Predicting Obesity Status

By: Ethan Lee, Erica Yee, Eric Jung, Ethan Shahzad, Yi Lan

Table of Contents

01

Introduction

Obesity Context and
Data Set Overview

Obesity

Obesity is a medical condition marked by an excessive accumulation of body fat, which can adversely impact health. It is typically defined as having a body mass index (BMI) greater than 30.

Many adults with obesity often experience other **significant chronic conditions**, such as diabetes and heart disease.

Various Factors contributing to obesity:

Environmental and societal:

- Food resources
- Physical activity
- Cultural influences

Genetic:

- Heart disease
- Hypertension
- Stroke

Obesity Status Original Data Set

Observations

Each observation represents an individual

Variables

Information about each individual (ex: age, gender, cholesterol, etc.)

02

Methodology

Data Cleaning and Modeling

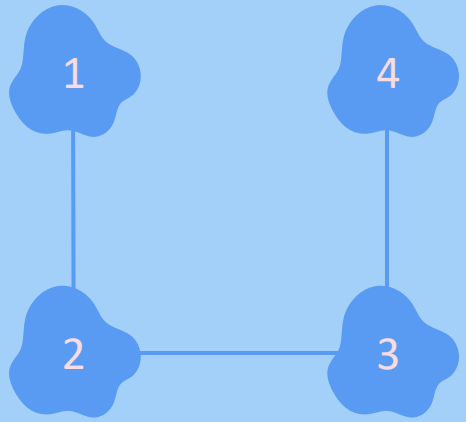
Methods

Clean Data

Fix NA's and outliers

Model Data

Create predictive models using cleaned data



Analyze Models

Assess testing misclassification rates

Compare Models

Choose simplest model with the best MCR

Data Cleaning: Dealing with NAs and outliers

1. For **Numerical Variables,** we replaced the missing values with the median of their respective variables.

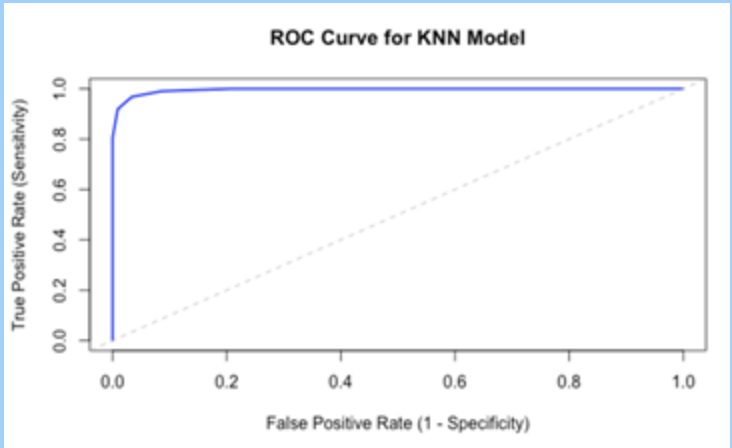
1. For **Categorical Variables,** we replaced the missing values to match the frequency of their respective variables.

1. For **Outliers,** we replaced them with the median values of their respective variables.

1. **Standardizing Predictor Variables**

Data Modeling: K-Nearest Neighbors

All Predictors with k = 5



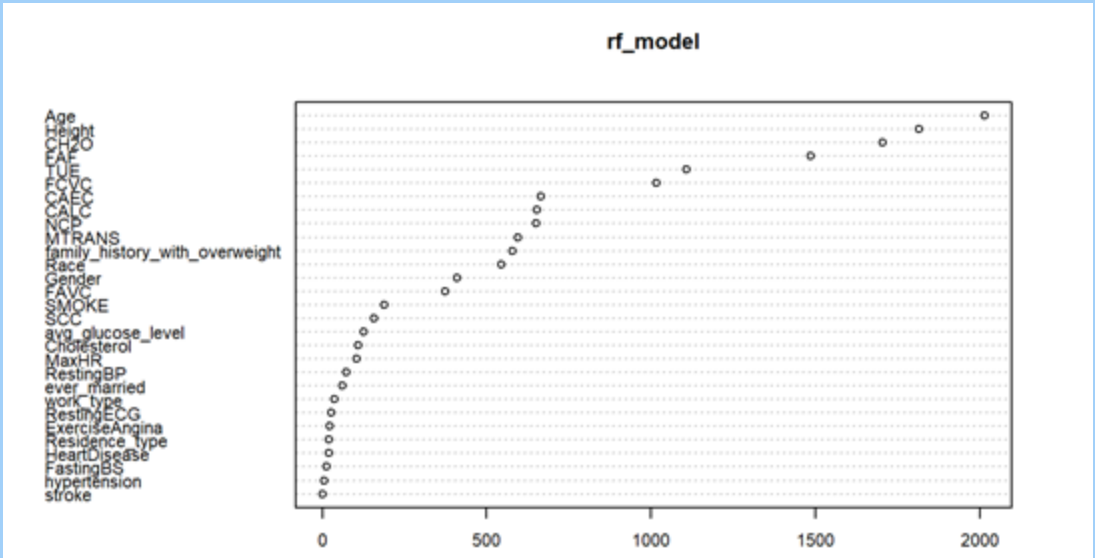
Misclassification Rate: 0.033

Prediction	Not Obese	Obese
Not Obese	18874	400
Obese	657	12083

Data Modeling: Random Forests

Full Model

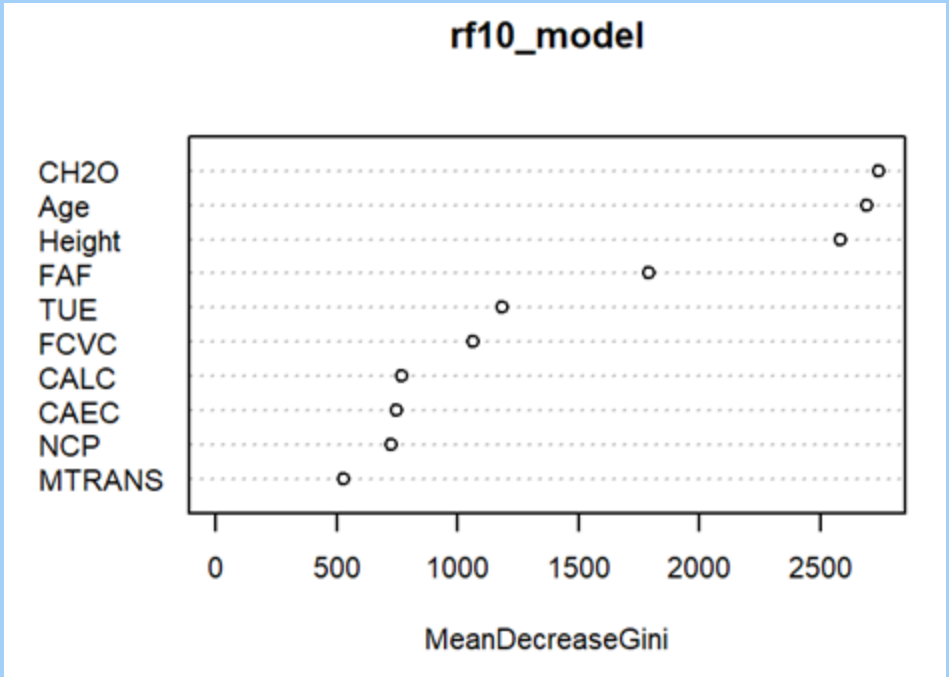
Variable Importance Plot



From the importance plot, we see that there are variables with little to no importance that we can cut in order to simplify our model

Data Modeling: Random Forests (10 Predictors)

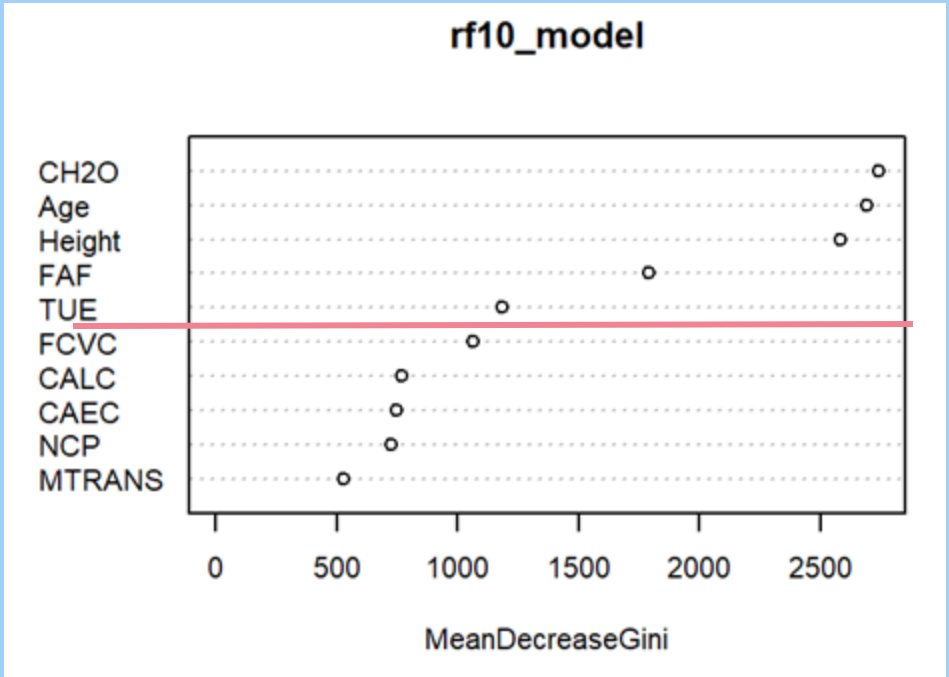
10 Predictor Model



Using these top 10 most important predictors, we get the misclassification rate **0%**

Data Modeling: Random Forests (5 Predictors)

5 Predictor Model



Using these top 5 most important predictors, we get the misclassification rate **0.787%**

Choosing the Best Model: Random Forests (5)

We want the simplest model with the lowest misclassification rate

With a MCR of **0.787%** and an accuracy score of **99.21289%**, our best model was the **Random Forests Model with 5 predictors**

Although our random forests model with 10 predictors yielded a MCR of 0% and an accuracy score of 1, we opted for the 5 predictor model as it still yields a very high accuracy while being much simpler.

03
Results & Discussions

Discussion: Important Predictors

The Most Important Predictors	1	2	3	4	5
	Daily Water Intake	**Height**	**Age**	**Physical Activity Frequency**	**Time Using Technology Devices**

04
Limitations & Conclusions

Limitations

The method of Random Forests chooses predictors based on how it can improve accuracy rather than predictors most related to the response variable

- Therefore, the model is strongly suited for prediction, but may not explain reasonable relations between variables
- Variables we thought would be more important (ex. caloric intake) did not make the top 5 predictors

Conclusion

While our 5-predictor Random Forest model achieved high accuracy, the interpretability of some predictors is unclear. Notable predictors like caloric intake and family history of obesity were omitted, while factors like water intake and time spent on technology were included, raising questions about their roles in obesity.

Further research should focus on determining why predictors such as water intake are influential in predicting obesity as this might give us greater insight into the behavioral and environmental factors behind obesity.

Overall, this analysis suggests that looking beyond traditional risk factors such as caloric intake into lifestyle habits may be beneficial for obesity prevention.

References

- https://www.kaggle.com/competitions/predicting-obesity-status

- Professor Almohalwas' Stats 101C Lectures