Analysis of Obesity Factors and Prediction Models

Dinesh Ram Veerappan Kosal, DeVos School of Management, Northwood University.

Abstract

This paper has been presented based on analysis conducted on various factors contributing to obesity levels in individuals. The dataset used was comprehensive which can help explore the relationships between multiple features such as demographics, dietary and other lifestyle factors along with their impact on obesity. The aim is to identify key predictors of obesity using exploratory data analysis, regression models, classification techniques and clustering algorithms. The findings provided insights into the complex nature which helps develop potential strategies which can help interventions in obesity prevention and management.

Problem statement

The dataset labeled as obtained from UCL dataset repository which is named as the file: "ObesityDataSet_raw_and_data_sinthetic.csv" contains information about various factors that may contribute to obesity. The goal of this analysis is to explore the relationships between these factors and obesity levels, develop predictive models to categorize individuals into different obesity categories, and identify patterns or clusters within the data. This aim of the analysis is to provide insights that could be valuable for public health initiatives, personalized health recommendations, and understanding the complex nature of obesity.

Introduction

Obesity is a growing global health concern with significant implications for individual well-being and public health systems. Understanding the factors that contribute to obesity is crucial for developing effective prevention and intervention strategies. This study utilizes a comprehensive dataset containing information on various demographic, dietary, and lifestyle factors to analyze their relationship with obesity levels.

The objectives of this study are:
1) To identify key factors associated with obesity through exploratory data analysis.
2) To develop and evaluate predictive models for obesity levels using regression and classification techniques.
3) To explore potential underlying patterns or groups within the data using clustering algorithms.
4) To provide insights that can inform targeted interventions for obesity prevention and management.

Methodology

This study employed a multi-faceted approach to analyze the obesity dataset, including exploratory data analysis, regression modeling, classification techniques, and clustering algorithms.

A. Exploratory Data Analysis

We conducted exploratory data analysis to understand the relationships between various factors and obesity levels.

The analysis of the dataset has produced multiple visualizations that can help define the dataset. And as shown in multiple images the graphs are mentioned to denote the feature and identify multiple aspects of the dataset.

The graph presented in Fig 1, is that graph of a correlation matrix that helps us explore the dataset and as for the values on the axis are the feature available in the dataset that help find the correlation between each pair.

The diagonal from top left to bottom right is red, indicating a perfect correlation of each variable with itself. A positive correlation of 0.84 suggests that as height increases, weight tends to increase.

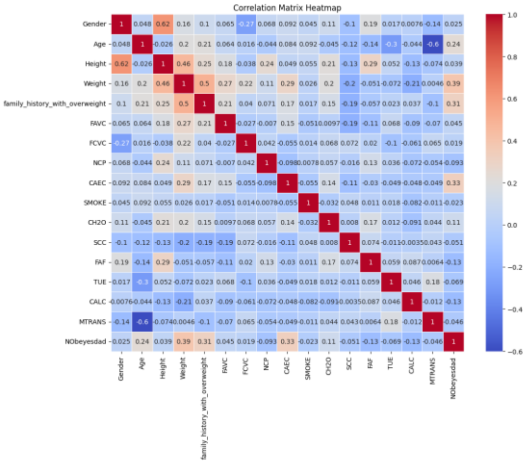

Fig. 1. Correlation matrix of key variables

From the figure, we also can see a moderate positive correlation of 0.46 between weight and age. It also shows a positive correlation of 0.39 which indicates a potential link between family history and obesity levels.

The heatmap also helps us identify relationships between variables, which can be useful for hypothesis testing, data analysis and understanding various potential factors that influence obesity.

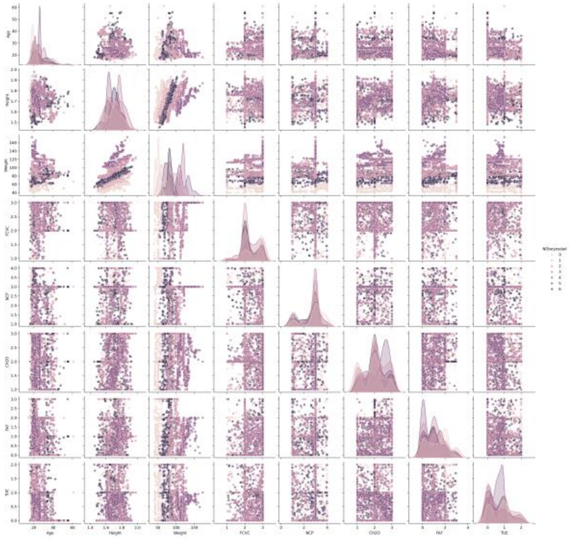

Fig. 2. Pair plot of selected variables

The Fig 2 shows a pair plot, which is a matrix of scatterplots that help visualize pairwise relationships between multiple variables in the Obesity dataset.

The diagonal plots show the distribution of each variable individually. Some appear to be continuous while other variables seem categorical.

The pair plot provides a comprehensive overview of the relationship between the variables in the dataset, which can be used for understanding obesity related factors for further analysis or modelling.

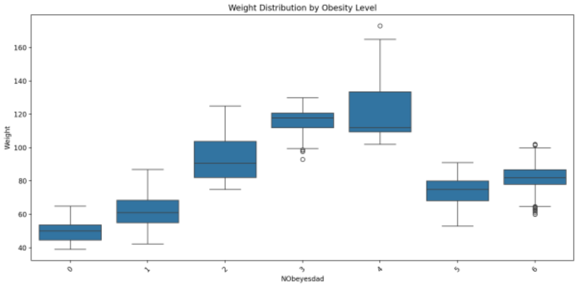

Fig. 3. Boxplot of weight distribution across obesity levels

The fig 3 shows the boxplot which provides insights into how weight relates to different obesity classifications, showing both the central tendencies and the spread of weight data for each category.

B. Regression Analysis

We performed linear regression to predict obesity levels based on the available features.

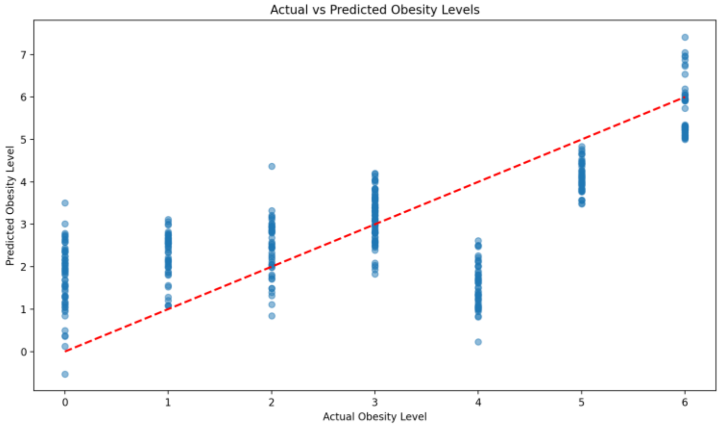

Fig. 4. Actual vs Predicted obesity levels

The fig 3 shows a scatterplot produced containing “actual obesity levels” vs “Predicted obesity levels”. The graph helps visualize the performance of a predictive model for obesity levels, showing both its successes and the limitations of accurately predicting obesity across different actual levels.

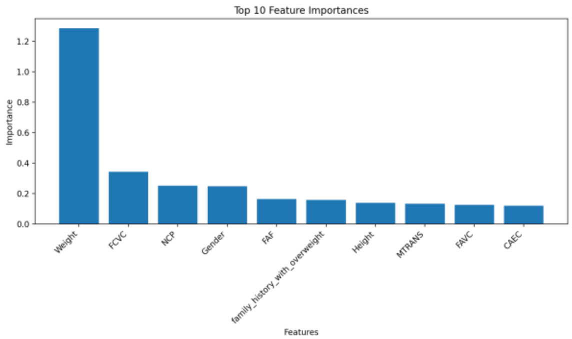

Fig. 5. Feature importance in predicting obesity levels

C. Clustering Analysis

We applied three clustering algorithms: K-means, Hierarchical Agglomerative Clustering (HAC), and DBSCAN to identify potential patterns in the data.

Fig. 6 visualizes the results of our clustering analysis.

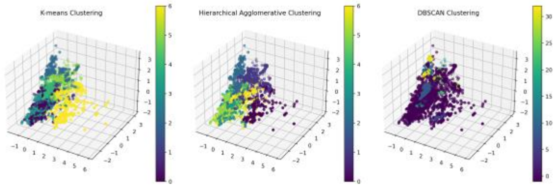

Fig. 6. Visualization of clustering results

Fig. 7 shows the distribution of samples across different clusters for each clustering method.

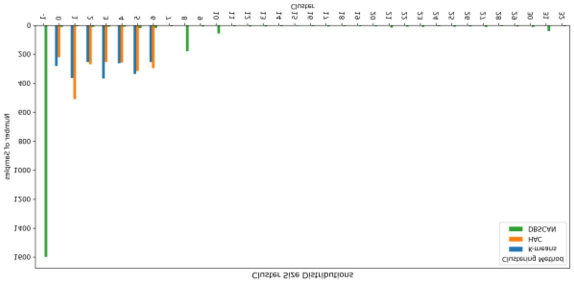

Fig. 7. Distribution of samples across clusters

Results

The performance of the models was evaluated using various metrics. For classification models, accuracy, precision, recall, and F1-score were used. The SVM model achieved an accuracy of 96%, while the Decision Tree model achieved 94%. For regression, the R-squared value was 0.5056, indicating moderate predictive power. Clustering evaluation was performed using the Silhouette Score, with K-means achieving the highest score of 0.1658. Visualizations such as confusion matrices and ROC curves were used to further assess model performance.

Discussion

The results of this study highlight the significant role of weight, dietary habits, and physical activity in predicting obesity levels. The classification models demonstrated high accuracy, indicating their potential for practical applications in obesity risk assessment. Challenges encountered included handling missing data and ensuring model generalizability, which were addressed through data preprocessing and cross-validation. Ethical considerations include the potential for bias in model predictions and the need for careful interpretation of results in real-world settings. The findings underscore the importance of personalized interventions and the potential for machine learning to inform public health strategies.

Conclusion

It can be seen that Weight, eating patterns, and physical activity are some of the major causes of obesity that this study has effectively uncovered. The models that have been built show a high level of accuracy in forecasting obesity levels, which may find use in tailored health evaluations and focused intervention tactics. Future studies could concentrate on investigating more sophisticated machine learning approaches, like deep learning, for better predicting performance and integrating longitudinal data to record temporal changes in obesity risk factors. Furthermore, examining how socioeconomic and environmental factors affect obesity may offer a more thorough comprehension of this intricate health problem.

References

1. Jiang, Y., Jiang, H., Feng, Y., Hu, Y., Jiang, Y., & Jiang, H. (2023). Predicting risk of obesity in overweight adults using machine learning models: A retrospective cohort study. PLoS One, 18(12). https://doi.org/10.1371/journal.pone.0294847

2. Kim, J., Lee, S., & Park, H. (2024). Prediction model of weight control experience in men with obesity: A cross-sectional study using the Korea National Health and Nutrition Examination Survey. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-70833-5

3. Ren, Y., Luo, X., Jiang, Q., Li, X., & Chen, G. (2023). Prediction models for children/adolescents with obesity: A systematic review and meta-analysis. Preventive Medicine, 172, 107614. https://doi.org/10.1016/j.ypmed.2023.107614

4. Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2021). Predicting Obesity in Adults Using Machine Learning Techniques Based on Cardiovascular Risk Factors. Frontiers in Nutrition, 8, 669155. https://doi.org/10.3389/fnut.2021.669155

Github Link: https://github.com/dineshram18/Final_Project_ML.git