Stats 101C Final Project

Predictive Analysis of Obesity Status

Abstract

Based on a Kaggle competition, the objective of this project is to predict someone's obesity status by feeding the individual's data through a statistical learning model. This report aims to provide a detailed analysis from start to finish of the creation of our classification model. An introduction, data analysis, methodology and models, discussion and limits, and overall results of the classification model are incorporated in this paper.

Our final model is a random forest model using five predictors: daily water intake, height, age, physical activity frequency, and time using technology devices. We achieved a Kaggle score of 1.0, with a final ranking of 2nd in the competition.

1. Introduction

Obesity, defined as a medical condition characterized by excessive body fat and a Body Mass Index (BMI) over 30, has emerged as a critical public health concern with profound health, social, and economic consequences. It serves as a major risk factor for chronic diseases such as Type 2 diabetes, cardiovascular disease, and certain cancers, while also contributing to mental health issues and reduced quality of life. Given its increasing prevalence, accurately predicting obesity status is of paramount importance to guide preventive strategies and mitigate its broader societal impact.

This project aims to address the growing challenge of obesity by leveraging a dataset curated from Kaggle and public health resources like the CDC. The dataset includes 29 predictive features, spanning demographic information, dietary habits, physical activity, and health markers, to classify obesity status (ObStatus). It comprises 32,014 training samples and 10,672 testing samples. By predicting obesity status, this project seeks to identify at-risk individuals early, enabling timely interventions and informing public health initiatives. Through data-driven approaches, we will perform feature selection and exploratory data analysis to gain insights into the underlying factors influencing obesity. Our mission was to introduce a classification model to predict the target variable diagnosis (ObStatus) in the testing data by selecting key variables. This work underscores the role of predictive modeling in healthcare, illustrating its potential to address pressing public health challenges and support evidence-based interventions to combat obesity.

2. Data Analysis

Before applying any classification techniques, we cleaned up the data and dealt with the missing values (NA values) in the dataset. Instead of simply erasing all the missing values, we used various techniques to ensure that no important observations were lost. For numerical predictors, we replaced the missing values with the median values, and for the categorical predictors, we replaced the missing values to align with the frequency of each category. Subsequently, when dealing with outliers in the data, we replaced them with the median values. Lastly, we made sure to standardize the predictor variables.

When beginning our data modeling, we used various classification techniques such as LDA, logistic regression, and lasso regression. However, the classification techniques that displayed the lowest misclassification rates were KNN and the random forest model.

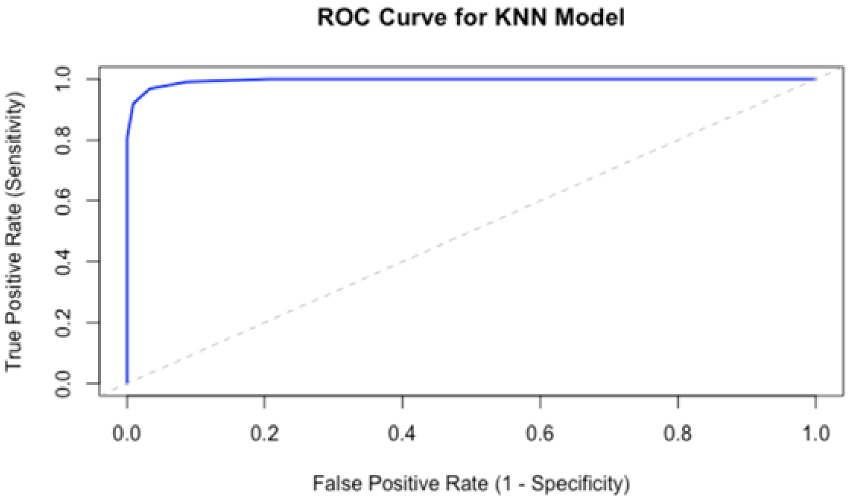
3. Method & Models

We explored various classification techniques, including linear discriminant analysis (LDA), logistic regression, and lasso regression, to evaluate their performance on our dataset. It is worth noting that the accuracy of the other models was approximately 70%, whereas the KNN and random forest models achieved over 90% accuracy after careful tuning of their hyperparameters. This significant improvement underscores the effectiveness of these two models in capturing the underlying patterns within the dataset.

Ultimately, we decided to focus on presenting the results of k-nearest neighbors (KNN) and random forest models.

3.1 KNN

The k-nearest neighbors (KNN) model was implemented using a cross-validation approach to optimize its performance. Specifically, we applied 5-fold cross-validation to ensure robust evaluation of the model's accuracy. The train function from the caret package was used to train the KNN model, with the response variable predicted based on all available features in the training dataset.

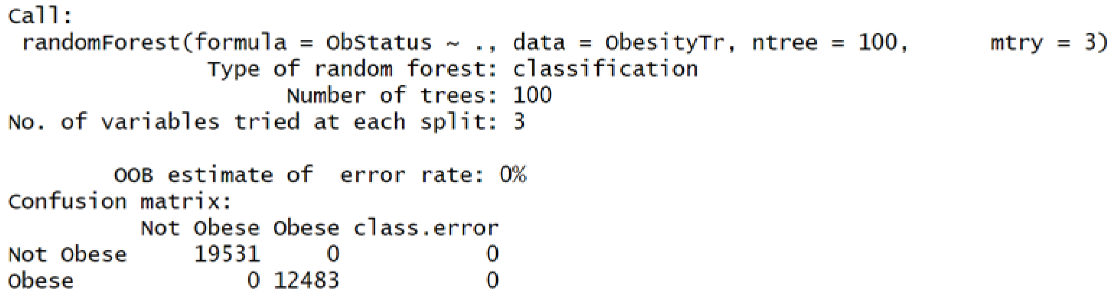


The graph above illustrates the performance of the model, with a curve positioned near the top-left corner indicating superior performance. As observed in this case, such a curve reflects the model's ability to achieve a high true positive rate while maintaining a low false positive rate. In contrast, the diagonal line represents the performance of random guessing, which serves as a baseline for comparison. The substantial distance between the curve and the diagonal line highlights that the KNN model significantly outperforms random guessing, demonstrating its robustness and effectiveness in distinguishing between classes.

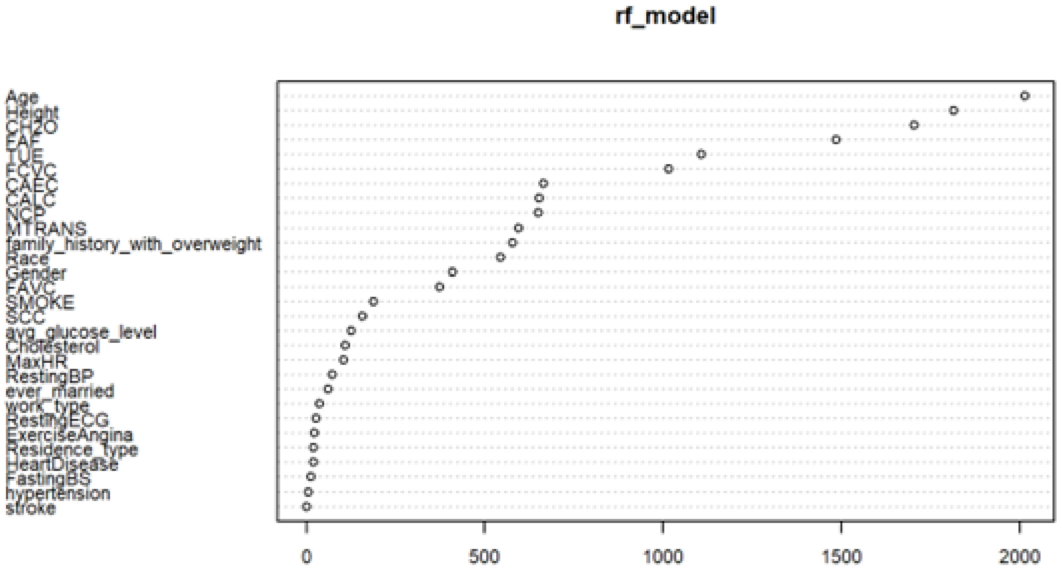
Ultimately, we obtained a model with an accuracy score of 0.97366.

3.2 Random Forests

For the random forest model, we initially employed a full model with all predictors (ntree = 100, mtry = 3) and observed an unexpectedly high accuracy of 100%. Following this, we proceeded to simplify the model.

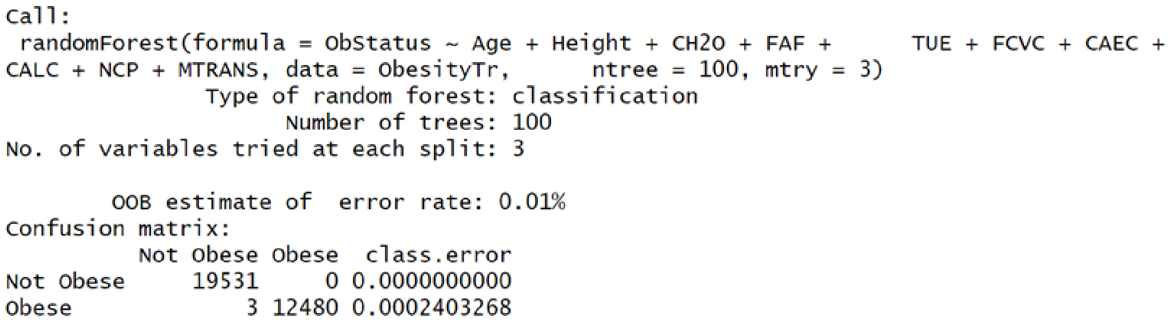


```
Call:
 randomForest(formula = ObStatus ~ ., data = ObesityTr, ntree = 100,      mtry = 3)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0%
Confusion matrix:
          Not Obese Obese class.error
Not Obese     19531     0           0
Obese             0 12483           0
```



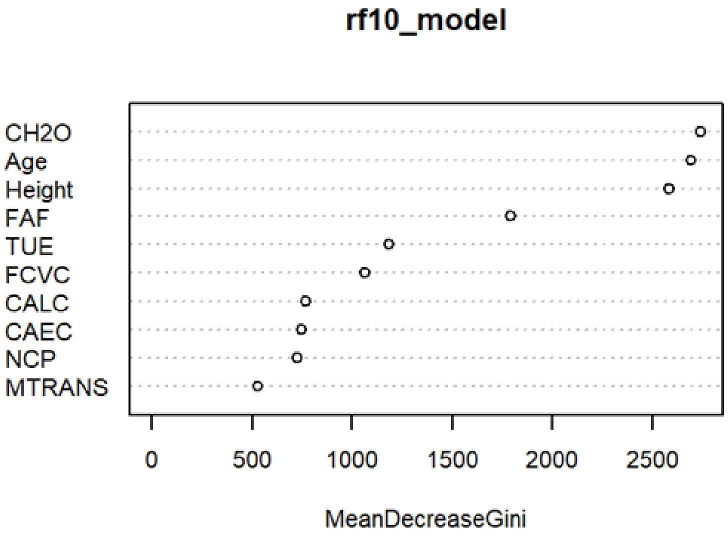
From the importance plot above, it is evident that several variables contribute little to no importance to the predictive performance of the full model. These variables can be effectively excluded to streamline the model, reducing complexity while maintaining or potentially improving accuracy. By focusing on the most relevant predictors, we aim to enhance the model's interpretability and ensure that unnecessary variables do not degrade its performance.

Using these top 10 most important predictors identified from the importance plot, we achieved an accuracy score of 1 (on Kaggle).



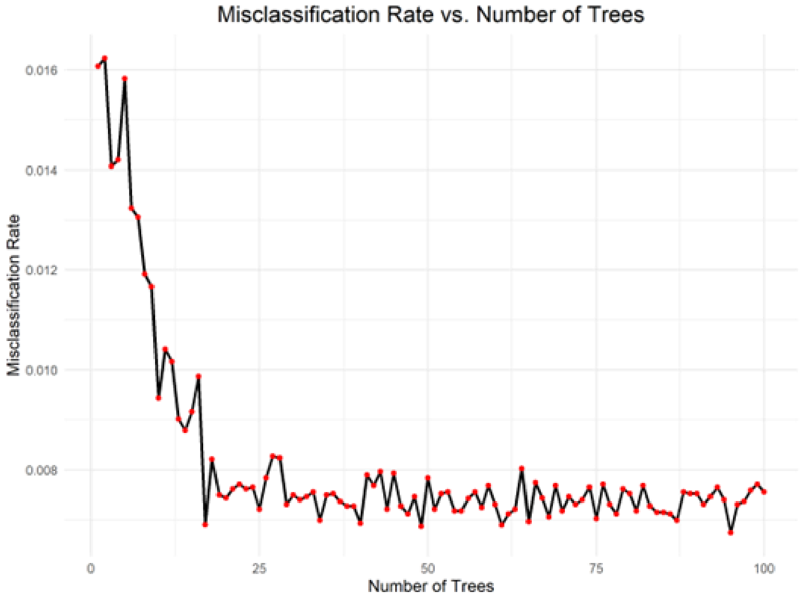
```
Call:
 randomForest(formula = ObStatus ~ Age + Height + CH2O + FAF +      TUE + FCVC + CAEC +
CALC + NCP + MTRANS, data = ObesityTr,      ntree = 100, mtry = 3)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.01%
Confusion matrix:
          Not Obese Obese  class.error
Not Obese     19531     0 0.0000000000
Obese             3 12480 0.0002403268
```

Thus, we further reduced the number of predictors to five.



With a Misclassification Rate (MCR) of 0.787% and an accuracy score of 99.213%, our optimal model was the Random Forest model utilizing five predictors. We aimed to achieve the simplest model with the lowest misclassification rate, and although the Random Forest model employing ten predictors achieved a perfect Misclassification Rate of 0% and an accuracy score of 1, we selected the five-predictor model due to its comparable high accuracy and significantly reduced complexity. This trade-off between simplicity and performance is advantageous, as it enhances interpretability and reduces the risk of overfitting while maintaining excellent predictive capability.

Finally, after further tuning, we set the number of trees (ntree) to approximately 25.



Here is our optimal model:

randomForest(formula = ObStatus ~ Age + Height + CH2O + FAF + TUE, data = ObesityTr, ntree = 25, mtry = 3)

4. Discussion & Limitation

Opting for the simpler 5 predictor random forests model, we obtained an impressive accuracy score of 99.21%. An analysis of variable importance found that the 5 most influential predictors contributing to obesity are daily water intake, height, age, physical activity frequency, and time spent using technology devices. Notably, daily water intake ranked as the single most important contributor. This was surprising amongst other factors that are more commonly associated with causing obesity such as whether an individual exercises or has a family history of obesity. This finding highlights a limitation of random forests, whose iterative training process prioritizes predictors that maximize model accuracy at each step, rather than those that have the strongest theoretical relationship with the response variable. Essentially, the model is strongly suited for prediction but falls short in explaining reasonable relationships between the variables. As such, the interpretability of predictors such as daily water intake are unclear. Additionally, random forests models are limited by their immense computational cost as a result of processing multiple decision trees and aggregating their results. This can make random forests less efficient for very large datasets or applications needing real-time predictions.

Other methods we employed predicted obesity with a lower degree of accuracy and came with their own limitations. For instance, logistic regression assumes linearity in the relationship between the response variable and the predictors. Similarly, KNN can be even more computationally expensive than random forests in the prediction phase and offers little interpretability. Therefore, despite its limitations, random forests emerged as the most effective model for predicting obesity in our analysis.

5. Conclusion

Overall, we determined that the most efficient model for this dataset would be a random forest model with 5 predictors. Throughout the modeling process, we figured out that although select models provided better accuracy than the final model, a model with leveled complexity and accuracy helped provide a more balanced result. During this project, we learned that the limitations of a model should be examined carefully, as it has a significant impact on prediction accuracy. When choosing a statistical learning model, it is important to weigh the benefits and limitations of the model against the type of dataset to get the best results. If further analysis were to be done on this project, we could try other types of models, as we selected a handful due to time restrictions and the nature of the competition. This would allow us to see if there are better-suited models to fit this type of dataset, having a more optimal complexity and higher accuracy (although our model achieved pretty strong results).

Despite the constraints, we are proud that we achieved high placement in the competition resulting from an effective computational model. All in all, this project was a great learning experience for the group and helped us learn to apply solutions to real-world problems, allowing us to greatly further our knowledge in the field.

References

Almohalwas, Akram Mousa. “Predicting Obesity Status.” *Kaggle*, www.kaggle.com/competitions/predicting-obesity-status/overview. Accessed 16 Dec. 2024.