Predicting Obesity Risk

A Machine Learning Approach Utilizing Diverse Health Factors

Eric Wei
Business Analytics
Wentworth Institute of Technology
Boston, Massachusetts, USA
weil1@wit.edu

ABSTRACT

Obesity is a significant global health issue and a major risk factor for numerous chronic diseases, including diabetes and cardiovascular conditions. Accurate prediction and early identification of individuals at risk are crucial for developing effective intervention and prevention strategies. This project leverages machine learning (ML) techniques to classify obesity risk levels using a comprehensive dataset of demographic, health, and lifestyle variables sourced from the U.S. National Health and Nutrition Examination Survey (NHANES).

KEYWORDS

- Machine Learning
- Obesity Prediction
- Logistic Regression
- K-Nearest Neighbors
- Random Forest

1 Introduction

Obesity is a pervasive and complex health issue with far-reaching consequences. According to the World Health Organization, worldwide obesity has nearly tripled since 1975, and it is a major risk factor for conditions such as heart disease, stroke, type 2 diabetes, and certain types of cancer [1].

This moves beyond simple correlations to build predictive models that can synthesize a wide array of data points. By leveraging machine learning, we can uncover complex, non-linear relationships between lifestyle factors and obesity risk that might be missed by conventional statistical methods. A reliable predictive model can be a powerful tool for public health officials, clinicians, and even individuals, enabling targeted prevention programs, personalized health recommendations, and more efficient allocation of healthcare resources.

The application of machine learning in healthcare and epidemiology has grown significantly. Current research in obesity prediction has successfully utilized various algorithms. For instance, studies have used decision trees and Random Forests to identify key predictors from demographic and behavioral data, often finding factors like physical activity level, sedentary time, dietary patterns, and socioeconomic status to be highly influential [2]. Other research has employed logistic regression for its interpretability, providing clear odds ratios for different risk factors.

A review of the literature indicates that ensemble methods like Random Forest often achieve high accuracy in this domain due to their ability to handle mixed data types and capture complex interactions [3].

2 Data

For this analysis, the NHANES Obesity Dataset was sourced from Kaggle [4]. This dataset derives from the National Health and Nutrition Examination Survey (NHANES), a program of studies designed by the National Center for Health Statistics

2.1 Source of dataset

This dataset is a subset of the U.S. National Health and Nutrition Examination Survey (NHANES), focusing on adults aged 18 to 80. The original NHANES data is collected through a complex, multistage probability sampling design to represent the non-institutionalized US population. The Kaggle version has been structured for machine learning tasks, with obesity status as the target variable.

2.2 Characters of the datasets

The dataset contains about 10,000 instances (rows) and 17 features (columns). The target variable, BMI_WHO, is a multi-

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA
© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00

class categorical variable with four levels: Underweight, Normal weight, Overweight, and Obese.

Name	Data Type	Units / Levels
BMI_WHO	Categorical (Ordinal)	Underweight, Normal, Overweight, Obese
Age	Continuous (Integer)	Years
Gender	Categorical (Binary)	Male, Female
Race1	Categorical (Nominal)	White, Black, Hispanic, etc.
Education	Ordinal	Levels
HHIncome	Ordinal	Levels
PhysActive	Categorical (Binary)	Yes, No
Smoke100	Categorical (Binary)	Yes, No
Diabetes	Categorical (Binary)	Yes, No
BPSysAve	Continuous (Float)	mmHg
TotChol	Continuous (Integer)	mg/dL
Alcohol12PlusYr	Categorical (Binary)	Yes, No
MaritalStatus	Categorical (Nominal)	Married, Widowed, Never Married, etc.
Work	Categorical (Ordinal)	Not Working, Looking, Working
Height	Continuous (Float)	Centimeters
Depressed	Categorical (Ordinal)	None, Several, Major, etc.

3 Methodology

For this analysis, three machine learning models were employed to classify individuals' obesity risk based on NHANES data. These models include Logistic Regression, K-Nearest Neighbors and Random Forest. The selection allows for a comparison between a straightforward baseline model and a powerful, non-linear classifier to determine the most effective approach for this prediction task. All the coding are deployed in python environment with necessary packages like: numpy, pandas, matplotlib, seaborn and sklearn.

3.1 Logistic Regression

Logistic Regression is a statistical and machine learning model used for binary and multi-class classification problems. Despite its name, it is a classification algorithm. It models the probability that a given input belongs to a particular category. The core of the model is the logistic function (sigmoid function), which maps any real-valued number into a value between 0 and 1, representing a probability.

Advantages of Logistic Regression model is that it’s highly interpretable. The coefficients of the model indicate the direction and magnitude of each feature's influence on the log odds of the outcome. It provides a calibrated probability score for the classification, not just a class label.

And the disadvantage is that the assumption of linearity. If the relationship between features and the log-odds is not linear, the model will have poor performance. So, I add other models as a comparison.

3.2 K-Nearest Neighbors

KNN classifies a new data point, it looks at the existing data points that are most similar (its "neighbors") to that new point. The new point is then assigned the most common class among those neighbors.

KNN does not build a formal model during the training phase. The "training" data is just stored in memory. The actual work of calculating distances and making a prediction happens only when you ask it to classify a new data point.

K is the most important hyperparameter, representing the number of neighbors to consider (e.g., K=5 means it looks at the 5 closest points). In this case, we will discover the best k value with cross validation and apply to the model training.

3.3 Random Forest

Random Forest is an ensemble learning method, specifically a Bagging (Bootstrap Aggregating) algorithm, used for both classification and regression. It operates by constructing a multitude of decision trees during training. For a classification task, the output of the Random Forest is the class selected by most of the trees.

The advantage is that it does not assume a linear relationship between features and the target so if the relationship is non-linear, it will outperform the Logistic Regression. And it can handle numerical and categorical features without the need for extensive pre-processing (like SVM).

The disadvantage is while it provides feature importance, the overall model is a "black box," and it is difficult to understand how a specific prediction was made.

3.4 Preprocessing

3.4.1 Clean data. After examining the CSV file, we found that the variable “Depressed” has three values: “None, Several and Most”. In pandas, it regards it as null value by default, so we must convert the value “None” to “No”. After that we cleaned the dataset with removing the rows which contain any NaN values.

3.4.2 Split dataset. We split the dataset into 80% train and 20% test data which includes 4439 rows of train data and 1110 rows of test data.

3.4.3 One-hot encoding and scaling the dataset. One-hot encoding creates separate binary columns for each category, allowing the model to assign a distinct, independent weight to each category. In this case, Logistic Regression and KNN model need one-hot encoding to ensure mishandling for nominal categories. Scaling is also important for the numerical variables in the dataset, it aims to prevent disproportionate impact the distance calculation compared to features with smaller scales.

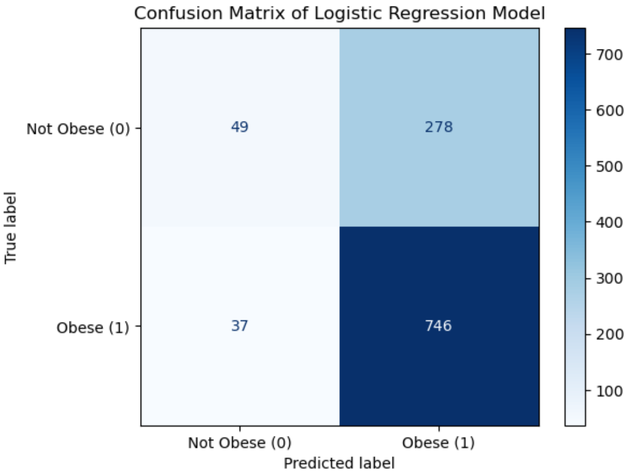
3.4.4 Convert target variables. Convert the continuous target variable into a binary classification problem, we use 1 to present obese and 0 to present not obese.

4 Results

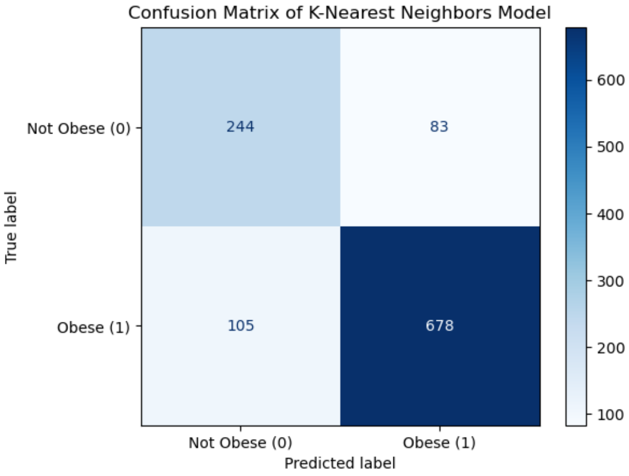
This analysis provides insights into obesity prediction using machine learning models. The following questions were answered using the methodologies detailed in this paper:

4.1 What is the accuracy of the three models

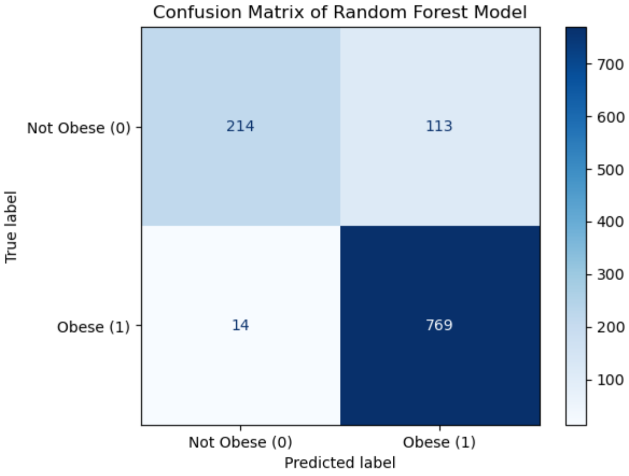
4.1.1 Logistic Regression Model. With the Logistic Regression model, we get an overall accuracy of 71.62%.



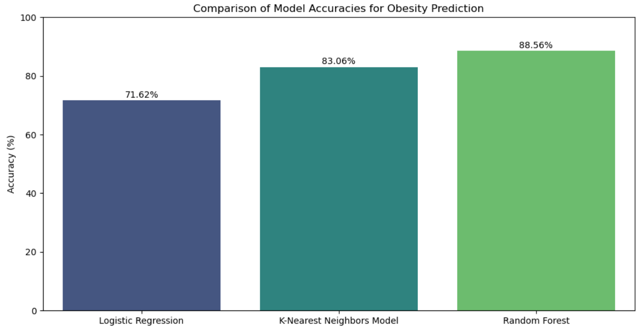
4.1.2 K-Nearest Neighbors Model. With the K-Nearest Neighbors Model, we get an overall accuracy of 83.06%.



4.1.3 Random Forest Model. With the Random Forest Model, we get an overall accuracy of 88.56%.



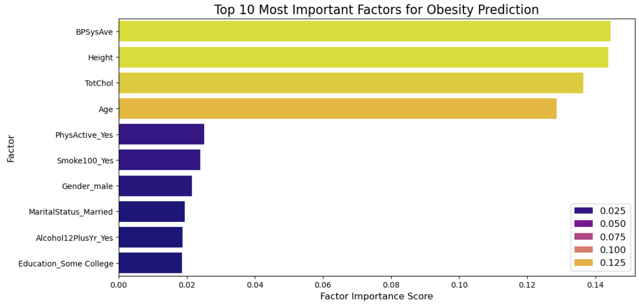
4.1.4 Comparison. The Random Forest Model achieved the highest overall accuracy, correctly classifying patients' obesity status approximately 88.56% of the time. This demonstrates that tree-based ensemble methods are highly effective for this specific prediction task, likely due to their ability to capture complex, non-linear relationships and interactions among the various health factors in the dataset better than the simpler Logistic Regression model or the distance-based KNN algorithm. The Logistic Regression model performed the weakest, suggesting that a simple linear relationship between the input features and the obesity target is insufficient to model the underlying data structure effectively.



4.2 What are the major factors of obesity?

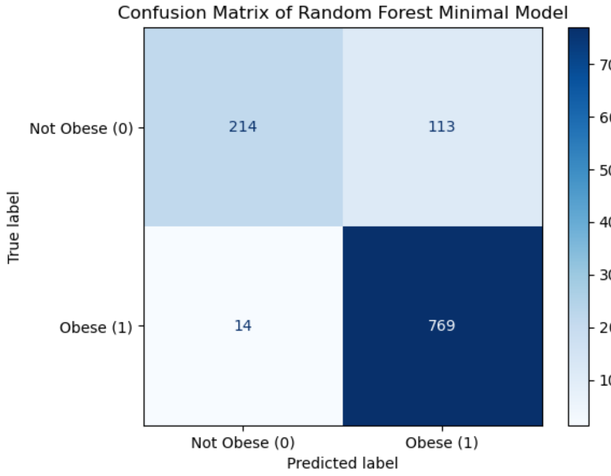
To determine the major factors contributing to obesity using the dataset, we can leverage the feature importance functionality built into the Random Forest model which was the best-performing model.

The four most significant factors (BPSysAve, Height, TotChol, and Age) have substantially higher importance scores (over 0.125 each) compared to the remaining six factors (all under 0.025). This suggests that core physiological measurements and age are far more predictive of obesity in this model than lifestyle choices (physical activity, smoking, alcohol use) or demographic data (gender, marital status, education level).



4.3 Can we build a simpler, highly effective model using fewer predicting variables?

4.3.1 Remove less-important variables. According to the analysis of 4.2, we can only keep the first 4 variables which are highly correlated with the target variable. And we get the same accuracy rate. Thus, we have strong evidence that a simpler, more efficient, and potentially more interpretable model using only those four readily available health factors is as effective as a model requiring dozens of data points.



4.3.2 Using PCA to reduce dimensionality. PCA (Principal Component Analysis) is an unsupervised dimensionality reduction technique. It transforms the original set of features into a new, smaller set of uncorrelated variables called principal components (PCs). But PCA may not help with understanding which health factors matter most, since we cannot interpret a principal component using the name of the original variables.

4.4 How can we prevent obesity by observing the results?

4.4.1 Health & Physiological Factors. BPSysAve (Systolic Blood Pressure Average) & TotChol (Total Cholesterol): These medical metrics have high importance scores, suggesting a correlation between existing cardiovascular risk factors and obesity. Regular health monitoring of blood pressure and cholesterol levels are important. Treating these conditions often involves diet and exercise modifications that simultaneously help with weight management.

4.4.2 Lifestyle Factors. PhysActive_Yes indicates that regular physical activity is a significant protective factor. A low score suggests a sedentary lifestyle is a risk. Smoke100_Yes / Alcohol12PlusYr_Yes suggest these behaviors are associated with obesity status in this model.

4.4.3 Non-modifiable biological factors. Age ang Height, these are Non-modifiable biological factors. Being aware that age impacts metabolism helps prioritize prevention strategies early in life. Height is used in the BMI calculation itself, so its importance is expected.

5 Discussion

This project utilizes Logistic Regression, K-Nearest Neighbors (KNN) and Random Forest and encounters several limitations inherent to each algorithm.

Logistic regression, while interpretable, struggled to capture complex, non-linear relationships in obesity data due to its fundamental assumption of linearity between features and the outcome, leading to potential underfitting and sensitivity to multicollinearity among correlated health metrics.

KNN suffered primarily from computational inefficiency, as it requires calculating distances to every data point during prediction time, making it slow with large datasets and sensitive to high dimensionality and careful feature scaling.

Random Forest proved robust and unlike other models, it doesn't need the scaling of the dataset, but it also presented different challenges. Its main weakness was a lack of transparency. Additionally, training the ensemble of trees was significantly more computationally intensive than simpler models.

6 Conclusion

The project aimed to build machine learning models (KNN, Logistic Regression, and Random Forest) using data from the NHANES dataset to predict obesity and identify its key risk factors. We also identified strong predictors for obesity and confirmed that certain lifestyle and clinical metrics are highly influential.

The real-world effect of these results is the ability to move from general health advice to targeted, data-driven prevention strategies of obesity prediction. Assist public health organizations in understanding how behavioral risks are changing over time across different populations, enabling better resource allocation for preventative programs.

REFERENCES

[1] World Health Organization. (2021, June 9). *Obesity and overweight*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[2] Dugas, L. R., Cao, G., Luke, A. H., & Koepp, G. (2020). A machine learning model to predict obesity risk in young adulthood using demographic and behavioral data. JAMA Network Open, 3(8), e2012765.
DOI: https://doi.org/10.1001/jamanetworkopen.2020.12765

[3] Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics, 99(6), 323–329.
DOI: https://doi.org/10.1016/j.ygeno.2012.04.003

[4] Daniel López Gutiérrez, NHANES Obesity Data. *Kaggle.*
https://www.kaggle.com/datasets/daniellopez01/nhanes-obesity-data