Predicting Obesity Status

Hanyang Zhou, Jamie Tian, Sinan Allahbachayo, Tyler Taylor, Yiting Wang

December 16, 2024

Contents

1 Abstract

This study investigates the prediction of obesity status using statistical methods as well as some machine learning techniques in the Kaggle dataset ”Predicting Obesity Status”. Using logistic regression and random forest models on the dataset after data processing, the study addresses the challenges of classifying obesity status using mixed numerical and categorical predictors.

Multicollinearity was assessed using the Variance Inflation Factor (VIF), which revealed no significant correlations among predictors. The random forest model had a great predictive performance, effectively capturing non-linear relationships in the dataset.

The study contributes information on data-driven health risk assessment, highlighting the potential of advanced computational techniques for the classification of obesity status. Future improvements include expanding demographic features and exploring additional machine learning approaches to improve predictive accuracy.

2 Introduction

Obesity, a medical condition of excess body fat, defined by BMI over 30, is a major public health concern due to its widespread impact and associated healthcare costs. Obesity severely compromises personal health in both short-term and long-term. It significantly increases the risk of developing several serious health conditions such as heart disease, type 2 diabetes, stroke, high blood pressure, and osteoarthritis, leading to decreased quality of life and potential premature death. Obesity also requires extensive utilization of public healthcare resources. According to the U.S. Center for Diseases Control and Prevention [**1**], obesity costs the US healthcare system almost $173 billion a year.

On the other hand, accurately predicting obesity allows for early identification of individuals at risk, enabling timely interventions and preventative measures to be taken, potentially reducing the development of obesity-related health complications later in life. Therefore, finding an effective predicting method for obesity status is an important part of the solution to the obesity problem.

This study analyzes datasets from the Kaggle contest ”Predicting Obesity Status” [**2**], aiming at classifying individuals’ obesity status as “Obese” or “Not Obese.” There are 29 predictors in total, including 18 categorical and 11 numerical predictors. The training and testing datasets contain 32,014 observations and 10,672 observations respectively. Statistics and machine learning techniques are used to perform comprehensive investigation and prediction on the dataset.

3 Methods

3.1 Data Preprocessing

Data preprocessing is important to ensure the quality and consistency of the data, making it suitable for analysis, leading to more accurate and reliable predictions. In our study, we checked for data consistency, handled missing data and examined for multicollinearity problems (section 3.3). We first checked that the variables in our test and training datasets match by finding the intersection of their column names, and ensured that the datasets have the same structure.

Next, we checked the missing values in the dataset. As we can see from the missing value tables (Figure 1), missing value is an important problem in both the training and testing datasets, with high mean missing values of 7.73% and 8.00% respectively. Handling those missing values helps maintain the integrity of the dataset. Imputing missing values ensures that the data remains consistent and suitable for analysis. For numerical variables, we replaced the missing values ‘NA’ in each variable with the mean of that variable. For categorical variables, we replaced the missing values ‘NA’ in each variable with the most frequent category in that variable (mode).

Missing Percentage Training

Feature	Percentage
Age	7.87
Gender	8.03
Height	7.93
family_history_with_overweight	7.96
FAVC	8.06
FCVC	8.04
NCP	8.07
CAEC	7.92
SMOKE	8.15
CH2O	7.96
SCC	7.97
FAF	8.00
TUE	8.10
CALC	8.17
MTRANS	7.78
Race	8.08
RestingBP	7.84
Cholesterol	7.98
FastingBS	7.98
RestingECG	8.17
MaxHR	8.08
ExerciseAngina	7.89
HeartDisease	8.11
hypertension	8.17
ever_married	7.93
work_type	8.33
Residence_type	7.97
avg_glucose_level	7.81
stroke	7.66
ObStatus	0.00

Missing Percentage Testing

Feature	Percentage
Age	8.06
Gender	8.29
Height	8.05
family_history_with_overweight	7.96
FAVC	8.00
FCVC	7.96
NCP	7.76
CAEC	8.17
SMOKE	8.19
CH2O	8.01
SCC	7.81
FAF	7.50
TUE	7.87
CALC	7.93
MTRANS	8.32
Race	7.66
RestingBP	8.02
Cholesterol	7.74
FastingBS	8.09
RestingECG	8.37
MaxHR	8.04
ExerciseAngina	7.92
HeartDisease	7.93
hypertension	7.75
ever_married	7.90
work_type	8.60
Residence_type	7.90
avg_glucose_level	8.07
stroke	8.13

Figure 1: Missing Value in Percentage for Training data (Left) and Testing Data (Right)

3.2 Choice of Models

We choose Logistic Regression as our exploratory approach to the relationship between predictors and obesity status, and then further improved the modeling performance in accurately predicting obesity status using the Random Forest method.

The advantage of a logistic regression classifier is that it is easy to implement and computationally efficient. It also provides insights into important contributing factors in obesity predictions. However, there are also drawbacks of this method. Logistic regression is sensitive to multicollinearity, meaning that correlated predictors can distort coefficient estimates. In section 3.3, we examined our data for multicollinearity problems. Logistic regression is also not flexible and cannot accurately capture the non-linear relationships, therefore, we didn't expect high performance of prediction accuracy.

In order to reach higher prediction accuracy, we also applied the Random Forest method. Random Forest is able to handle a mix of numerical and categorical predictors well, and is also very accurate with large datasets. These characteristics make it a suitable choice for our obesity status dataset.

Therefore, to best leverage the simplicity of logistic regression and the accuracy of the Random Forest method, we first apply Logistic Regression to the data, aiming to explore the relationship between variables. Then we apply the Random Forest method for a more accurate prediction of individual obesity status.

For a binary classification problem, some other conventional methods are LDA, QDA and KNN. Theoretically, these models can also be used in modeling mixed data with both numerical and categorical variables. However, it should be noted that we have more categorical predictors (18) than numerical ones (11), and many of our categorical predictors in the obesity data set are binary. This data structure makes it difficult to satisfy the normality assumption for LDA and QDA, find linear discriminants with LDA, or separate classes in the data space using distance metrics with KNN. Therefore, these methods are not prioritized in our study.

3.3 Multicollinearity

Multicollinearity refers to that several independent variables in a model are correlated. Multicollinearity among independent variables will result in less reliable statistical inferences. Variance Inflation Factor (VIF) is a statistical metric used to measure the degree of multicollinearity between independent variables in a regression model, where $\text{VIF} = 1$ stands for an ideal case of no correlation. Conventionally, $1 < \text{VIF} \leq 5$ is generally acceptable and $\text{VIF} > 5$ stands for high correlation between predictors, indicating multicollinearity problems. In this study, we used our logistic regression model with all 29 variables to check for multicollinearity. We found that all predictors have a VIF value between 1 and 2, indicating that there are no multicollinearity problems in the training data.

4 Model Testing

4.1 Full Logistic Regression Model

First, we developed a Full Logistic Regression model using the obesity status variable as the response variable and ALL 29 predictors in the training dataset. This model was used to predict the probability that an individual is obese or not obese and categorize them based on the more likely option.

	Not Obese	Obese
Not Obese	16526	5114
Obese	3005	7369

Figure 2: Confusion Matrix for the Full Logistic Regression Model

We calculated the error rates for both classes of the obesity status variable: ”Not Obese” and

”Obese”. The error rates for each class are as follows:

- The error rate for class “Not Obese” is approximately 0.2363.
- The error rate for class “Obese” is approximately 0.2897.
- The error rate for both classes is approximately 0.2536.

The training error rate of the model was found to be approximately 25.4%. The training accuracy is approximately 74.6%. Our evaluation on the model showed that the model had had an okay performance with a somewhat high accuracy given the simplicity of the model. However, this model suffered from high dimensionality and therefore was more difficult interpret than the next model we used.

By sorting the predictors by p-value, we found which predictors were considered significant in the Full Logistic Regression Model. The predictors with the most significance, including height, vegetable consumption frequency, number of main meals, daily water intake, physical activity frequency, cholesterol, average glucose level, gender, family history of overweight, high caloric food frequency, consumption of food between meals, consumption of sweet drinks, transportation methods, race, FastingBS, exercise habits, and stroke history, are suggested to have the greatest impact on the accuracy of the model in predicting obesity status.

4.2 Partial Logistic Regression Model

Next, we developed a partial logistic regression model using the obesity status variable as the response variable and ONLY 5 predictors in the training dataset. We started from a full lasso regression model and removed the least significant predictor(s) step by step and recreated a new partial lasso regression model with the remaining predictors each time. We found that the most balanced Lasso logistic regression model used ONLY 5 predictors. This model was used to predict the probability that an individual is obese or not obese and categorize them based on the more likely option.

	Not Obese	Obese
Not Obese	16984	5813
Obese	2547	6670

Figure 3: Confusion Matrix for the Partial Logistic Regression Model

We calculated the error rates for both classes of the obesity status variable: ”Not Obese” and ”Obese”. The error rates for each class are as follows:

- The error rate for class “Not Obese” is approximately 0.2550.
- The error rate for class “Obese” is approximately 0.2763.
- The error rate for both classes is approximately 0.2611.

The training error rate of the model was found to be approximately 26.1%. The training accuracy is approximately 73.9%. This is only about 0.7% worse than the full logistic regression model. Our evaluation on the model showed that the model had had a similar performance to Model 1.1 even though this model has significantly lower dimensionality and thus much more interpretable. This model is the best balance of sacrificing accuracy for very simple interpretability.

By using step-wise lasso regression as a feature selection technique, we found which predictors were considered the most significant in the partial logistic regression model. The predictors with the most significance, including daily water intake, consumption of food between meals, physical activity frequency, consumption of high caloric food frequency, and choice of transportation method, are suggested to have the greatest impact on the accuracy of the model in predicting obesity status.

4.3 Random Forest Model with Full Predictors

We developed a Random Forest model using the obesity status variable as the response variable and all 29 predictors in the training dataset. The model was trained with 1000 trees. A subset of predictors was randomly selected at each split. The size of this subset was equal to the square root of the total number of predictors in the dataset. The random sampling of predictors helps to reduce overfitting.

	Not Obese	Obese
Not Obese	19529	2
Obese	175	12308

Figure 4: Confusion Matrix for the Full Random Forest Model

We calculated the error rates for both classes of the obesity status variable: ”Not Obese” and ”Obese”. The error rates for each class are as follows:

- The error rate for class “Not Obese” is approximately 0.0001.
- The error rate for class “Obese” is approximately 0.0140.
- The error rate for both classes is approximately 0.0055.

The training error rate of the model was found to be approximately zero. The training accuracy is approximately 100%. Our evaluation on the model showed that the model had a good performance with high accuracy.

By sorting the predictors by mean decrease in accuracy, we found the importance of each predictor in the Random Forest model. The predictors that have the greatest mean decrease in accuracy, including race, height, age, physical activity frequency, and caloric intake, are suggested to have the greatest impact on the accuracy of the model in predicting obesity status. Resting electrocardiogram (ECG), fasting blood sugar, hypertension, residence type, and stroke are predictors that are suggested to have the least importance on the model's accuracy.

5 Results

The analysis highlights the effectiveness of three models—full logistic regression, partial logistic regression, and random forest—in predicting obesity status. Each model was evaluated on accuracy, interpretability, and computational efficiency, with significant differences observed in performance.

5.1 Full Logistic Regression Model

The full logistic regression model utilized all 29 predictors and achieved a training accuracy of 74.6% with an error rate of approximately 25.4%. This model provided valuable insights into key predictors such as height, vegetable consumption, physical activity, and family history of overweight. However, the model's high dimensionality complicated interpretation, making it less suitable for practical applications despite its baseline accuracy.

5.2 Partial Logistic Regression Model

Using lasso regression for feature selection, the partial logistic regression model retained only five predictors: daily water intake, high-calorie food consumption frequency, transportation methods, physical activity frequency, and consumption of food between meals. Despite reduced complexity, this model performed comparably to the full model with a slightly lower accuracy of 73.9%. Its simplicity and enhanced interpretability make it a balanced choice for understanding the critical factors influencing obesity.

5.3 Random Forest Model

The random forest model demonstrated superior predictive power, achieving an almost perfect training accuracy of 100% and a negligible error rate of 0.0055%. Its capacity to handle non-linear relationships and mixed data types allowed it to outperform logistic regression models significantly. Key predictors influencing this model's accuracy included race, height, caloric intake, and physical activity frequency. However, the model's complexity limited interpretability, and slight instability was observed in predictions due to random sampling of features.

5.4 Comparative Performance

The random forest model emerged as the most accurate, suitable for high-stakes applications requiring precision. However, logistic regression models, especially the partial model, provided valuable interpretability and simplicity for understanding the dataset's structure. These complementary models highlight the trade-offs between accuracy and usability.

6 Discussion

6.1 Limitations

Although the overall accuracy is high, we acknowledge that our accuracy could still be improved, and there are issues in our model and our process for obtaining the desired model.

In terms of the data cleaning process, our approach of handling missing data by replacing missing values with the mean or the most frequent category might introduce bias, especially when missing values are not similar to the mean or mode.

In terms of modeling, logistic regression suffers from the issue of linearity as it assumes a linear relationship between target variables and the predictor. This would limit its effectiveness as if the relationships are more complex, alternative models or classifiers may be more appropriate. For random forest, it is computationally complex for large dataset. Also, it is hard to interpret the model and its individual predictions. More importantly, it suffers from a lack of stability as variations with variables or the random seed used could lead to different predictions. In addition, a limited number of models were tried due to time constrains. Other classification methods, like Support Vector Machines (SVM) or Gradient Boosting Machines (GBM), were not used, and they could potentially enhance the performance of our models.

From the confusion matrix of our prediction of training data, the error rate for “Not Obese” is lower for that of “Obese”, meaning that our model is more accurate in classifying obese than not obese. In real world context, detecting obese is more important as is allows for early intervention or medical support which could lead to better management of potential health risks.

6.2 Conclusions and Future Improvements

From this project, we had the opportunity to understand and apply machine learning techniques learned in class to a real-world context, focusing on variables that could cause obesity. During the presentation and the report, we went through the process of data cleaning, feature selection, model testing and model construction. We particularly realized the importance of the choice of model and its integral role in predicting the accuracy of the results. We also recognized that there is always a trade-off when making choices between models. Logistic regression, despite having the advantage of simplicity and computational effectiveness, may not be able to capture the non-linear patterns present in our dataset. On the other hand, random forest model, having a higher accuracy for capturing the nuanced interactions in the data, may suffer from computational expensiveness and instability.

For future improvements, we plan to expand our dataset by including a broader set of features, especially demographic variables like race and age to consolidate our understanding of obesity across populations. We also aim to incorporate more longitudinal data which could enable us to spot the trend of obesity over a longer period of time.

Moreover, the adoption of more machine learning techniques, especially Support Vector Machines (SVM) and Gradient Boosting Machines (GBM), will also be helpful. This is because these modeling techniques could be more effective at handling the high dimensionality and heterogeneity of our data, which could potentially reduce bias and offer higher accuracy for spotting non-linear relationships. By utilizing more techniques, we can better compare the prediction accuracy of each method and leverage the collective power of multiple models to improve our prediction.

In essence, this project is a crucial experience for us to learn more about machine learning and practice our skills with real-world dataset, which set the stage for our future learning and research.

7 Reference

[1] Centers for Disease Control and Prevention. (2024, May 16). About Obesity. Obesity. https://www.cdc.gov/obesity/php/about/index.html.

[2] Akram Almohalwas. (2024, October 30). Predicting Obesity Status. Kaggle. https://www.kaggle.com/competitions/predicting-obesity-status/overview.