Stats 101C Final Project

Predicting Obesity Status

Sarina Doss, Sukie Yeung, Jordyn Fuchs, Maia Smolyanov

Abstract

For this Kaggle project, the overall goal is to utilize various statistical models in order to predict an individual's obesity status based on the given Obesity data set. This report details the steps of model creation, including an introduction to the problem, exploratory data analysis, cleaning missing values, selecting important features, model construction, analysis of results, and limitations to improve on in the next steps.

The final model is trained using a Random Forest approach, incorporating the following variables: Height, Race, Physical Activity Frequency (FAF), Number of Main Meals (NCP), Daily Water Intake (CH2O), Caloric Intake (CALC), Frequency of Vegetable Consumption (FCVC), Mode of Transportation (MTRANS), Smoking Habits (SMOKE), Time Using Technology Devices (TUE), Consumption of Food Between Meals (CAEC), Family History with Overweight, Frequent Consumption of High-Calorie Foods (FAVC), and Gender.

The resulting model has an accuracy of 99.615% and a 0.385% misclassification rate on the testing data set, resulting in an 11th place ranking on Kaggle.

1. Introduction

Obesity is a medical condition characterized by excess body fat and a body mass index (BMI) over 30, and is quickly escalating to become a national health crisis. One in three adult Americans is classified as overweight, while nearly two in five are classified as either obese or severely obese. However, obesity is not merely a matter of excess weight or body image; it is linked to a range of severe health issues that can have devastating consequences, including Type 2 diabetes, high blood pressure, heart disease, strokes, joint problems, liver disease, gallstones, respiratory challenges, and many more (NIH).

In 2019 alone, obesity-related medical expenditures reached approximately $173 billion, underscoring its pervasiveness among Americans. The main factors influencing obesity can be classified into environmental, societal, and genetic. In particular, demographic data indicates that race plays a crucial role in obesity prevalence, with Hispanic adults and non-Hispanic Black adults being the most affected groups at rates of 44.8% and 49.6%, respectively.

This project works with the Obesity dataset, which contains data from multiple resources, including Kaggle and Centers for Disease Control and Prevention (CDC). The dataset comes split with a training data set of 32,014 observations and testing dataset of 10,672 observations. The dataset consists of 29 numeric and categorical variables spanning lifestyle factors and demographic information that influence the likelihood of obesity in an individual. At a glance, some numeric variables include age, frequency of vegetable consumption and resting blood pressure, while categorical variables include gender, smoking history, and family history. Together, these variables provide us with a strong foundation to build a model that holistically assesses the most crucial predictors of this prevalent condition.

The goal of this project is to develop a model incorporating relevant variables to predict an individual's obesity status – either Obese or Not Obese.

2. Data Analysis

A. Investigating missing values – NA Values

Our exploratory data analysis first examines missing values in the data set. Figure 1 illustrates the proportion of missing data for every variable, which remains relatively low at approximately 8%. Due to the relatively small amount of missing data, we attempted to remove all rows with missing (NA) values. However, we soon found that while each variable has few

missing values, removing all rows with a missing value in any column led to the removal of almost 90% of our observations.

Specifically, this approach led to the removal of 74,274 values from the training data set and 24,759 values from the testing data set. We decided against this approach, and chose to take an imputation approach that allows us to preserve as much data as possible while still addressing the issue of missing values. For categorical variables, we replaced NA values with the mode of the variable and for numerical variables, we replaced missing values with the mean of the respective column. While developing the model, we also experimented with replacing NA values in numeric columns with the median instead of the mean, and found that this improved the overall results. This approach preserves the integrity of our dataset and strikes a balance between data preservation and accuracy.

Age	0.078716
Gender	0.080340
Height	0.079278
family_history_with_overweight	0.079559
FAVC	0.080559
FCVC	0.080402
NCP	0.080683
CAEC	0.079247
SMOKE	0.081496
CH2O	0.079621
SCC	0.079684
FAF	0.079965
TUE	0.080965
CALC	0.081652
MTRANS	0.077778
Race	0.080840
RestingBP	0.078372
Cholesterol	0.079840
FastingBS	0.079840
RestingECG	0.081745
MaxHR	0.080777
ExerciseAngina	0.078934
HeartDisease	0.081090
hypertension	0.081683
ever_married	0.079278
work_type	0.083307
Residence_type	0.079653
avg_glucose_level	0.078091
stroke	0.076591
ObStatus	0.000000

Figure 1: Percentage of NA values per variable

B. Variable Selection

After imputing missing values, we focused on selecting the most influential variables to maximize accuracy without overfitting. We developed a baseline Random Forest model with 50 trees and all 29 predictors, which resulted in a starting accuracy of 98.182%. While this accuracy was promising, using all predictors led to a high risk of overfitting, which would limit the generalizability of our model.

We fine-tuned our model to find the optimal number of variables, extracting the importance scores of each variable from our baseline Random Forest model with 100 trees. We experimented with using top 20 predictors (99.390% accuracy), top 10 predictors (98.388%) and top 15 predictors (99.419%). Given our results, using 15 predictors appeared to strike a balance between bias and variance.

The top 15 predictors were Height, Age, Race, Number of Main Meals (NCP), Caloric Intake (CALC), Physical Activity Frequency (FAF), Mode of Transportation (MTRANS), Daily Water Intake (CH2O), Frequency of Vegetable Consumption (FCVC), Consumption of Food Between Meals (CAEC), Family History with Overweight, Gender, Smoking Habits (SMOKE), Frequent Consumption of High-Calorie Foods (FAVC), and Time Using Technology Devices (TUE).

Variable	Importance
Height	39.56709
Age	38.19910
Race	34.68565
NCP	32.56339
CALC	30.40443
FAF	27.24320
MTRANS	24.01084
CH2O	22.77628
FCVC	22.63095
CAEC	22.17939
family_history_with_overweight	20.94001
Gender	19.44257
SMOKE	17.90208
FAVC	17.46655
TUE	16.57910

Figure 2: Top 15 Predictors

C. Number of Trees

The second parameter for the Random Forest is the number of trees. We began with a baseline model comprising 100 trees, which yielded an accuracy of 99.250%. To systematically evaluate the effect of varying tree counts on misclassification rates, we conducted an analysis that is visually represented in Figure 3.

As we increased the number of trees, we observed a trend where the misclassification rate decreased, indicating improved model performance. Specifically, when we expanded the tree count to 200, the model achieved an accuracy of 99.475%, demonstrating a significant reduction in misclassification compared to the baseline model.

To further refine our approach, we also tested configurations with fewer trees. For instance, models with 50 trees and 150 trees were evaluated, revealing accuracies of 98.182% and 99.350%, respectively. This data suggests that while increasing the number of trees generally contributes to enhanced accuracy, diminishing returns may occur beyond a certain point.

In our final optimization step, we replaced NA values in numerical columns with the median instead of the mode. This adjustment led to a notable improvement in model performance, culminating in an accuracy of 99.615% when utilizing 200 trees and the top 15 predictors.

In summary, our analysis indicates that a Random Forest model with 200 trees provides an optimal balance between predictive accuracy and computational efficiency, as evidenced by its superior performance metrics and reduced misclassification rates.

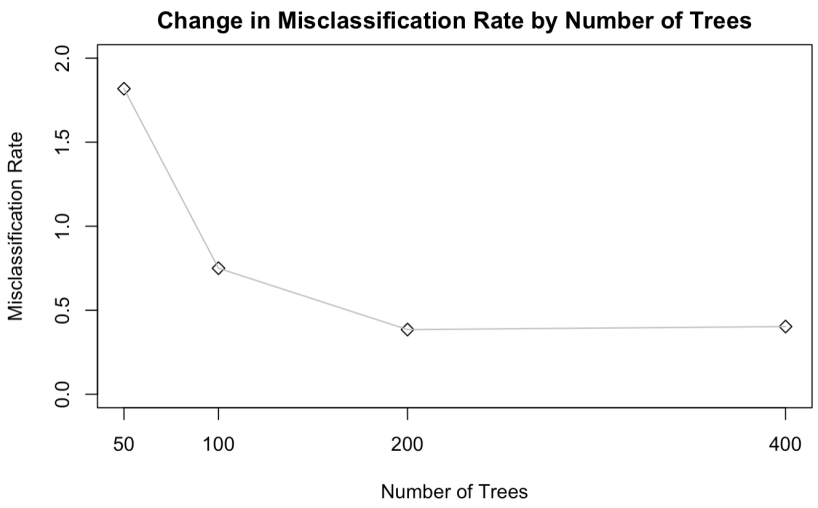

Figure 3: Number of Trees

3. Methods and Models

A. K-Mean Clustering

On our first attempt at creating a predictive model we chose to use a K-Means Clustering model. In this model, we chose to use all the predictors and then test them on this model. In the end, this resulted in only a 43.6% prediction accuracy rating. This is likely because the data points did not appear to be visibly clustered which is not ideal in an unsupervised classification.

B. Gradient Boosting

Taking the K-Mean Clustering we then tried to use a gradient boosting model. In this model, we used all the predictors and 150 estimators. We also used a 0.3 learning rate and a max depth of 9. In the end, this model had a 99.212% prediction accuracy. This means that it predicted most of the obesity statuses correctly.

C. Single Classification Tree

We then tried to see what our accuracy would be by using a Single Classification Tree. In this model we allowed all of the predictors to be used in a single classification tree. In the end, the tree only ended up including CH2O, FAF, Age, Height, and CALC and their branches. This model had only a 77% prediction accuracy with a 22.16% misclassification error rate making it not suitable for us.

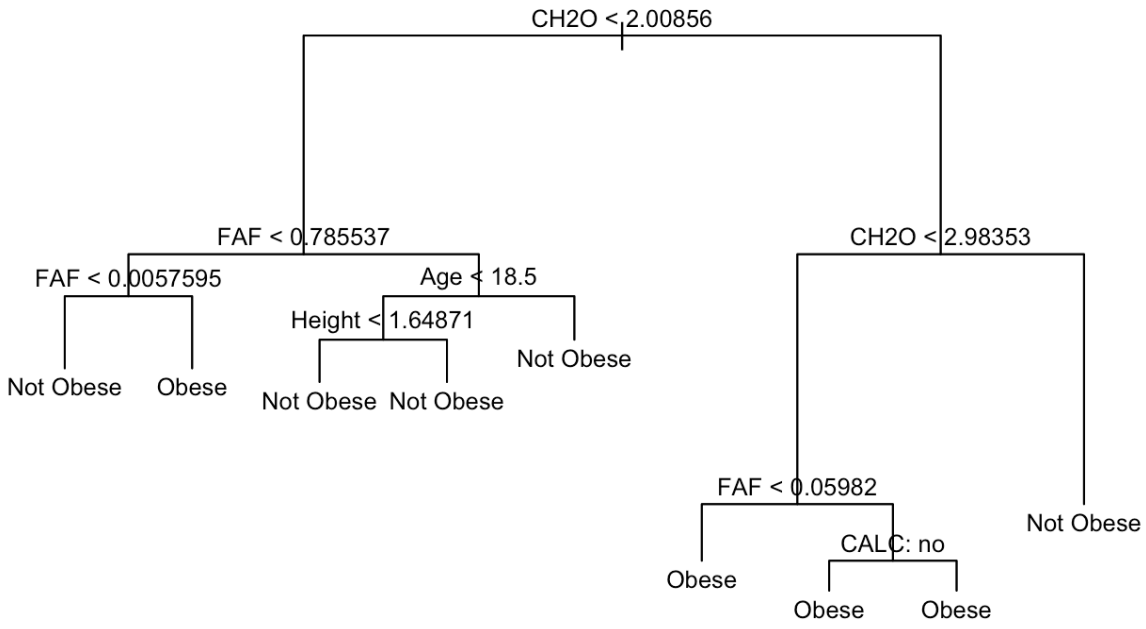

Figure 4: Single Classification Tree

D. Random Forest

Finally we ended up using a Random Forest Model as our final classifier after strategic fine-tuning and adjustment. Originally, we started with a random forest model using all of the predictors (bagging method) and 50 trees. This resulted in an accuracy of 99.250%. With this highly accurate classification rate, we chose to use this as our method of choice. The random forest model is best suited for our dataset because this model works well with high dimensionality, is robust to missing data, handles various data types, and conducts feature selection. This already shows improvement from our Gradient Boosting model, and we continued to tune this model to make it more accurate in predicting obesity status.

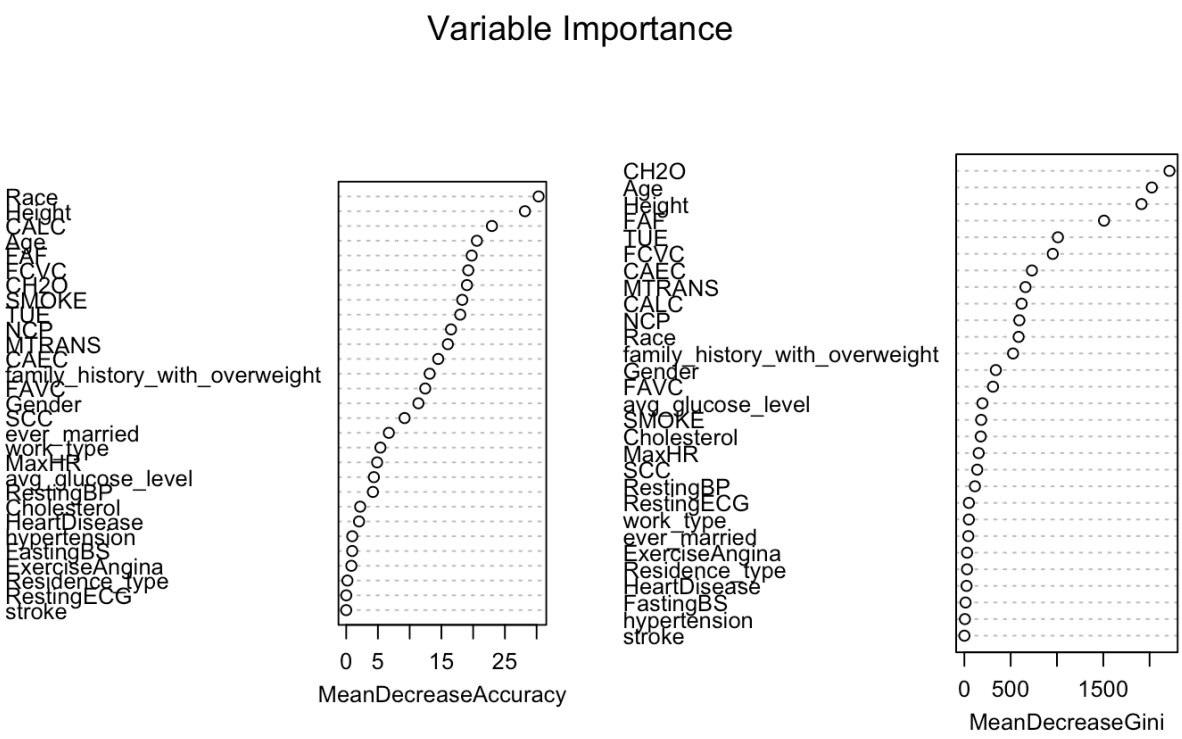

Figure 5: Variable Importance with all Predictors and 50 Trees

Using the top 15 predictors and 200 trees, we obtained a classification accuracy of 99.475%. This was our best accuracy so far, but we continued making adjustments to see if we

could improve any further. We replaced the NA values in the numerical columns with the median instead of the mean, and this resulted in an accuracy of 99.615%. This is the highest accuracy we were able to obtain, and we utilized this model as the final iteration for the project.

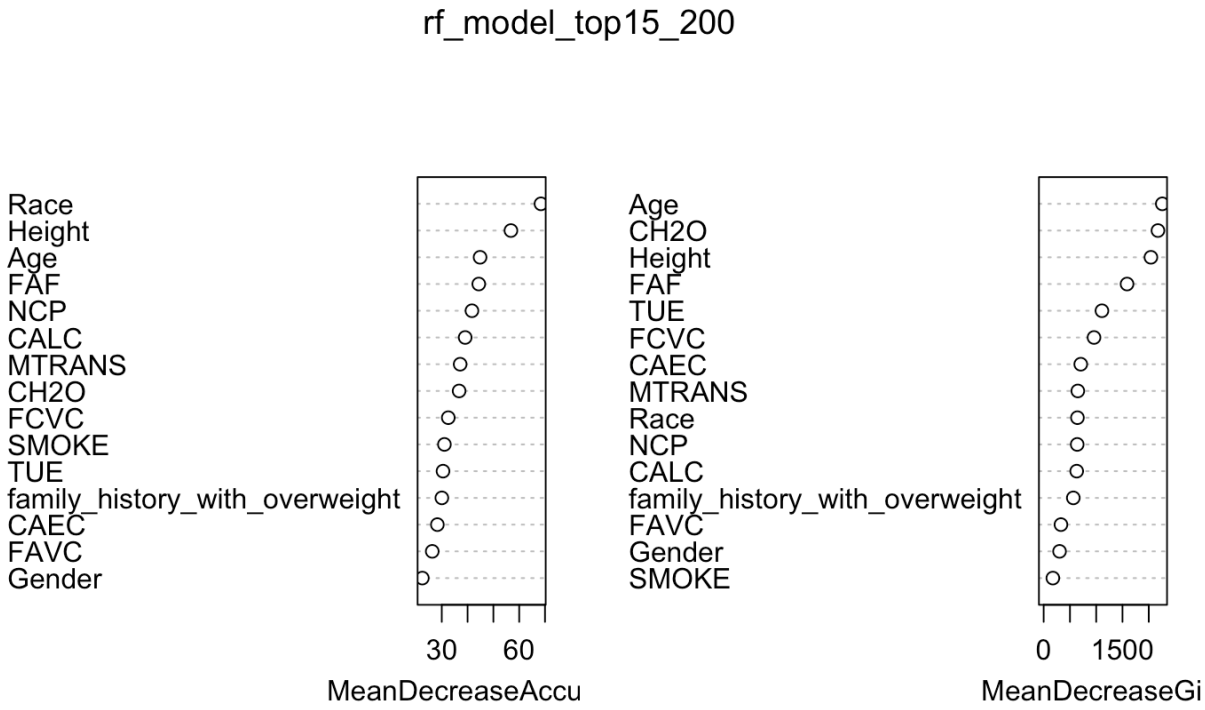

Figure 6: Variable Importance with 15 Predictors and 200 Trees

4. Discussion and Limitations

While we got 11th overall in the Kaggle competition, we acknowledge there are some limitations to our approach that we can address through further exploration and more complete data collection. The first limitation is that there is data set bias due to the limited demographic diversity in the data which could affect the generalizability. To improve on this in our next steps, we could expand the dataset to include a more diverse population. Another limitation that we faced was the missing data handling. While our methods of using the mean, median, and mode

helped address the missing values, these methods still may not fully capture true variability when it comes to the NA values. To improve upon this limitation we could explore more advanced imputation methods such as multiple imputation or machine learning-based techniques. Finally, there is no external validation due to the lack of testing on independent datasets in this project. To improve upon this in the future we could validate the model using an external dataset from different regions and with different populations to assess the generalizability of the model.

5. Conclusion and Recommendations

The development and refinement of our predictive model for obesity classification underlines the crucial role of machine learning in addressing pressing health challenges. By leveraging a comprehensive dataset and iteratively testing various algorithms, we achieved a final Random Forest model with an exceptional accuracy rate of 99.615% on the test data. This result highlights the robustness of our approach, particularly in identifying key lifestyle and demographic predictors of obesity.

Our findings emphasize the significance of behavioral and dietary factors in determining obesity status. Variables such as caloric intake (CALC), physical activity frequency (FAF), and daily water intake (CH2O) emerged as the most influential predictors, reinforcing the well-established connection between lifestyle choices and weight management. Other important contributors included the number of main meals (NCP) and the frequency of high-calorie food consumption (FAVC), which further underline the multifaceted nature of obesity as a health issue. While demographic variables such as age, height, and gender provided supplementary insights, they were less impactful compared to lifestyle habits, which reinforces the need for tailored interventions that prioritize behavior modification.

In conclusion, this project demonstrates the potential of machine learning to tackle complex health issues like obesity. Our Random Forest model, driven by behavioral and lifestyle predictors, offers actionable insights into obesity prevention and management. By addressing the identified limitations and pursuing the outlined recommendations, we can build upon this foundation to create more comprehensive and impactful solutions in the fight against obesity.

7. Acknowledgement

We would like to express our appreciation to Professor Akram Mousa Almohalwas for providing us with an excellent and informative experience in Stats 101C this fall.

References

Almohalwas, Akram Mousa. “Kaggle Competition Fall 2024”. Dec, 16, 2024.

National Institute of Diabetes and Digestive and Kidney Diseases. (2017). *Overweight & obesity statistics*. U.S. Department of Health and Human Services. Retrieved December 12, 2024, from https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity