An Analysis of Determinants for Obesity Classification

A Machine Learning Approach to predict obesity

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
- Random Forest
- Public Health

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

*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

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, two machine learning models were employed to classify individuals' obesity risk based on NHANES data. These models include Logistic Regression, a simple and interpretable linear model, and Random Forest, a more complex ensemble method. Their 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.

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 Random forest to be a comparison.

3.2 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.

4 Results

In this part, you need to select a reasonable way to deliver the result of your topic. For example, equation or numerical results, or visualization of your result. You also need to provide a clear explanation of all results and how to understand the results. If there exist any unexpected results, please explain why or possible cause of this special result. You can use subsection 4.1, 4.2, … to separate your results.

4.1 Heading Level 2

Example format: In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.

3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

1.1.1 Heading Level 3. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

1.1.1.1 Heading Level 4. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

5 Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

6 Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what's the effect of this result on the real-world?

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here.

REFERENCES

Use the following ACM Reference format for your citation

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

[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