Statistical Analysis for Obesity Level Prediction

Team 009 Pranavi Chintala[1], Sudheer Kumar Reddy Batthina[1], Bindu Priya Bayyatham[1], Sainath Aittla[1]

Abstract

Obesity arises due to an excessive accumulation of body fat, which can lead to both physical and psychological challenges. Impaired physical functionality has been linked to diminished quality of life in areas including societal distress, sexual function, self-esteem, and occupational well-being. Over recent decades, the incidence of obesity has risen sharply across various demographics, affecting all ages, sexes, and ethnic groups. Notably, individuals with BMI > 40 kg/m^2 have increased in greater proportion compared to those in BMI < 35 kg/m^2. Given the severe health risks linked with obesity, it is crucial to devise effective interventions to counteract these obesity-inducing factors. This includes the implementation of Public Health Planning and Policy and health education programs, which are particularly vital. This research utilized a dataset from an open-source platform, comprising patient records with 17 variables, available through.The efficacy of machine learning techniques in predicting obesity levels was also explored, evaluating model performance based on accuracy.

Keywords: Obesity, Machine Learning, Personalised Recommendations

1. Introduction

Obesity is a chronic disease marked by excess fat that harms health, increasing risks of type 2 diabetes, heart disease, certain cancers, and quality of life. Diagnosis relies on BMI calculation, with a BMI over 25 indicating overweight and over 30 indicating obesity. Although age and gender influence BMI categorization, it involves complex interactions between genetic, environmental, and behavioral factors.

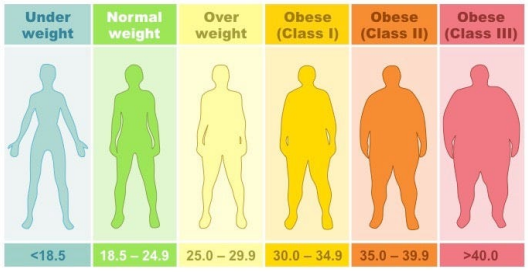

Figure 1: Types of Obesity

Detection of obesity is crucial as the early identification allows for timely intervention and management, potentially preventing or mitigating associated health complications. Detecting obesity also helps healthcare professionals assess an individual's risk for various chronic diseases. Furthermore, by identifying obesity at the population level, public health efforts can be targeted towards implementing preventive measures and policies to curb its prevalence and associated health burdens.

It is important to understand the multifaceted nature of obesity problems as it is not simply a matter of personal choice or willpower; it is influenced by a complex interplay of genetic, environmental, and socio-economic factors. Addressing obesity requires comprehensive, multi-sectoral approaches that encompass individual behavior changes, community interventions, and policy initiatives aimed at promoting healthy lifestyles, improving access to nutritious foods, and creating supportive environments for physical activity.

In this study, we aim to estimate the obesity levels in individuals from the countries of Mexico (70% of adults overweight or obese), Peru (20% of adults are obese), and Colombia (19% obesity rate).Initially, we intend to do exploratory data analysis along with data visualisation. Following we use stepwise forward and backward subset selection methods for feature selection. Then, we plan to use machine learning models such as Logistic Regression, Naive Bayes, LDA, QDA, KNN, Decision Tree, Random Forest and Multi-Perceptron model to estimate the obesity levels with the performance metric set to AIC or accuracy. Further, we will validate the model using K-fold Cross Validation.Finally, we aim to develop a web application with dual functionality: collecting data to test the best model and providing personalized suggestions.

2. Literature Review

A decent amount of research and model development have been done in predicting the obesity level based on different predictors. A few research articles are mentioned below:

M. Kivrak [1] and S. Kitis [2] explored using deep learning to predict obesity levels based on eating habits

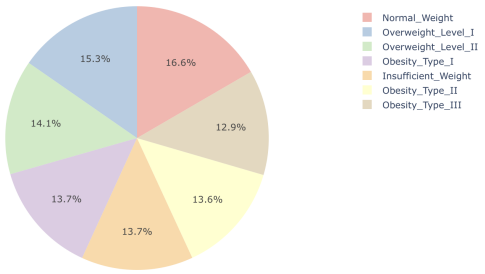

Figure 2: Obesity Levels

and physical condition. While [1] emphasizes the need for effective interventions to address the increasing prevalence of obesity, [2] emphasizes the urgent need for effective interventions to address the rising prevalence of obesity. F. Musa [3] explored machine learning techniques to predict obesity status using different algorithms, aiming to develop an automated predicting model for disease occurrence. The research demonstrates promising results, with the Gboost classifier achieving the highest accuracy of 99.05% among the classifiers used. A. Hapfelmeier [4] explored a novel approach for variable selection in Random Forests based on permutation tests, which demonstrates superior performance in distinguishing relevant from irrelevant variables.

A. R. Lubis [5] examined the use of the Euclidean distance formula in the K-Nearest Neighbor (KNN) classification method, comparing it with normalized Euclidean, Manhattan, and normalized Manhattan distances to optimize nearest neighbor distance calculations for improved results. S. Garba [6] presented a method to enhance obesity level classification accuracy using boosting and bagging techniques with decision trees and naïve Bayes models, validated through empirical evaluation. F. H. Yagin [7] used neural network models and feature selection algorithms like chi-square, F-Classify, and mutual information to predict obesity levels. Bayesian optimization techniques were employed to optimize model hyperparameters effectively. X. Zou [8] enhanced traditional logistic regression for binary classification by optimizing the Sigmoid function and reducing training iterations using the gradient descent method.A. R. Lubis [9] reviewed the challenges of learning from imbalanced datasets in classification tasks, discusses existing methods to handle such imbalances, and explores future trends that may improve outcomes.

S. Yadav [10] explored the efficacy of k-fold cross-validation versus hold-out validation across dataset sizes, presenting experimental findings that k-fold can be preferable for achieving high-quality classification results. F. M. Palechor [11] introduced a dataset for estimating obesity levels based on dietary habits and physical condition, comprising 17 attributes and 2111 records from Mexico, Peru, and Colombia, with synthetic and real data. A. Kerkadi [12] evaluates the link between lifestyle factors like physical inactivity and diet, and obesity among adolescents in Qatar, using data from the Arab Teens Lifestyle Study to show significant associations between lifestyle habits and general and abdominal obesity.

3. Dataset

The dataset utilized for analysis was obtained and encompasses data for estimating obesity levels among individuals aged 14 to 61 years, representing diverse eating habits and physical conditions in Mexico, Peru, and Colombia. It comprises patient records containing 17 variables and a total of 2111 entries.A detailed explanation of the variables is given in Table I.

4. Exploratory Data Analysis

4.1. Data Transformation

Given that the predictor variables, including Gender, CALC, FAVC, SCC, SMOKE, Family history with overweight, CAEC, MTRANS, and NObeyesdad, consist of categorical data, we opt factor conversion to convert all these predictors datatype to factors.

4.2. Data Preprocessing

This stage involved significant preprocessing tasks such as handling extreme values and missing data analysis, which are crucial to ensure the quality and integrity of the data. The preprocessing steps help in refining the dataset, making it suitable for the modeling process by reducing noise and handling anomalies.

4.3. Feature Selection

After performing the forward and backward stepwise selection, our top five features consistently include Weight, Height, Age, CALCSometimes, and CALCFrequently.

5. Plan of Actions

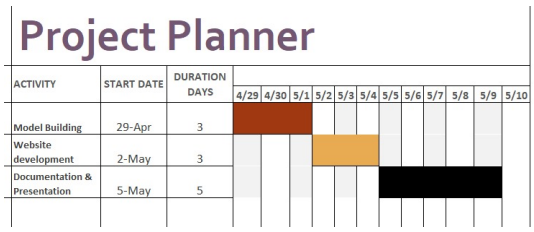

Figure 3: Obesity Levels

Predictor	Description
Obesity	Level Target (1:Insufficient Weight (BMI<18.5), 2:Normal Weight (18.5 to 24.9), 3:Overweight (25 to 29.9), 4:Obesity Type I (30 to 34.9), 5: Obesity Type II (35 to 39.9), 6: Obesity Type III (BMI>40))
Age	Age
Gender	Gender (1:male, 0:female)
Height	Height
Weight	Weight
History	Family History have overweight (1:Yes, 0: No)
FAVC	Eat High Caloric Food Frequently (1:Yes, 0: No)
FCVC	Frequency Eating Vegetables (1:Never, 2:Sometimes, 3:Always)
NCP	Number of main meals (Between 1 y 2, Three, More than three)
CAEC	Consumption of food between meals (0:No, 1:Sometimes, 2:Frequently, 3:Always)
Smoke	Smoking (1:Yes, 0: No)
CH2O	Consumption of water daily (Less than a liter, between 1and 2 L, More than 2 L)
SCC	The attributes related to the physical condition are: Calories consumption monitoring (1:Yes, 0: No)
FAF	Physical activity frequency (Not have, 1 or 2 days, 2 or 4 days, 4 or 5 days)
TUE	Time using technology devices (0-2 hours, 3-5 hours, More than 5 hours)
CALC	Alcohol Consumption (0:No, 1:Sometimes, 2:Frequently, 3:Always)
MTRANS	Mode of Transport (1:Automobile, 2:Motorbike, 3:Bike, 4: Public Transportation, 5:Walking)

Table 1: Predictors along with their description

5.1. Completed Activities

Previously, we have completed the exploratory data analysis along with data visualisation, preprocessesing and transformation. Also, we have completed forward and backward step-wise Feature selection.Now we are presently in model building phase.Simultaneously we are working to develop the webapp.

5.2. Under Construction

For the implementation of an optimized model using the dataset described in the study on obesity levels, we aim to refine our approach to enhance predictive accuracy and efficiency. This involves leveraging the comprehensive exploratory data analysis (EDA) already conducted to identify the most influential variables and relationships that affect obesity levels. With the insights gained from the EDA, we will focus on optimizing the machine learning model. Key to this optimization is the fine-tuning of hyperparameters such as number of nodes, number of layers, which will be accomplished using techniques like grid search and cross-validation. These methods ensure that our model not only performs well on training data but also generalizes effectively to unseen data. We aim to reduce overfitting, thereby improving the model's robustness. The goal is to deploy a model that accurately predicts obesity levels, providing a tool that can assist in personalized health interventions and contribute to broader public health strategies.

6. Proposed Methodolgy

Here, we are using Logistic Regression, Naive Bayes, LDA, QDA for interpretability and for the performance we are taking into consideration SVM, Decision Tree, Random Forest and Percpetron model.

7. Results and Discussion

Under Construction

8. Summary and conclusions

Acknowledgements

9. References

[1] M. Kivrak, "Deep Learning-Based Prediction of Obesity Levels According to Eating Habits and Physical Condition," vol. 6, no. 1, p. 2021, doi: 10.52876/jcs.939875.
[2] S. Kitis and H. Goker, "Detection of Obesity Stages Using Machine Learning Algorithms," Anbar Journal of Engineering Sciences, vol. 14, no. 1, pp. 80–88, May 2023, doi: 10.37649/AENGS.2023.139350.1045.
[3] F. Musa, F. Basaky, and O. E.O, "Obesity prediction using machine learning techniques," Journal of Applied Artificial Intelligence, vol. 3, no. 1, pp.

24–33, Jun. 2022, doi: 10.48185/JAAI.V3I1.470.
[4] A. Hapfelmeier and K. Ulm, “A new variable selection approach using Random Forests,” Comput Stat Data Anal, vol. 60, no. 1, pp. 50–69, 2013, doi: 10.1016/j.csda.2012.09.020.
[5] A. R. Lubis, M. Lubis, and A.- Khowarizmi, “Optimization of distance formula in K-Nearest Neighbor method,” Bulletin of Electrical Engineering and Informatics, vol. 9, no. 1, pp. 326–338, Feb. 2020, doi: 10.11591/eei.v9i1.1464.
[6] S. Garba, M. Abdullahi, U. A. Umar, and N. T. Wurnor, “Obesity Level Classification Based on Decision Tree and Naïve Bayes Classifiers,” Journal of Science and Technology (SLUJST, vol. 3, no. 1, pp. 113–121, 2022.
[7] F. H. Yagin et al., “Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique,” Applied Sciences 2023, Vol. 13, Page 3875, vol. 13, no. 6, p. 3875, Mar. 2023, doi: 10.3390/APP13063875.
[8] X. Zou, Y. Hu, Z. Tian, and K. Shen, “Logistic Regression Model Optimization and Case Analysis,” Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019, pp. 135–139, Oct. 2019, doi: 10.1109/ICCSNT47585.2019.8962457.
[9] A. Ali, A. Ralescu, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A review,” Classification Int. J. Advance Soft Compu. Appl, vol. 5, no. 3, 2013, Accessed: Apr. 16, 2024. [Online]. Available: https://www.researchgate.net/publication/288228469
[10] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” Proceedings - 6th International Advanced Computing Conference, IACC 2016, pp. 78–83, Aug. 2016, doi: 10.1109/IACC.2016.25.
[11] F. M. Palechor and A. de la H. Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico,” Data Brief, vol. 25, p. 104344, Aug. 2019, doi: 10.1016/J.DIB.2019.104344.
[12] A. Kerkadi et al., “The Relationship between Lifestyle Factors and Obesity Indices among Adolescents in Qatar,” Int J Environ Res Public Health, vol. 16, no. 22, Nov. 2019, doi: 10.3390/IJERPH16224428.

References