perform the best, but in regard to interpretability and adding context to our data rather than predictability it was far more powerful.

In regard to predictability, the other three models were all similar, but had slightly different importance values and rankings for the variables. Random forest had CAEC and FAVC at the top with SCC, CH2O, and TUE at the bottom. Of these, CAEC or consumption of food between meals and FAVC or frequent consumption of high caloric food were ranked at the top of importance with CAEC in particular sticking out the most. When looking at the gradient boosting model FAF or physical activity frequency is the most important with CH2O being the second most important. For reference, CH2O or daily water consumption was one of the least important variables for random forest. In addition, FAVC is also ranked third to last in the gradient boosting model with SCC (calories consumption monitoring) and SMOKE (smoker or not) being ranked almost identical at last. Similar to random forest, XGBOOST ranks CAEC at the top of importants by a large margin as well with SCC and SMOKE falling at the bottom. Note that SMOKE was also at the bottom of random forest, but was removed since that helped improve the model.

CHAPTER 5

Limitations and Conclusion

5.1 Limitations

When conducting a study centered around a subject such as obesity and its prediction, it is essential to understand the limitations, shortcomings, and disclaimers in regard to the overall study. One of the major parts that should be recognized, is that due to the nature of the study and its primary focus of predictability it does not establish any causal claims or relationships between any of the predictors and obesity. The predictive models generated within our research seeks to find and identify patterns based on the data available from the UCI Machine Learning Repository and by no means is intended to demonstrate or showcase causation. On the topic of the data, the models are reliant on the data set that is provided and is built around that particular sample of data. In the case of our data set, the data is centered around individuals from Colombia, Peru, and Mexico and the models created with respect to this data set have not been tested against different regions of the world, so we are unaware of how much diversity the models cover. In addition, as mentioned early in the study, obesity is a multifaceted condition that is impacted by both genetics and one's environment and it is important to note that the environment changes. As time progresses societal norms will change, technology will advance, lifestyles will change and with that in mind the model is only as relevant as the data.

One of the major shortcomings of the study is the use of BMI to identify obesity. There are cases regarding people who weightlift where a person may have a BMI over 30, but may not actually be obese due to muscle being the reason for their weight and not actual fat. In the future, using a study that uses body fat percentage instead of BMI to assess obesity would be interesting in order to compare the most robust model of each and see how the predictors shift overall and which one works best on average.

5.2 Conclusion

Machine learning and its models continually allow for a deeper understanding in topics regardless of the subject. In this paper, the application of machine learning was leveraged to determine the predictability of obesity given various lifestyle factors of people and investigate the variables affiliated. The study began with an exploratory data analysis to reveal potential patterns or relations between the variables present and the classification of obesity. With the process of exploratory data analysis completed a simple random forest model was fit with every variable available in order to establish a baseline understanding of what to expect the importance of variables to look like.

The analysis then fully shifted toward the supervised learning methods logistic regression, random forest, gradient boosting, and XGBoost with all models aside from logistic regression fully focused on predictability rather than interpretability and variable understanding. Logistic regression does not have the predicting power of ensemble methods, but the odds-ratio from it provided essential knowledge as to how the category the predictors fall in impact obesity prediction. For instance, the log odds from the logistic regression tells that individuals that usually take an automobile as transportation have 5.59 times the odds of being obese than those who walk or bike and individuals that take public transport have 7.30 times the odds than those who walk or bike. Without interpretability in mind, the models in the study were found to be well at predicting obesity with one model, XGBoost, predicting obesity at an accuracy of 83.9%. Bare in mind that for reference the issue of obesity is multifaceted as 50% of the variance in weight is explained by genetics and the other half by their environment meaning that prior to conducting the study the main objective was to get as close to 50% accuracy as possible whereas the actual model exceeded expectations by about 34% [BLM15].

REFERENCES

[BLM15] Molly S. Bray, Ruth J.F. Loos, Jeanne M. McCaffery, Charlotte Ling, Paul W. Franks, George M. Weinstock, Michael P. Snyder, Jason L. Vassy, and Tanya Agurs-Collins. "NIH working group report—using genomic information to guide weight management: From universal to precision treatment." *Obesity*, **24**(1):14–22, 2015.

[CDC22a] CDC. "Defining Adult Overweight and Obesity.", Jun 2022.

[CDC22b] CDC. "Obesity is a Common, Serious, and Costly Disease.", Jul 2022.

[CPB18] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. "Random forest versus logistic regression: a large-scale benchmark experiment." *BMC Bioinformatics*, **19**(1):1–14, Jul 2018.

[Kha20] Neetika Khandelwal. "A Brief Introduction to XGBoost." *Towards Data Science*, Jul 2020.

[LB03] R. J. F. Loos and C. Bouchard. "Obesity - is it a genetic disorder?" *Journal of Internal Medicine*, **254**(5):401–425, 2003.

[Mas22] Tomonori Masui. "All You Need to Know about Gradient Boosting Algorithm - Part 1. Regression." *Towards Data Science*, Feb 2022.

[Nas17] Vladimir Nasteski. "An overview of the supervised machine learning methods." *Horizons. b*, **4**:51–62, 2017.

[NK13] A Natekin and A Knoll. "Gradient boosting machines, a tutorial. 7 (December).", 2013.

[RUS16] Santhanam Ramraj, Nishant Uzir, R Sunil, and Shatadeep Banerjee. "Experimenting XGBoost algorithm for prediction and classification of different datasets." *International Journal of Control Theory and Applications*, **9**(40):651–662, 2016.