PREDICTING OBESITY LEVELS USING MACHINE LEARNING AND DEEP LEARNING METHODS

SUBMITTED BY GROUP 3

Title:

Predicting Obesity Levels Using Machine Learning and Deep Learning Methods

(Infosys Springboard 5.0 Project)

Team Members:

1	Kachibhotla Naga Sai Siriakshaya
2	Naga Manoj Kumar
3	Adithya Prasad G
4	Janhavi Bhandare
5	Kalavakunta Uday Venkata Krishna
6	Boreddy Supriya Reddy
7	Ruthvik Chowdary

Project Guide:

Narendra Kumar

Date:

November 2024

1. Introduction:

This project aims to develop a predictive model to assess obesity levels based on eating habits and physical conditions using machine learning (ML) and deep learning (DL) techniques. By leveraging the "Obesity based on eating habits and physical conditions" dataset from Kaggle, our objective is to build a system that classifies obesity levels and helps identify at-risk individuals based on their lifestyle patterns. The outcome will be a model that predicts obesity levels such as insufficient weight, normal weight, overweight, and various types of obesity.

2. Problem Statement:

Obesity is a growing global concern, and the ability to predict obesity levels based on lifestyle habits can help with early intervention. This project focuses on analyzing the eating habits and physical conditions of individuals to predict their obesity levels. The dataset provides information on various features like food consumption habits, physical activity, and the use of technology devices.

3. Dataset Overview:

In this project, we explore multiple datasets related to obesity and lifestyle factors. Below are the details of the datasets:

3.1 Primary Dataset (Selected):

"Obesity based on eating habits and physical conditions" from Kaggle, created by Lesumit Kumar Roy. This dataset provides a comprehensive set of 17 attributes related to eating

habits, physical activity, and demographic information. It is suitable for our project's objective of predicting obesity levels accurately.

Dataset Link (Finalized):

Obesity Risk EDA & Prediction Dataset

3.2 Explored Datasets (Not Selected):

- **Dataset 1**: Link to Dataset 1

  Reason for not selecting: Upon exploring this dataset, we found that it lacked certain key attributes such as detailed eating habits and physical condition factors, which are crucial for our project's objectives. Additionally, the data quality and structure did not align well with the requirements for building predictive models.

- **Dataset 2**: Link to Dataset 2

  Reason for not selecting: This dataset provided limited information on physical conditions and was missing critical obesity class labels. The available features were not detailed enough for building accurate machine learning models focused on obesity prediction.

- **Dataset 3**: Obesity Dataset

  Reason for not selecting: Although this dataset was comprehensive, it had several missing values and inconsistencies, which would have required extensive data cleaning, potentially affecting model accuracy.

  Due to these limitations, we opted to proceed with the primary Kaggle dataset, which offered more

comprehensive and high-quality data relevant to our project.

3.3 Attributes:

3.3.1 Attributes related to eating habits:

- Frequent consumption of high-caloric food (FAVC)
- Frequency of vegetable consumption (FCVC)
- Number of main meals (NCP)
- Consumption of food between meals (CAEC)
- Daily water consumption (CH20)
- Alcohol consumption (CALC)

3.3.2 Attributes related to physical condition:

- Calories consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)

3.3.3 Other variables:

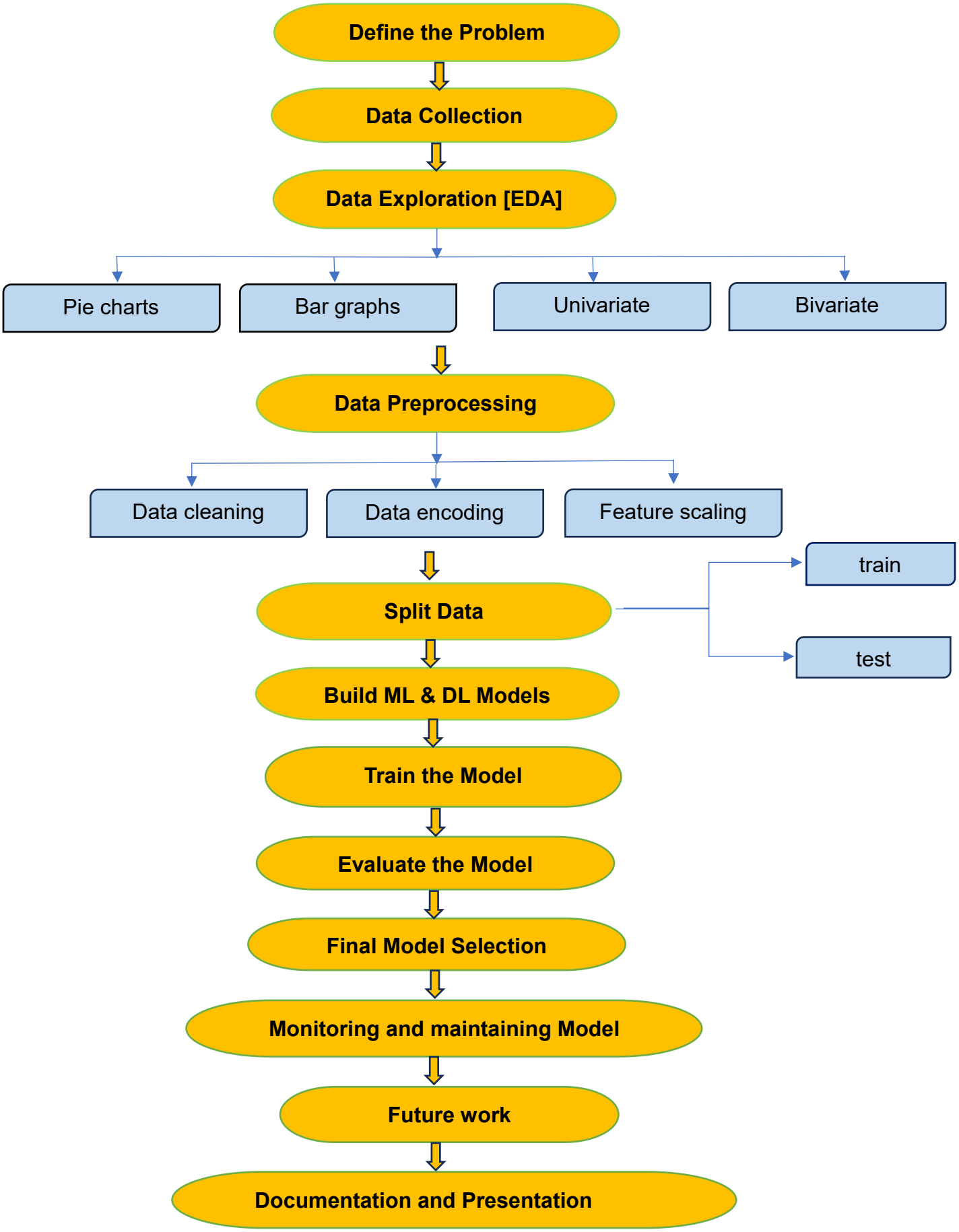
- Gender, Height, Weight

3.3.4 Target variable:

N Obesity (Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, problem

Obesity Type II, Obesity Type I

4. Flow Chart:


Define the Problem
Data Collection
Data Exploration [EDA]
Pie charts
Bar graphs
Univariate
Bivariate
Data Preprocessing
Data cleaning
Data encoding
Feature scaling
Split Data
train
test
Build ML & DL Models
Train the Model
Evaluate the Model
Final Model Selection
Monitoring and maintaining Model
Future work
Documentation and Presentation

5. Data Collection:

The dataset was collected from Kaggle and contains information about various lifestyle attributes affecting obesity levels. The data was labeled based on WHO and Mexican norms, and the class variable "NObesity" was created to classify individuals' obesity levels.

6. Data Preprocessing:

6.1 Data Cleaning:

The columns were renamed for better readability. For example, "FAVC" was renamed to "High caloric food," and "NObeyesdad" was renamed to "Obesity_level."

6.2 Data Encoding:

Categorical variables like gender, transportation, and obesity level were label-encoded to convert them into numeric format for machine learning models.

6.3 Feature Scaling:

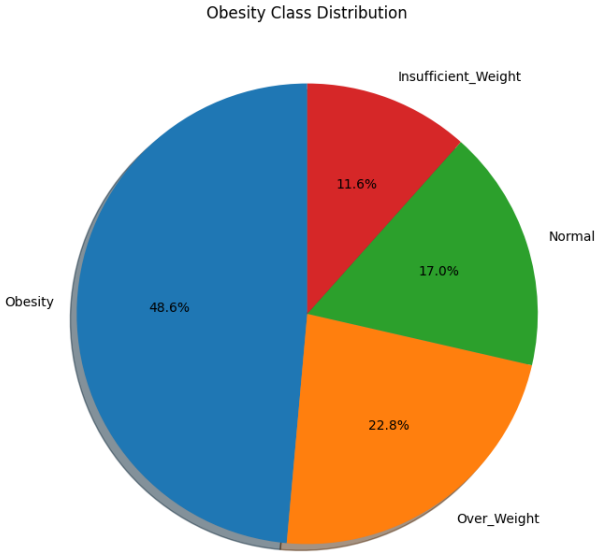
Features like age, height, weight, and others were scaled using StandardScaler to ensure all features have equal importance during training.

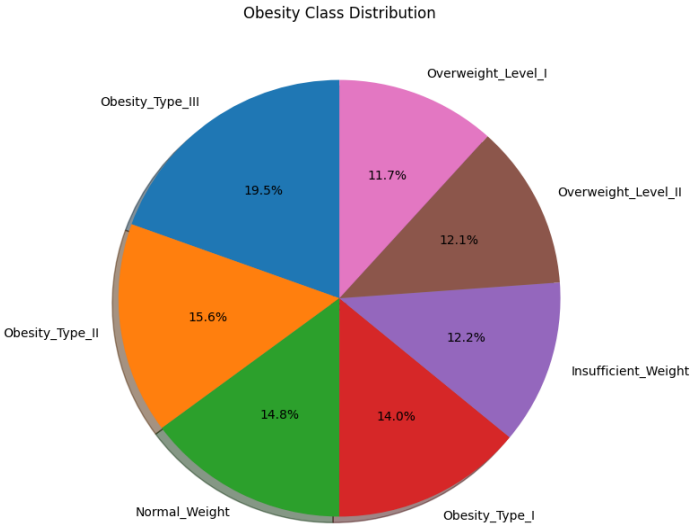
7. Exploratory Data Analysis (EDA):

Several visualizations were generated to better understand the data and the distribution of obesity levels:

7.1 Pie Chart: Obesity Class Distribution

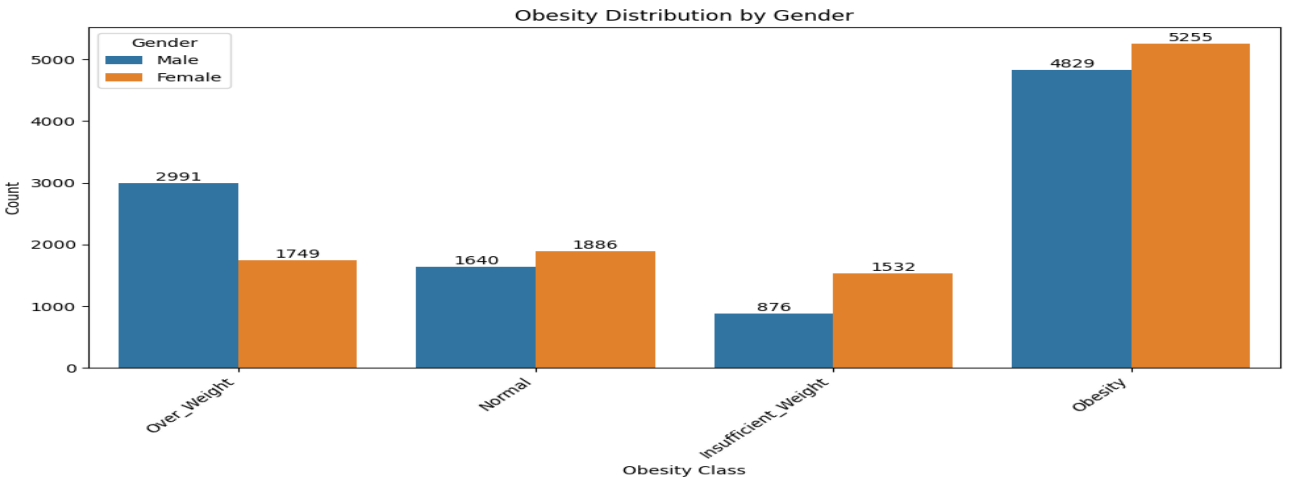
pie chart of 7 classes and 4 classes to illustrate the distribution of obesity classes as percentages, providing a clear visualization of their proportions within the dataset.

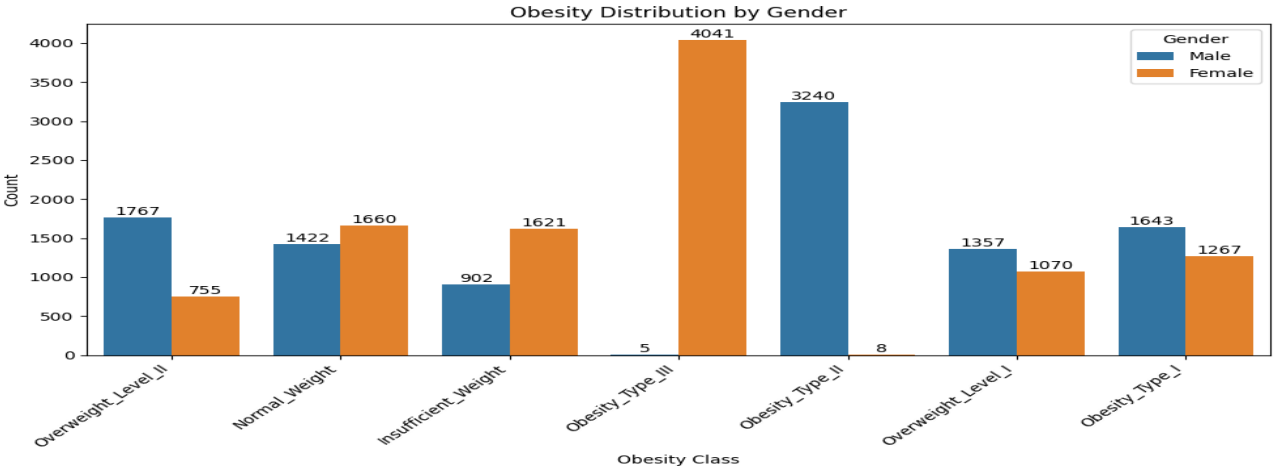




7.2 Bar Graph: Gender-Specific Obesity Distribution

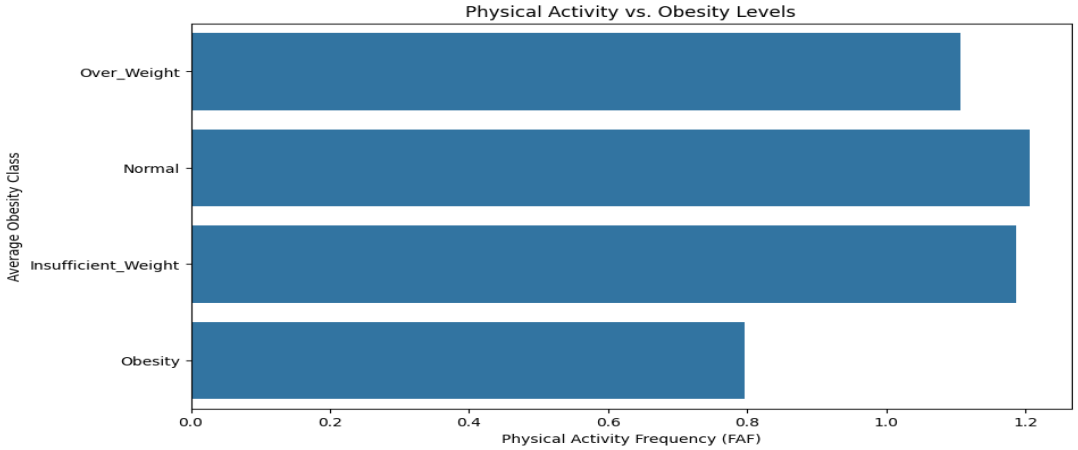
Bar graphs of 7 classes and 4 classes were generated to visualize the distribution of obesity across males and females, highlighting gender-specific patterns in the dataset.

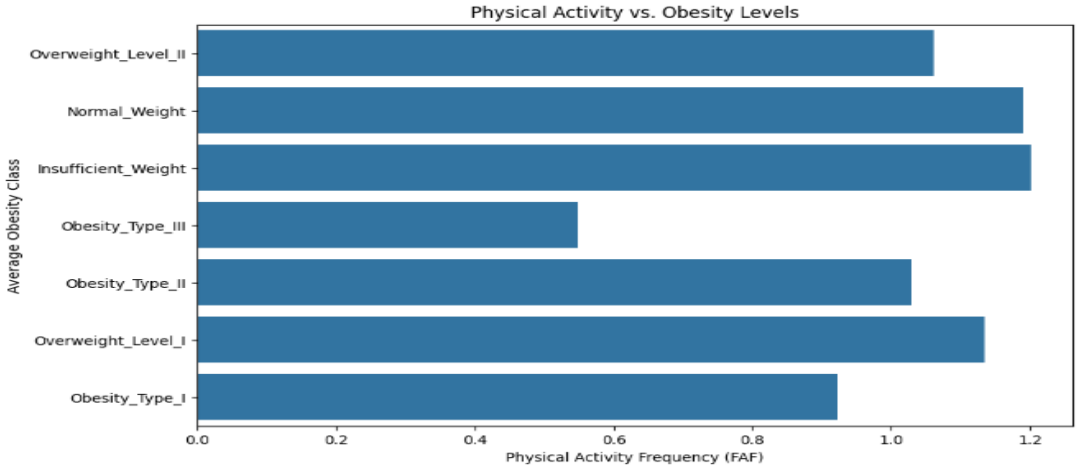




7.3 Bar Plot: Physical Activity and Obesity Correlation

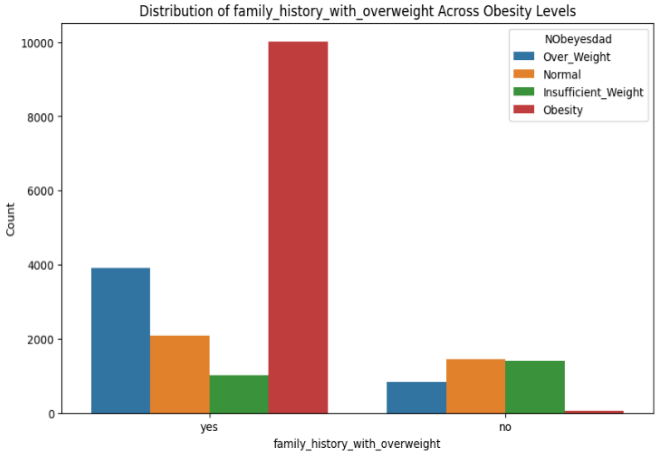
A bar plot of 7 classes and 4 classes was generated to depict the relationship between physical activity levels and obesity, highlighting potential correlations.

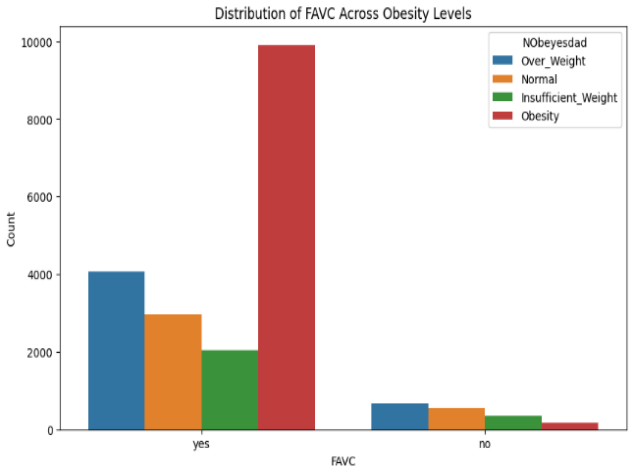


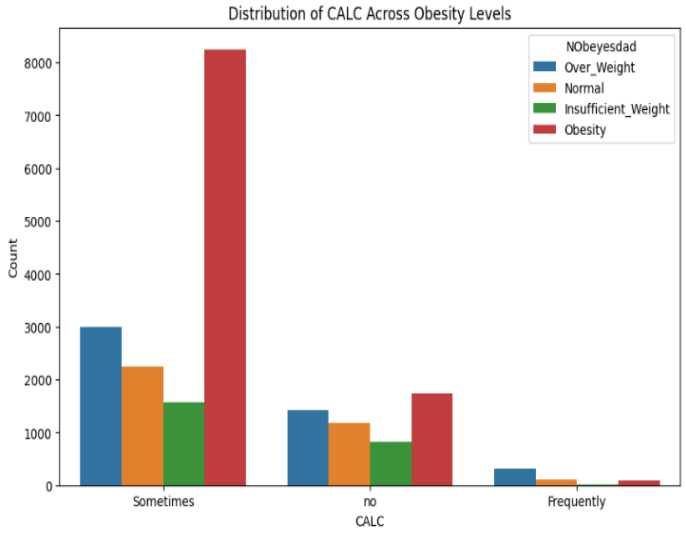


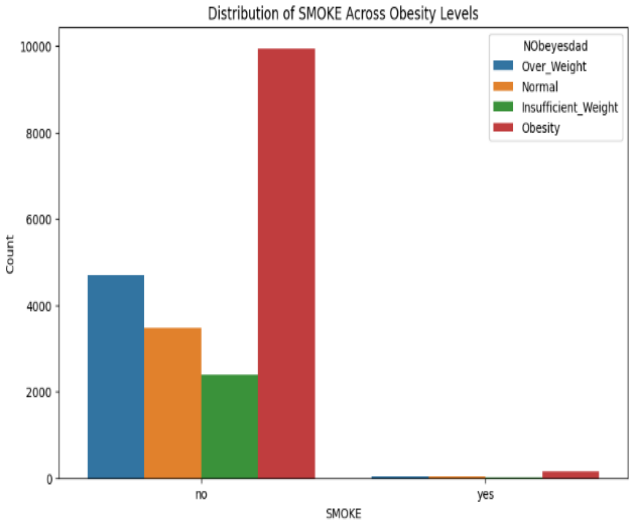
7.4 Count Plots: Behavioral Habits Across Obesity Levels

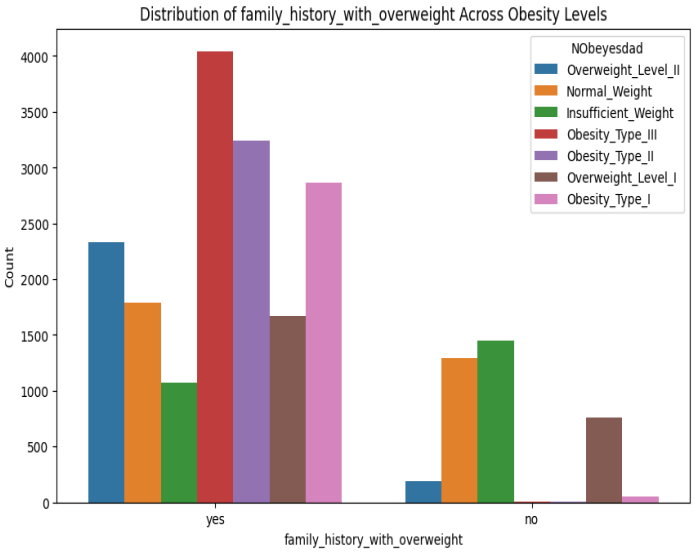
A count plot of 5 obesity levels and 3 behavioral habit categories was generated to visualize the distribution of physical activity, dietary habits, and sleep patterns across varying obesity levels, revealing possible trends.

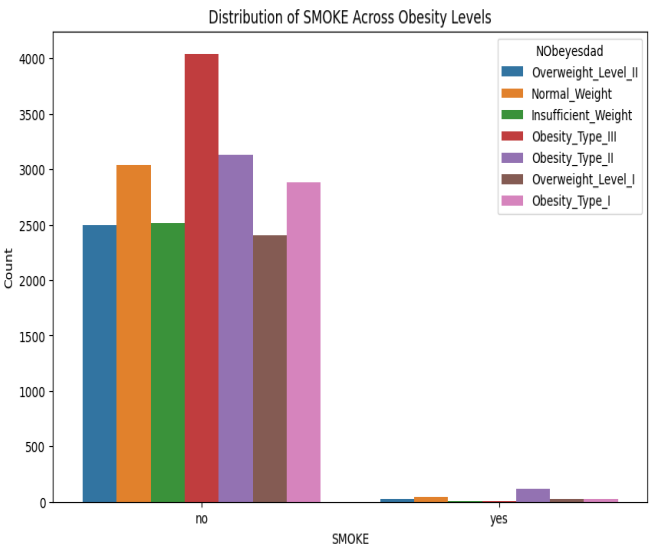


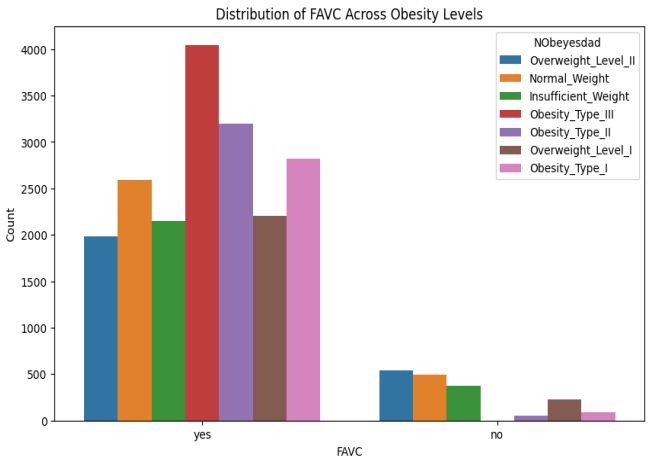


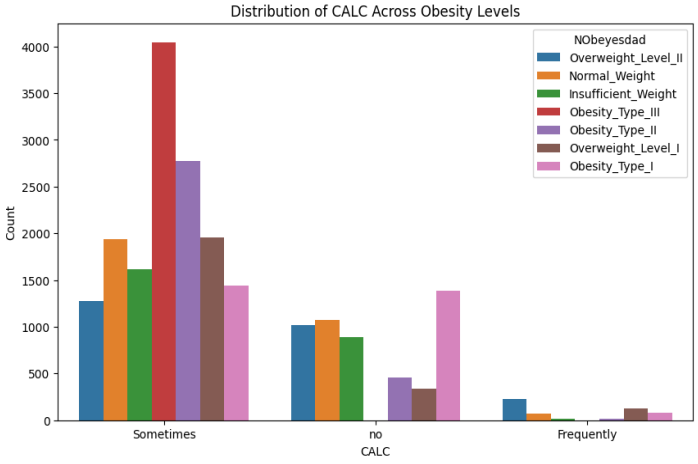


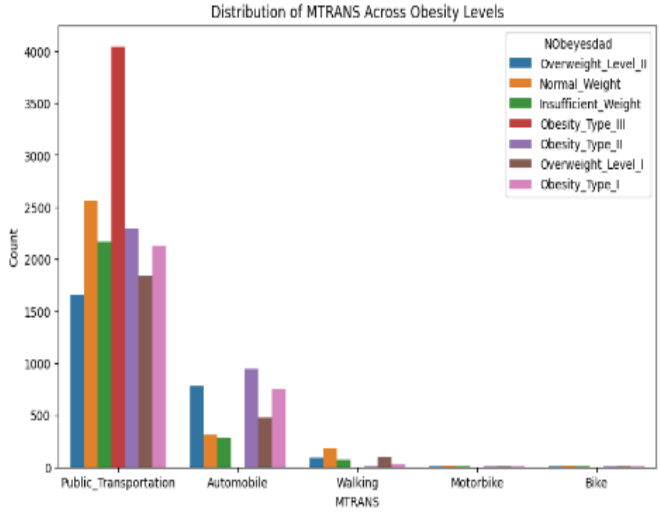


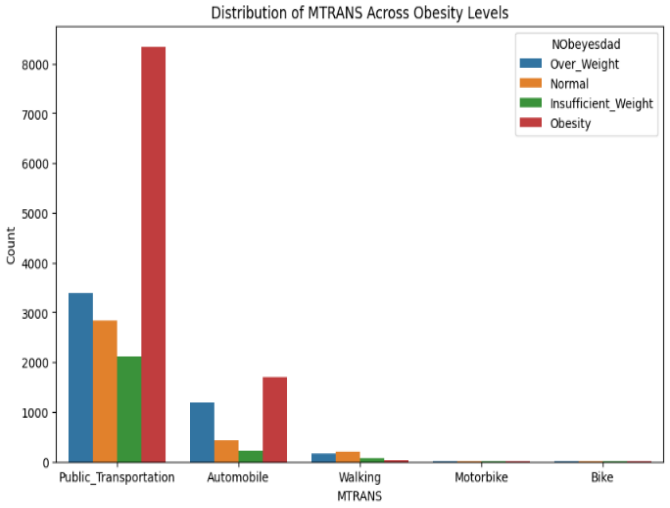








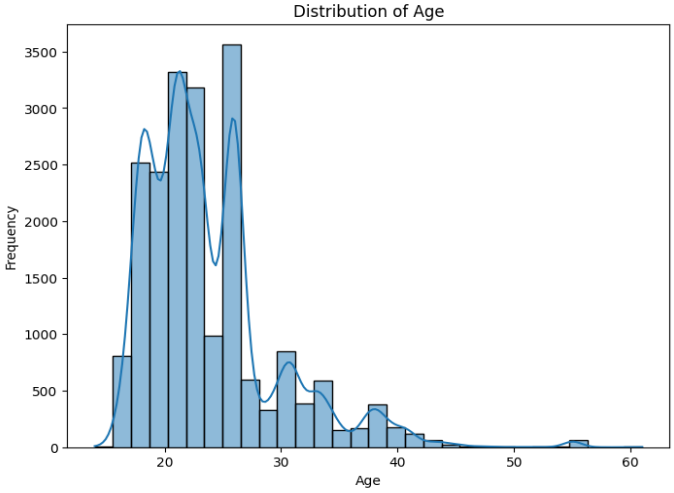


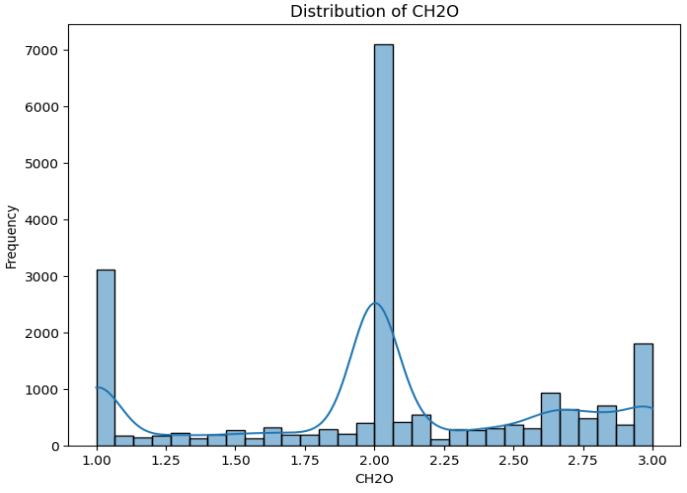


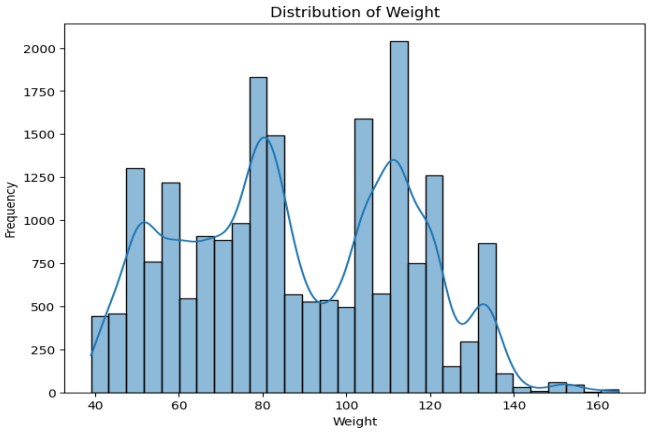
7.5 Univariate Analysis:

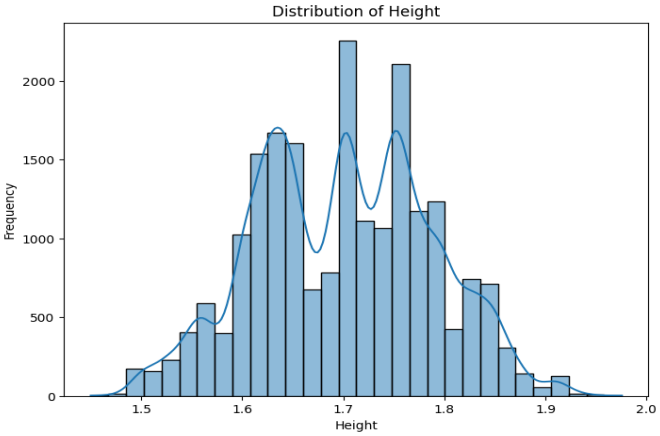
Univariate analysis examines a single variable to summarize its distribution, central tendency, and spread, using tools like histograms and boxplots to identify patterns or anomalies.

Since it takes a single value to summarize and both classes are from same dataset, it gives same graphs.



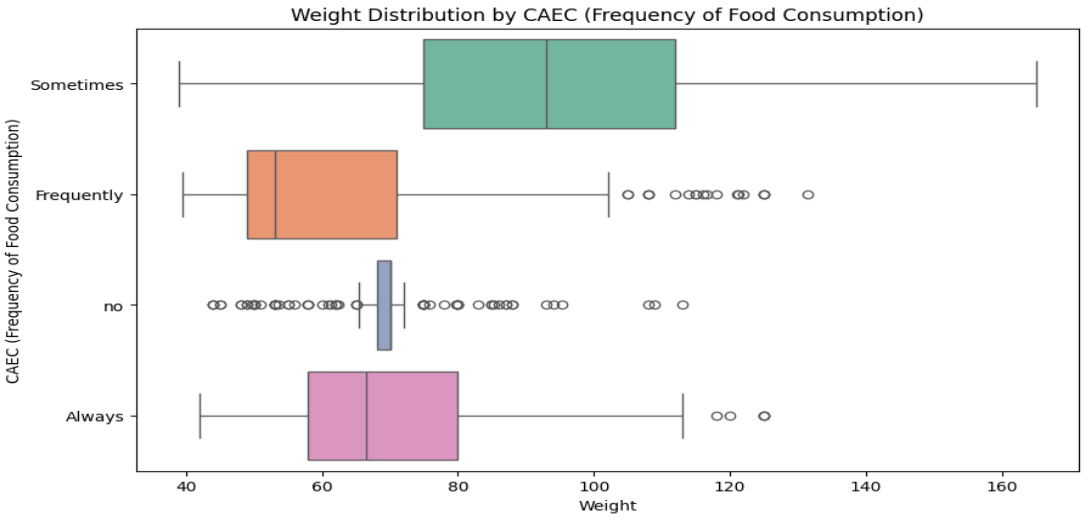


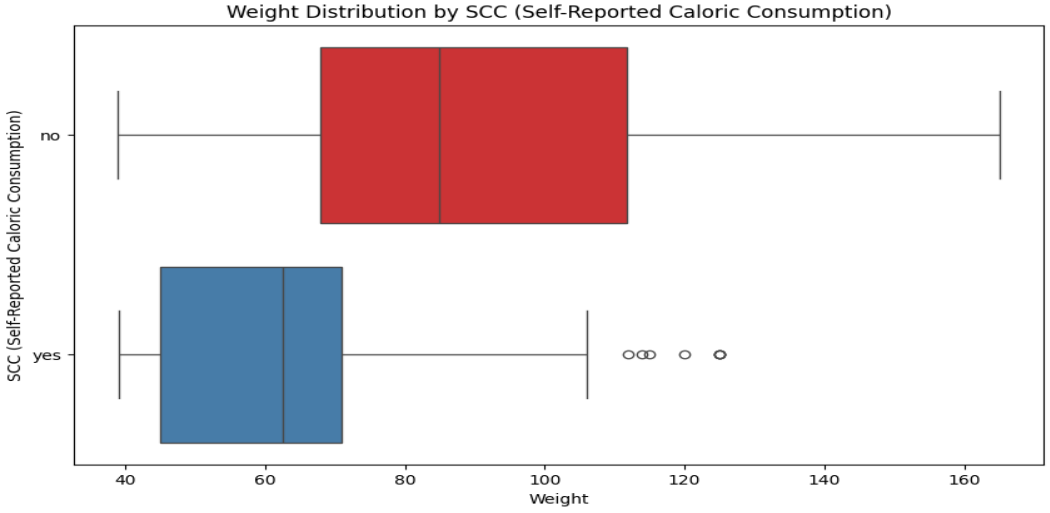




7.6 Bivariate Analysis:

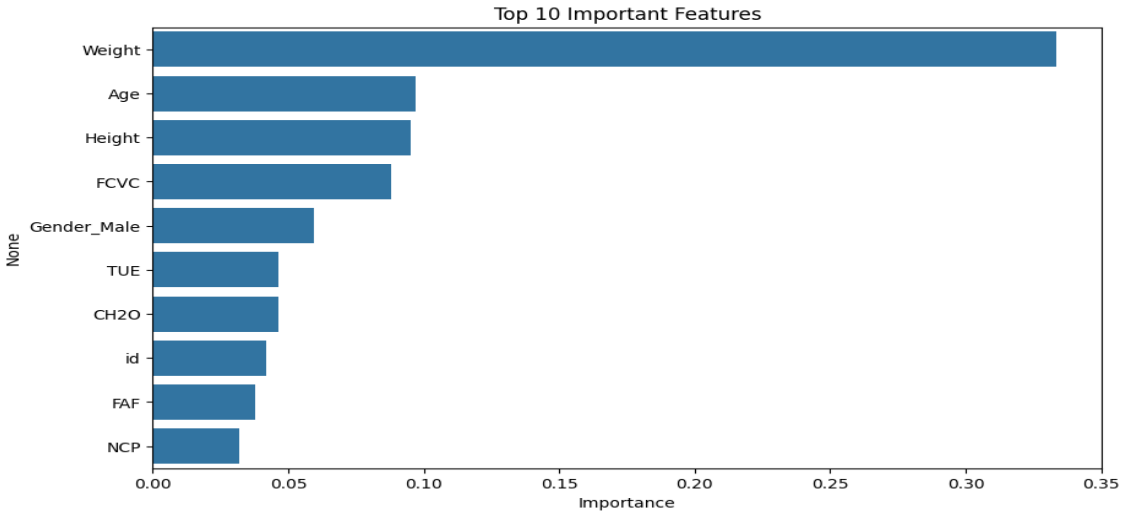
Bivariate analysis examines the relationship between two variables, exploring their associations, correlations, and interactions. Also gives same output for both datas.

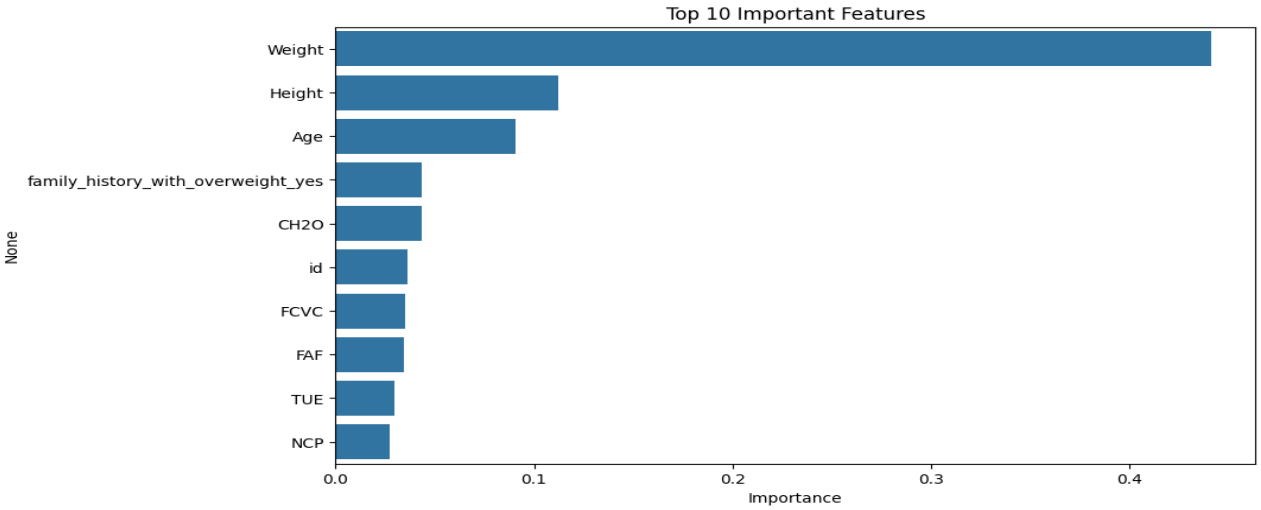




7.7 Important Features:

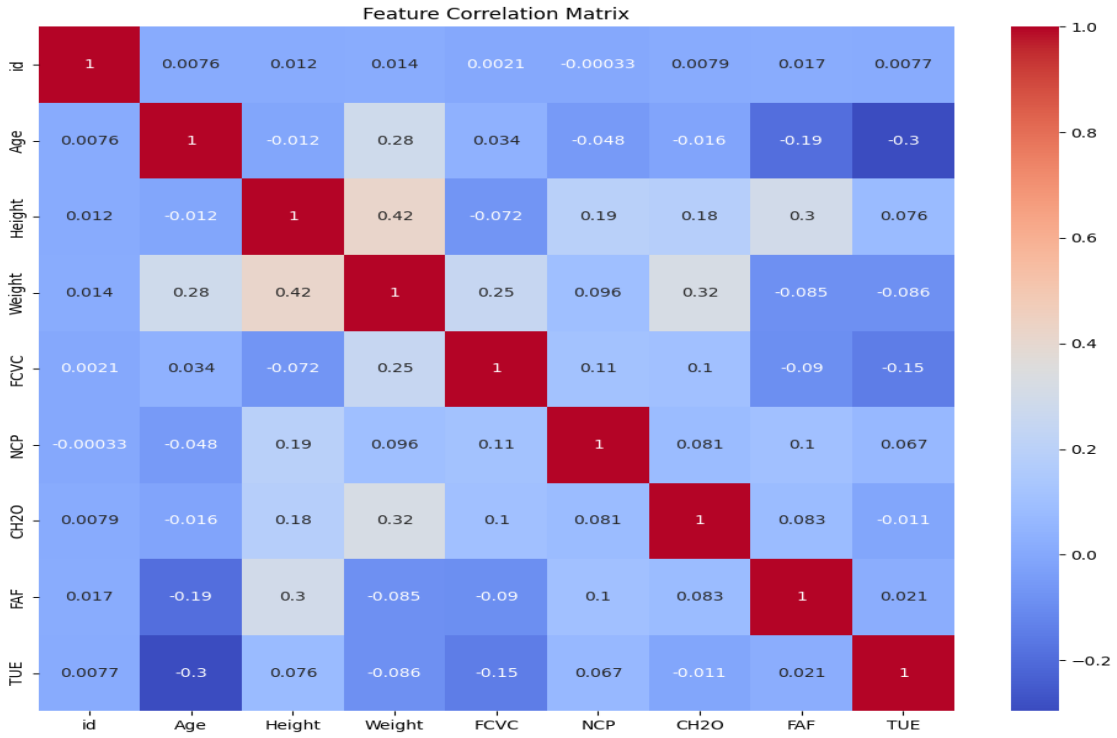
A bar plot of the top 10 important features was generated to highlight their significance in predicting obesity levels, emphasizing key contributors.





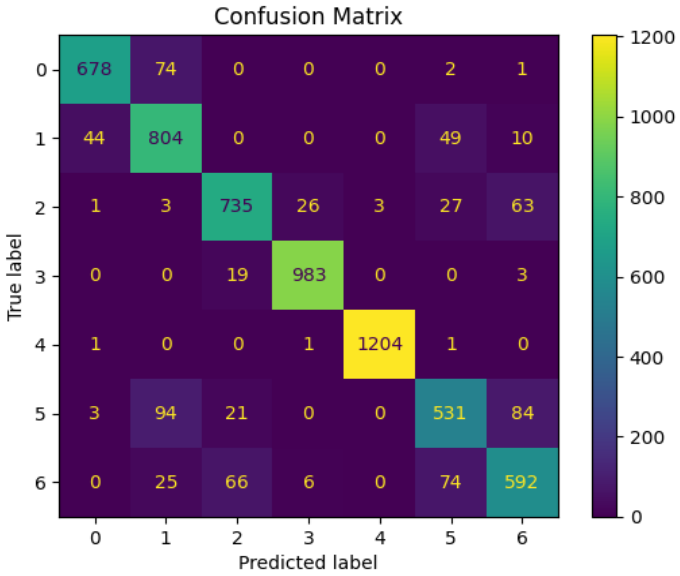
7.8 Correlation Matrix:

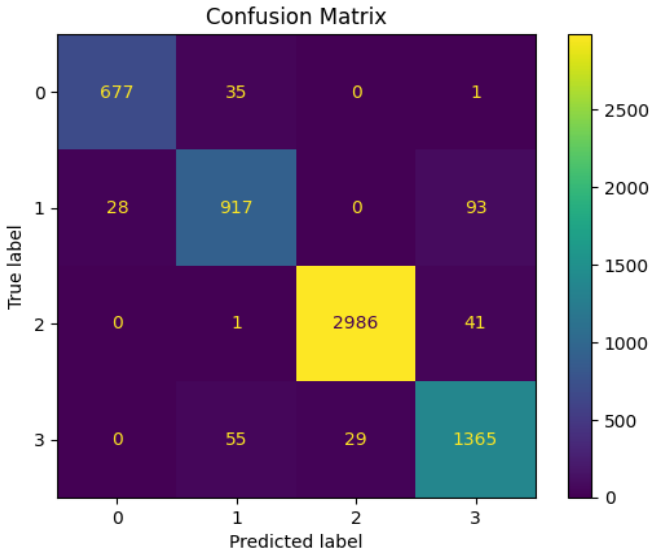
A correlation matrix was generated to visualize the relationships between features, revealing consistent patterns and similarities across both classes of data.



7.9 Confusion Matrix:

Confusion matrices were generated for both the 7-class and 4-class datasets, providing a detailed comparison of predicted versus actual classifications and highlighting performance differences across class distributions.





8.Model Building:

Model	Description	Test Accuracy (7 classes)	Test Accuracy (4 classes)	Additional Notes
Logistic Regression	Multinomial logistic regression model built using processed dataset.	85%	85%	Confusion Matrix, classification report, and accuracy scores analyzed for evaluation.
K-Nearest Neighbors (KNN)	KNN classifier with varying neighbors; best accuracy achieved with 1 neighbor.	84%	82%	Plot showing accuracy based on number of neighbours generated.
Random Forest Classifier	Random forest model trained with max depth of 10.	90.48%	94%	Confusion Matrix, classification report, and accuracy scores analyzed for comparison.
XGBoost	XGBoost model trained with n_estimators of 100.	90.63%	99.81%	XGBoost is an algorithm resulting in improved model accuracy.
LightGBM	LightGBM is a fast, distributed, and high performance gradient boosting framework.	90.54%	95%	LightGBM is designed for efficient and scalable machine learning.
ANN	ANN model trained with less no.of layers.	88.5%	88%	ANNs learn and approximate any complex network.

9. Final Model Selection and Evaluation:

After comparing the performance of the Logistic Regression, KNN, Random Forest classifiers, XGBoost, LightGBM, and ANN models we choosen the XGBoost model for its better accuracy and higher precision in handling multi-class classification.

10. Web Application:

- Developed a web application to predict obesity levels based on user input.
- **User Choice:**
  It is intergrated with both 7 categories prediction and 4 categories predioction where user can choose anyone of them
  then it will be redirected to the respective model.
- **Input:**
  Here user is requested to provide details about their Gender, Height, weight, FAVC, FCVC, NCP, CAEC, CH20, CALC, SSC, FAF, TUE, MTRANS.
- **Prediction:**
  And after this based on user choice Obesity level is predicted.
  - If user choice is 7 classes then it will predict the obesity level in this Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III.
  - If user choice is 4 classes then it will predict the obesity level in this Insufficient_Weight, Normal Weight, Over_Weight, Obesity.
- **System Architecture:**
  The web applications consists of a frontend built using HTML5, CSS, and a backend developed using Flask. The XGBoost model is integrated into the backend, leveraging its scalability and performance.
- The application offers an user-friendly interface for entering the input data, with the predicted obesity level shown after the processing is completed.

🖇 **Application Demo Link**

11. Obstacles:

1. **Data Collection**: Issues in finding the large dataset and lack of suitable features in the dataset.
2. **Data Preprocessing**: Data entry errors undermine data reliability and trustworthiness. Noisy data affected the model performance.

3. **Model building**: Imbalanced datasets lead to baised models, Insufficient data quality compromises model accuracy.

12. Conclusion:

The Obesity Level Prediction project successfully demonstrates the potential of machine learning in healthcare, specifically in predicting obesity levels. By leveraging the XGBoost algorithm, this web application achieves 90.63%(7 categories) and 99.81%(4 categories) accuracy, outperforming traditional statistical models. This innovative solution provides a user-friendly interface for individuals to assess their obesity risk.

13. Future Work:

- Early intervention and prevention of obesity-related diseases.
- Personalized healthcare and wellness strategies.
- Will collaborate with healthcare professionals for clinical validation.
- Integration of machine learning in healthcare decision-making.
- Explore other machine learning algorithms for comparison.

14. References:

- Obesity Prediction Dataset
- Obesity Data Set
- Obesity Risk EDA & Prediction Dataset
- Obesity Classification with Extra Trees (100% Accuracy)