Stats 101C Final Project

Predictive Analysis of Obesity Diagnostic

Kirtan Bhatt, Derek Diaz, Michael Gureghian, Daren Sathasivam

Abstract

The objective of this study was to effectively predict obesity status in individuals given the influence of 29 numeric, categorical, and ordinal predictors. A random forest model was fitted with 15 features derived from a cross-section of VIF screening, AIC stepwise regression, and random forest variable importance for feature selection. This random forest indicated that daily water intake, age, height, and frequency of physical activity are the most valuable predictors in modelling obesity status; height, age, and race are also the most influential given mean decrease in accuracy. The training accuracy of this model was evaluated at 98.49% via 10-fold cross validation, and its testing accuracy was higher at 99.06%, indicating a well-fitted model.

Keywords: Predictive Modeling, Classification, Feature Selection, Variance Inflation Factor, Data Imputation, Random Forest, K-Nearest Neighbors(KNN), Exploratory Data Analysis, Obesity Prediction, Stepwise Regression, Logistic Regression

1. Introduction

Obesity is a pressing health matter reaching unprecedented levels in recent times across the globe. Obesity has been found to be linked with a variety of chronic conditions, diseases, and 13 types of cancer. The World Health Organization also estimates that almost 650 million adults were classified as obese in 2016, highlighting the status of obesity as a dangerous and widespread threat to public health. While the causes are multifaceted, looking at genetic, behavioral, and environmental factors from an avenue of complex analysis provides promise towards understanding and predicting obesity risks. Using data-driven insights, we can address the growing obesity epidemic by magnifying common prevalent factors.

2. The Data

The dataset used in this study provides 29 predictors related to individual lifestyle, health, and demographic attributes to predict obesity status of an individual, our target variable. This includes a diverse array of features including key variables such as age, gender, and race capturing demographic information, as well as health-related metrics like height, resting blood pressure, and cholesterol level. Behavioral factors like caloric intake, frequent consumption of high-calorie food, and physical activity frequency shed light on dietary and activity habits. Environmental influences are captured through variables such as family history of overweightness, residence type, and work type. Additionally, indicators of chronic conditions and risk factors including heart disease, hypertension, and stroke provide further context to individuals' health profiles.

Our target (response) variable, obesity status, categorizes individuals as either obese or not obese, forming the basis for classification in our study. This rich dataset allows for an

in-depth exploration of the relationship between these predictors and obesity, to further develop learning models to better understand and predict obesity status.

3. Analysis

A. Data Cleaning

After getting our hands on the data, our first priority was to ensure the rows and columns were clean in order to move forward with effective analysis. Upon inspection, we observed that approximately 8% of the values in each predictor column were missing or NA values. To account for this, we applied Multiple Imputation by Chained Equations (MICE), a robust method to handle incomplete data. MICE iteratively imputes missing values by individually modeling each relationship, creating multiple plausible datasets to be combined for a final complete data frame in the end. This approach was advantageous in the context of our study, as it minimized bias and variability, while giving us the ability to incorporate the full range of predictors without inaccuracies. By using this imputation method, we ensured the missing values were accounted for, enhancing the accuracy and reliability of future analyses into obesity risk factors.

B. Exploration

To gain initial insights on the relationship between obesity status and the predictors in our data, we first performed linear model analysis using all variables. This helped us get introductory visibility on which were the most influential predictors of obesity, and we chose to take a closer look into family history with overweight and frequent consumption of high-calorie food. Not only did these predictors exhibit significant associations with our target, but overall drew interest in our study as common indicators of the growing epidemic.

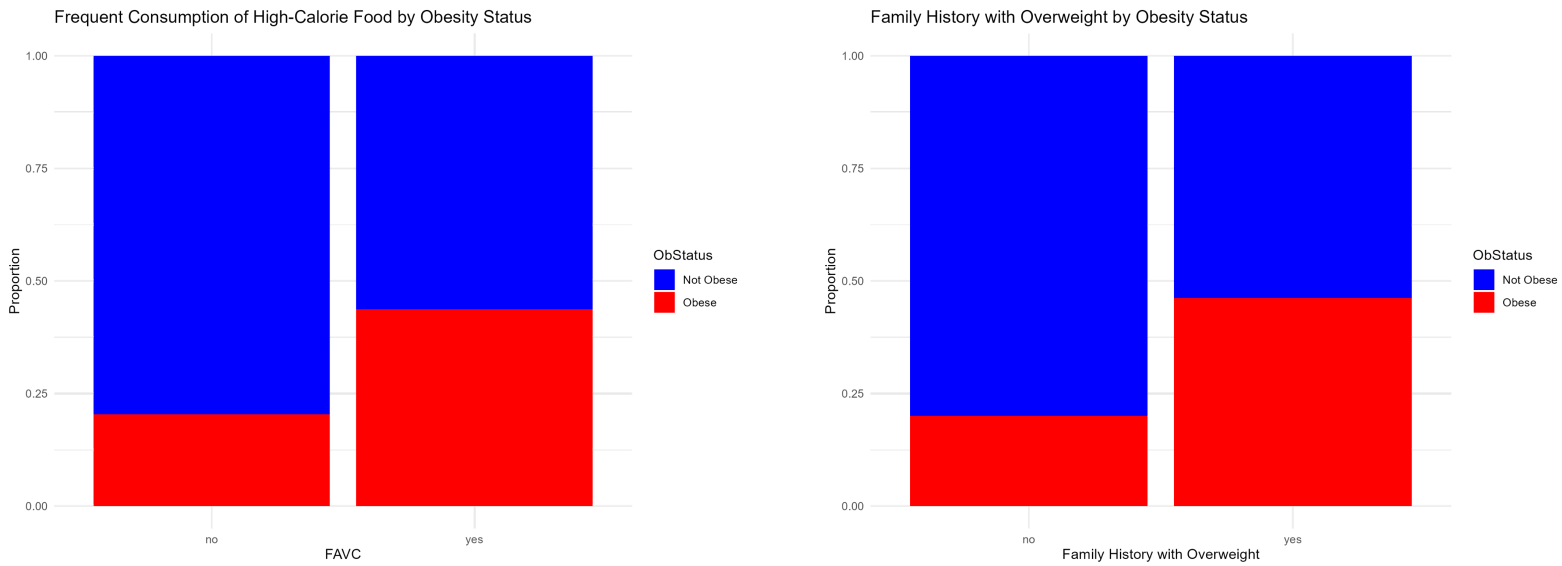

Fig 1. Proportion of Obese in relation to High-Caloric Consumption and Family History

The plot for Family History of Overweight strongly supports the notion that individuals with obesity history in their family are more likely to be obese themselves. Specifically, the proportion of individuals classified as Obese is noticeably higher among those who reported a family history of overweightness compared to those who did not. This trend highlights the combined influence of genetic predispositions and shared behavioral factors within families. The results align with existing research suggesting that familial obesity risk can stem from inherited traits affecting metabolism or appetite, as well as from shared dietary and lifestyle habits. These findings reinforce the significance of considering family history as a key predictor in obesity.

With Frequent Consumption of High-Calorie Food, the plot also indicates a strong association with obesity. Individuals who regularly consume such foods are more likely to be at obesity risk compared to those with minimal to no consumption. Plotting the relationship of this

predictor highlights the critical role of dietary habits in obesity risk, encouraging a stronger focus on nutritional education and intervention aimed at reducing high-calorie food consumption.

Overall, both of these predictors attracted early interest to explore the strongest effects on obesity in our study. In order to draw classification predictions at the highest possible accuracy, we wanted to further weigh our variable selection to construct the strongest sample of key predictors.

C. Variable Selection

After performing exploratory data analysis, we proceeded with variable selection to identify the most relevant predictors to our model. By choosing which predictors/features to include in our model, we are able to eliminate any variables that provide little to no effect on the obesity status of an individual, and consolidate a more representative sample of predictors with the strongest association.

a. Multicollinearity Screening

To address multicollinearity, we calculated the Variance Inflation Factor (VIF) for each predictor. Utilizing a threshold of 5, we flagged predictors with VIFs greater than this threshold as exhibiting multicollinearity.

CALCSometimes	CALCno	CALCFrequently
246.061863	222.837699	64.677704
work_typePrivate	work_typeGovt_job	`work_typeSelf-employed`
36.959361	23.754789	12.686801
work_typeNever_worked	RaceWhite	RaceHispanic
5.744232	4.713543	3.762754

Fig 2. Nine largest VIF values obtained from all predictors.

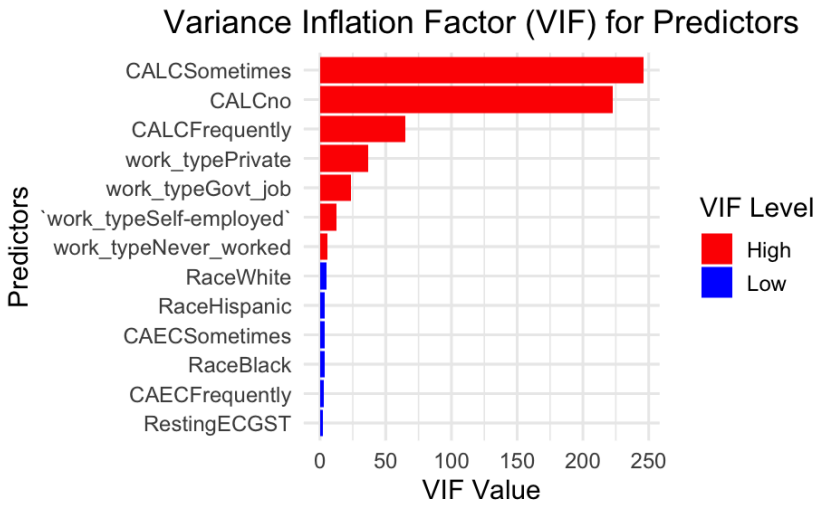

Fig 3. Plot for the largest VIF predictors

As observed in Fig 2 and Fig 3, the largest VIF values derived from all predictors consist mainly of two different predictors: CALC and work_type. Due to their high multicollinearity, we will consider these two predictors as problematic and would be removed from the final model as a result. This initial step in screening for multicollinearity reduces redundancy in the dataset, ensuring that predictors included in the model are independent of each other.

b. Stepwise Selection

Next, we applied stepwise regression using the Akaike Information Criterion (AIC). This iterative process assessed the contribution of each variable to the overall model performance. The stepwise regression process began with a full model and iteratively evaluated each predictor. At each step predictors were removed based on their contribution to the model's AIC value at each step. By doing so, the regression model selects the predictors that minimize the AIC value, which balances model fit with complexity.

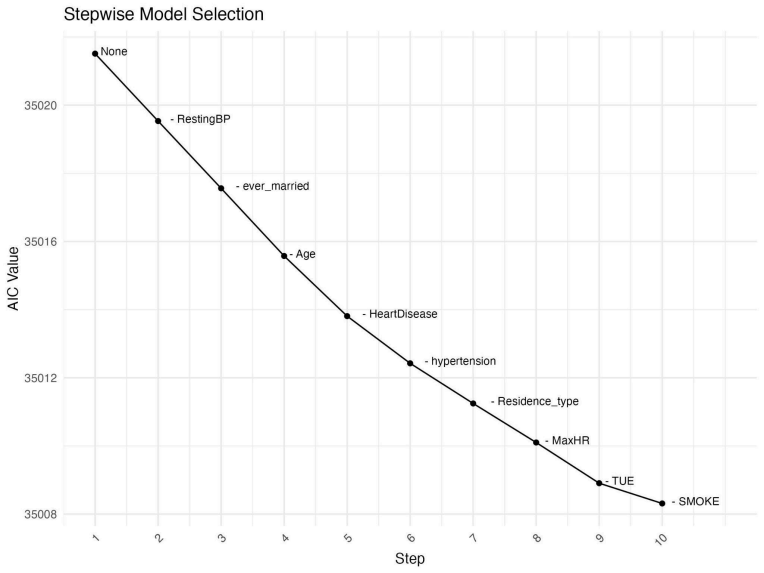

Fig 4. Plot for Stepwise Regression Model

The figure visualizes the AIC Value, and labels each predictor removed at each step. The selected variables explain the variation in the response variable with an optimal balance between simplicity and accuracy.

The stepwise regression model selected the following predictors:

- Numerical Predictors:
    - Height, Age, FAF, NCP, CH2O, Cholesterol, avg_glucose_level
- Categorical Predictors:
    - Race, MTRANS, CAEC, family_history_with_overweight, FAVC, Gender, SCC, ever_married

The predictors that have been selected will be considered for the final model.

c. Random Forest Variable Importance

Lastly, we evaluated predictor importance using the Random Forest function. This method can be used to find the variable importance value which is a metric indicating a features' contribution to prediction accuracy.

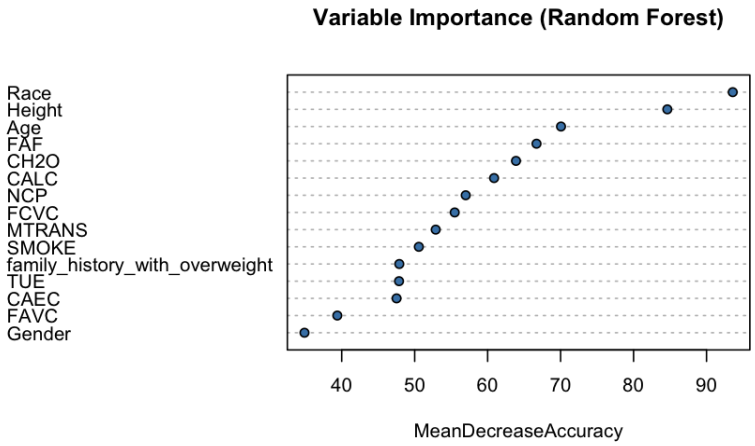

Fig 5. Variable importance plot.

Utilizing the randomForest library in R, we created a variable importance plot (Fig 5). As observed from the plot, the top predictors by mean decrease in accuracy were visualized, highlighting variables such as Race, Height, Age, FAF, and CH2O being the most impactful to the model.

d. Final Selected Predictors

Taking into account the results from the variable inflation factor screening, stepwise selection using AIC, and random forest variable importance, we narrowed the variables to be used for our final model. Although Cholesterol and avg_glucose_level weren't indicated by the variable selection methods, we believed these factors to have an impact on a person's chance of having obesity. Therefore, we finalized our model with the following 15 predictors:

- Selected numerical predictors:
    - Height, Age, FAF, NCP, CH2O, Cholesterol, avg_glucose_level
- Selected categorical predictors:
    - Race, MTRANS, CAEC, family_history_with_overweight, FAVC, Gender, SCC, ever_married

These selected predictors strike a balance reducing dimensionality from the original 29 predictors to the final 15 predictors and retains the most informative variables for classification.

4. Methods & Models

The data we have been given consists of 32,014 observations for the training dataset and 10,672 observations for the training dataset, which is roughly a 75-25% split. By using the training data for our model, we would produce the predicted obesity status utilizing the testing data. We repeated these steps for each model attempted and obtained the accuracy rate for each model to then select our best model based on the accuracy obtained.

A. Logistic Regression Model

Logistic Regression is a parametric and supervised learning algorithm used for binary problems in classification. The model predicts the probability that an observation belongs to a specific class by fitting the data to a logistic function. This was our first choice model to gain some initial insights on testing prediction accuracy. Ultimately, the accuracy when put into the Kaggle competition was around 66% for all logistic regression models, predicting obesity at an ineffective rate compared to future models we chose.

B. KNN Model

The K-Nearest Neighbors (KNN) algorithm is a non-parametric, instance-based learning method where predictions are based on the majority class of the k-nearest neighbors in the training dataset. Using the caret library in R, we scaled the numeric variables using the preProcess function to ensure all variables were on a similar scale. This prevents variables from dominating the distance calculations if they have wide ranges. The model was trained using a 10-fold cross-validation strategy to minimize overfitting. Then using a tune grid to test various k-values, the best performing k-value being 5 which would then be used for the KNN model.

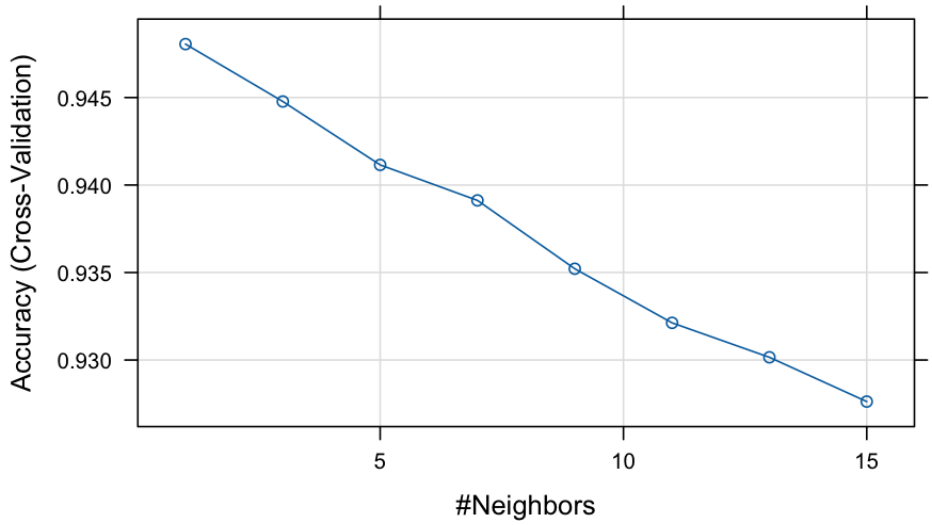

Fig 6. KNN model's accuracy with different number of neighbors using k = 5

In Fig 6, it can be observed that as the number of neighbors increases, the accuracy of the model's predictions gradually decreases due to the predictions becoming more generalized. The accuracy based on the testing dataset through the KNN using k = 5 was 94.805% and the kaggle score was 94.986%.

C. Random Forest Model

Random Forest is an ensemble learning method based on multiple decision trees. It reduces overfitting by averaging predictions from multiple trees and provides robust variable importance metrics. Data scaling was not required for this model since it is unaffected by various variable ranges. Using our selected 15 predictors, we trained the random forest model using 10-fold cross validation.

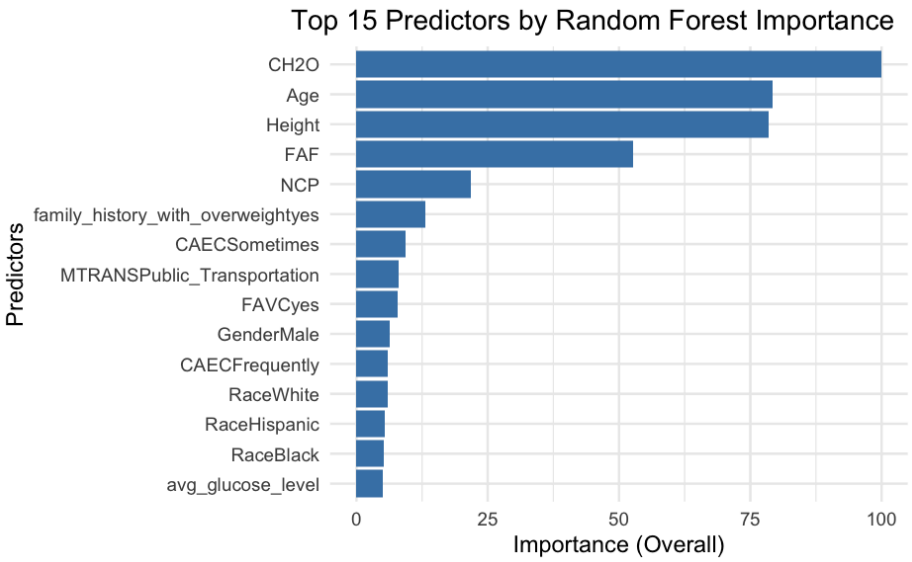

Fig 7. Top 15 predictors obtained from the Random Forest model

Based on the testing dataset, we were able to obtain an accuracy of 98.938% and our kaggle accuracy was 99.062%. Thus, this being the highest accuracy from the model attempted, we selected this as our final model.

5. Discussion and Limitations

While our random forest model achieved a 99.06% accuracy and KNN reached 94.3% accuracy, there are several aspects to consider when interpreting our results. The high accuracy from both models highlights the efficiency of our selected predictors, derived from a combination of multicollinearity screening using VIF values, stepwise regression with AIC, and random forest variable importance rankings. These methods allowed us to identify and focus on a concise set of 15 predictors out of the original 29, balancing model performance and interpretability.

Additionally, our models also rely on imputation of all missing values through MICE, however exploring additional methods may have assisted us in improving the accuracy of our results.

Another limitation stems from the complexity of the initial dataset. The dataset contained categorical variables with multiple levels such as work_type and MTRANS. Encoding these variables can introduce additional predictors in some models that cannot take categorical variables such as the KNN model. This complexity emphasizes the trade-off between using all available information versus simplifying the dataset for interpretability.

While KNN and random forest were our most effective models, other methods can and should be evaluated to further investigate the patterns and predictability of the data.. Future studies could explore ensemble methods combining multiple models for better performance and stability.

6. Conclusion and Recommendation

This study has explored the effectiveness of combined feature selection techniques, different methods of imputation, and advanced models to achieve high predictive accuracy classification. Furthermore, engaging in the modeling process with a structured approach (exploratory data analysis, data cleaning, variable selection, model construction, model assessment) was effective towards building a holistic understanding of the data.

This analysis showed the value of systematically reducing predictors using variable selection to balance interpretability and performance. By utilizing KNN and random forest models, we were able to obtain relatively high accuracies, which demonstrates the effectiveness of the variable selection approach for this dataset. Furthermore, performing variable selection allowed us to identify and focus on the most significant driving factors of obesity, providing valuable insights into which predictors have the greatest impact on the outcome. This understanding is crucial for informing targeted interventions and policy-making.

Overall, this project allowed us to strengthen our abilities in real-world data analysis, emphasizing the importance of matching specific models to the various possible dataset characteristics and balancing accuracy with interpretability. This experience will assist us in our future endeavors.

7. Acknowledgement

Thank you to Professor Akram Mousa Almohalwas for his support and for his work teaching Statistics 101C in the 2024 Fall quarter; without his guidance this study would never have been possible.

References

National Heart, Lung, and Blood Institute. “Overweight and Obesity Causes.” *National Institutes of Health*, 24 August 2022, www.nhlbi.nih.gov/health/overweight-and-obesity/causes.

Song, Jiyoung, et al. “The Global Prevalence of Obesity: Patterns and Trends.” *Frontiers in Public Health*, vol. 10, 2022, article 998782, www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.998782/full.

Van Buuren, S. *mice: Multivariate Imputation by Chained Equations* (version 3.16.0), Comprehensive R Archive Network, 2024, https://cran.r-project.org/web/packages/mice/mice.pdf.

World Obesity Federation. “Obesity Classification.” *World Obesity*, www.worldobesity.org/about/about-obesity/obesity-classification.