Estimation on Obesity Levels

Ahmed Riasat
Department of Computer Science
BRAC University
Dhaka, Bangladesh
Email: ahmed.riasat@g.bracu.ac.bd.com

***Abstract*—This paper presents the development of a machine learning model for predicting obesity levels based on demographic, lifestyle, and nutritional data. Using a comprehensive dataset with 2111 entries and 17 features—including age, weight, height, and various behavioral factors—the study employs preprocessing techniques, exploratory data analysis, and both classification and clustering methods. Experimental results indicate that tree-based models such as Decision Trees and Random Forests achieve high accuracy, while K-Means clustering, supported by principal component analysis, reveals meaningful groupings in the data.**

***Index Terms*—Obesity estimation, Machine learning, Data science, Classification, Clustering, PCA, Support vector machine.**

I. Introduction

Obesity has emerged as a major public health challenge due to modern sedentary lifestyles and unhealthy eating habits. The objective of this project is to develop a predictive model that estimates obesity levels using real-world data encompassing age, weight, height, family history, and related lifestyle factors. This data-driven approach aims to support early intervention strategies and promote healthier living.

II. Dataset Description

The dataset was obtained from the UCI Machine Learning Repository [1] and comprises 2111 data points with 17 features. These features include both categorical (e.g., Gender, family_history_with_overweight) and continuous variables (e.g., Age, Weight, Height). The target variable, *NObeyesdad*, categorizes individuals into seven obesity-related classes.

III. Correlation Analysis

Correlation analysis was performed to understand relationships among the features. Notable correlations were observed, for instance, between family history of overweight and obesity levels as well as between caloric tracking and weight gain. These insights played a crucial role in feature selection and understanding the data distribution.

IV. Class Distribution

The target variable exhibits seven distinct weight classes ranging from Insufficient Weight to Obesity Type III. Visualization of class distribution (e.g., via count plots) revealed the frequency of each category and assisted in addressing potential imbalances during model validation.

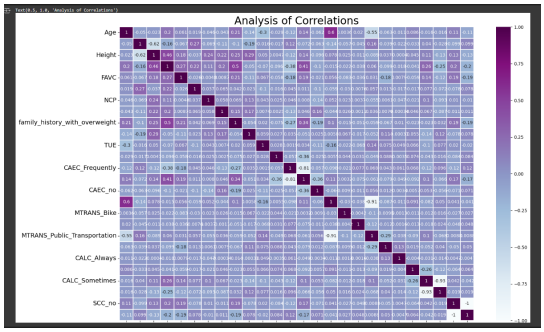


Fig. 1: Corelation Analysis of Features.

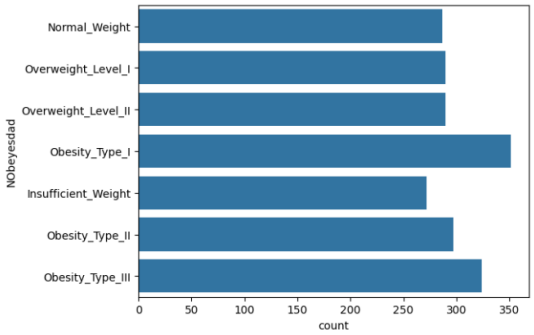


Fig. 2: Class Distribution.

V. Dataset Preprocessing and Feature Scaling

The preprocessing phase involved:

- There are no missing values in the dataset.
- Encoding categorical features using one-hot encoding.
- Standardizing numerical features with techniques such as log scaling via `RobustScaler`, due to their positively skewed nature.
- Based on relationship between obesity and features they were transformed and encoded.

VI. Exploratory Data Analysis (EDA)

EDA was conducted to explore relationships among different features and the target obesity levels. Visualizations highlighted trends such as the impact of gender, family history, and dietary habits on obesity. These analyses informed the selection of machine learning models and preprocessing methods.

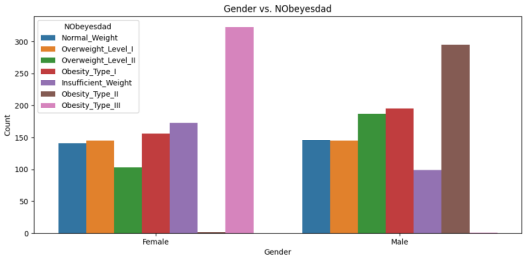


Fig. 3: Gender has some relation to being overweight.

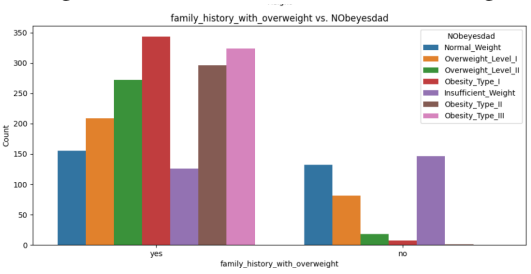


Fig. 4: The data clearly suggests that family history with overweight has direct part in a person being overweight.

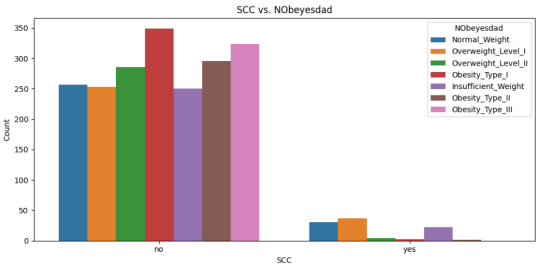


Fig. 5: SCC feature indicates if a person tracks their calorie intake. It is evident from the chart below is that not tracking calories lead to gaining weight.

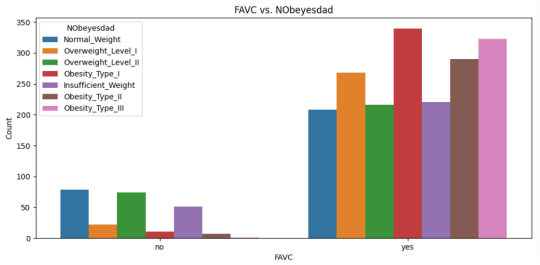


Fig. 6: FAVC indicates eating high calorie foods. The chart show us that high calorie diet is related to being overweight.

VII. Dataset Splitting

For model development, the dataset was partitioned into training (80%) and testing (20%) sets. This split ensured that the model was both well-trained and capable of generalizing to unseen data.

VIII. Model Training and Testing

We tested multiple classification models:

- **Decision Tree** – It is a tree structured algorithm where it splits data into subsets based on features, creating a flowchart-like structure. Each node represents a feature and each branch a decision, with each leaf representing an output. It optimizes splits using impurity measures such as 'Gini' (default) or 'Entropy'.

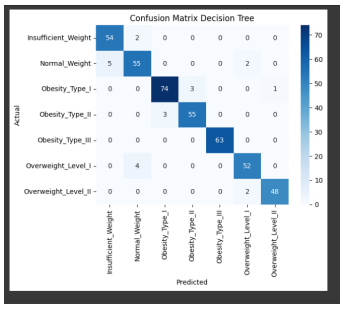


Fig. 7: Performance overview of Decision Tree.

- **Logistic Regression** – It is a statistical ML algorithm that utilizes ordinal regression for multi-class tasks. It uses the sigmoid logistic function to map predicted values and finds a linear relationship between input features and the target label.

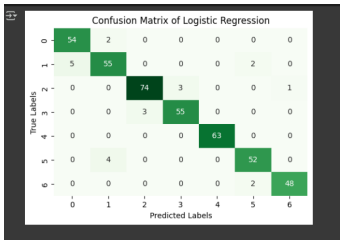


Fig. 8: Performance overview of Logistic Regression.

- **K-Nearest Neighbors (KNN)** – It is a simple non-parametric ML algorithm that makes predictions by storing the training data and predicting the output by finding the K nearest neighbors in the data.

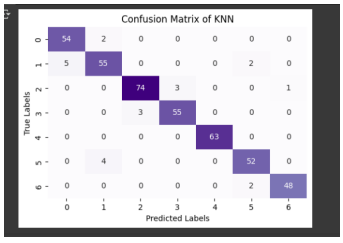


Fig. 9: Performance overview of K-Nearest Neighbors.

- **Random Forest** – An ensemble method that works by aggregating predictions from multiple decision trees to enhance accuracy and reduce overfitting.

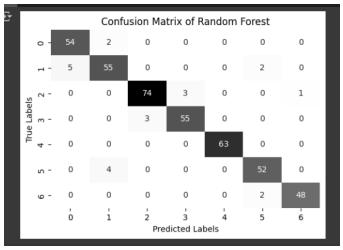


Fig. 10: Performance overview of Random Forest.

- **Gaussian Naive Bayes** – A probabilistic classifier based on Bayes' theorem, assuming feature independence.

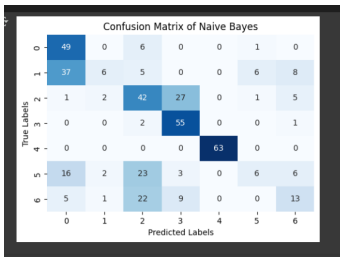


Fig. 11: Performance overview of Naive Bayes.

- **Support Vector Machines (SVM)** – Implemented with an RBF kernel to handle the nonlinear nature of the data.

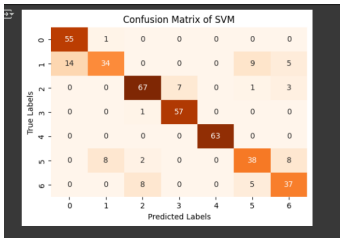


Fig. 12: Performance overview of Support Vector Machines.

Performance metrics such as accuracy, precision, recall, and confusion matrices were used for model evaluation. The experimental results indicated that Decision Tree and Random Forest models achieved the highest accuracy (0.94) and f1 scores (0.95).

IX. Clustering and Visualization

K-Means clustering was applied to the data to uncover hidden groupings. In order to visualize these clusters, Principal Component Analysis (PCA) was used to reduce the feature space to three dimensions. The clusters were then visualized in a 3D scatter plot, illustrating how the data naturally partitions into groups that correspond well with the underlying obesity levels.

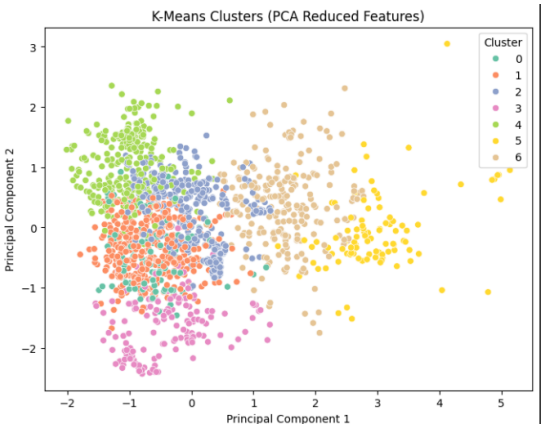
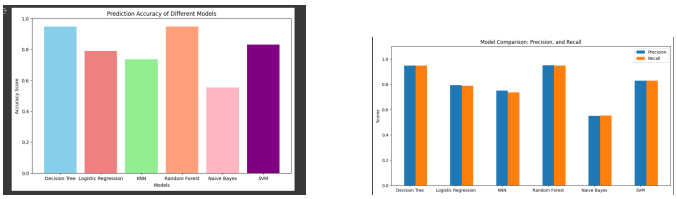


Fig. 13: Clustering of data using PCA and K-Means.

X. Result Analysis and Model Selection

A comparative analysis of all models was performed. Based on the accuracy, precision, and recall scores, tree-based models (Decision Tree and Random Forest) emerged as the most promising. The clustering results further validated the discriminatory power of the selected features.



(a) Model Comparison (b) Performance Metrics

Fig. 14: Comparison of models based on various evaluation metrics.

XI. Conclusion

The project successfully demonstrates the application of machine learning to the estimation of obesity levels using real-world data. By integrating classification and clustering techniques along with robust preprocessing and visualization steps, the proposed approach not only predicts obesity with high accuracy but also provides insights into the influences of various lifestyle factors. Future work will focus on further refining the models and exploring advanced ensemble methods.

Acknowledgments

The author thanks the UCI Machine Learning Repository for providing the dataset and acknowledges the support from the faculty of the Data Science course.

References

[1] Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.