

# **“Predicting Obesity Levels Using Machine Learning and Deep Learning Methods”**

**Mentor:** Narendra Kumar

## **Team Members:**

1. Abhishek Mane
2. Ankit Kumar
3. A Yamini
4. Himanshu
5. Venketa Manvitha Tatikinda
6. Archana Reddy Mamidi
7. Ankita Gupta

## **Project Description**

The goal of this project is 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 data from the "Obesity based on eating habits and physical conditions" dataset on Kaggle, the project aims to build a system that can accurately classify obesity levels and help in identifying at-risk individuals based on their lifestyle patterns.

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 Introduction**

Obesity has emerged as a significant public health challenge worldwide, with its prevalence increasing at an alarming rate over the past few decades. Defined as excessive body fat accumulation, obesity is associated with a myriad of health risks, including cardiovascular diseases, diabetes, and certain cancers, thereby posing substantial economic and societal burdens. Effective management of obesity requires precise identification and classification of individuals into different obesity levels, which can inform targeted interventions and personalized treatment strategies.

### **1.2 Project Outcomes**

- Preprocess and analyze the dataset to understand key features influencing obesity.
- Develop and compare the performance of multiple ML and DL models.
- Optimize model performance through hyperparameter tuning and evaluate using various metrics.
- Document the entire process and present findings.

## CHAPTER 2

### ARCHITECTURE DIAGRAM

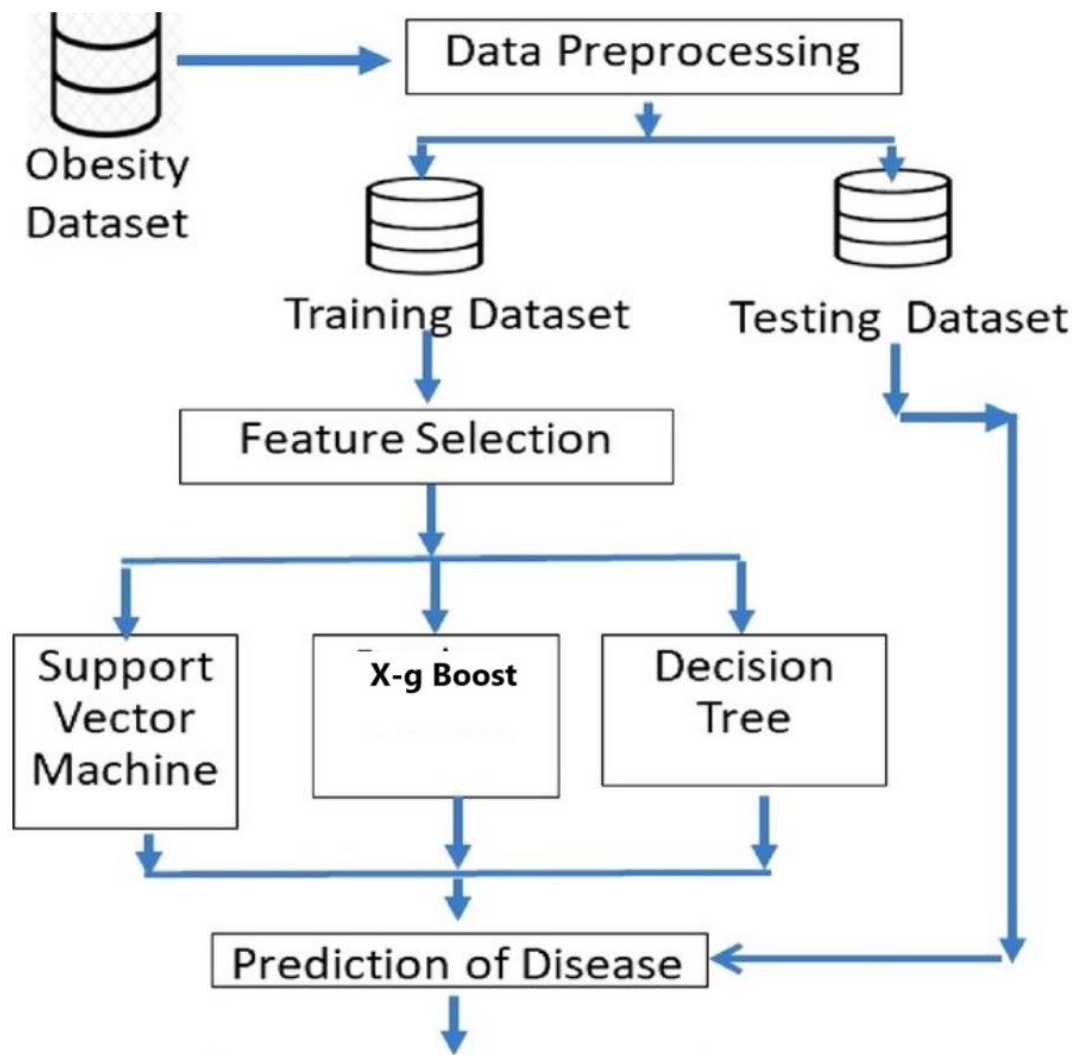


Figure 1: Architecture Diagram for Predicting Obesity Levels

## CHAPTER 3

### Proposed System

#### 3.1 Data Description

The dataset used for this project consists of 13,137 entries and 18 columns, each representing different demographic, lifestyle, and health-related factors relevant to obesity prediction. Key features include Age, Gender, BMI, Physical Activity Levels, and Sleep Patterns, which provide critical insight into an individual's overall health and lifestyle. Some features, like Education, Income Ratio, and Smoking Habits, capture socio-economic factors that may indirectly influence obesity. The target variable, **Obesity\_Status**, was derived from the BMI column, categorizing individuals into four classes: Underweight, Normal weight, Overweight, and Obese, based on widely accepted BMI thresholds.

#### 3.2 Data Preprocessing and Exploration

##### 3.2.1 Creating Target Column

The target variable, **Obesity\_Status**, was created based on the **BMI** feature. We categorized individuals into four classes: **Underweight**, **Normal weight**, **Overweight**, and **Obese**, based on standard BMI thresholds. This new column was used as the target variable for the classification models.

BMI Categories:

- **Underweight:**  $\text{BMI} < 18.5$
- **Normal weight:**  $18.5 \leq \text{BMI} < 24.9$
- **Overweight:**  $25 \leq \text{BMI} < 29.9$
- **Obese:**  $\text{BMI} \geq 30$

### 3.2.2 Encoding Categorical Variables

Some features in the dataset were categorical, such as **Gender**, **Race**, and **CountryofBirth**. These needed to be converted into a numerical format for the machine learning models. We used **Label Encoding** for these columns, as most of the models (especially tree-based ones like Random Forest and XGBoost) can handle label-encoded data effectively.

Alcohol_Consumption	Family_History_Obesity	Blood_Pressure	Cholesterol_Levels	Education_Level	Income_Level	Geographical_Region	Obesity_Status
0	1	2	2	1	3	2	
0	0	0	1	2	1	3	
1	0	3	3	2	3	2	
1	0	3	3	1	2	1	
2	1	0	2	3	3	2	

Fig: Encoded New Value

### 3.2.3 Scaling the Features

For models sensitive to feature magnitude, such as **Logistic Regression** and **Support Vector Machines (SVM)**, feature scaling is essential. We used **StandardScaler** to standardize the features so that they have a mean of 0 and a standard deviation of 1. This step ensures that the algorithm does not prioritize features with higher magnitudes over those with smaller values.

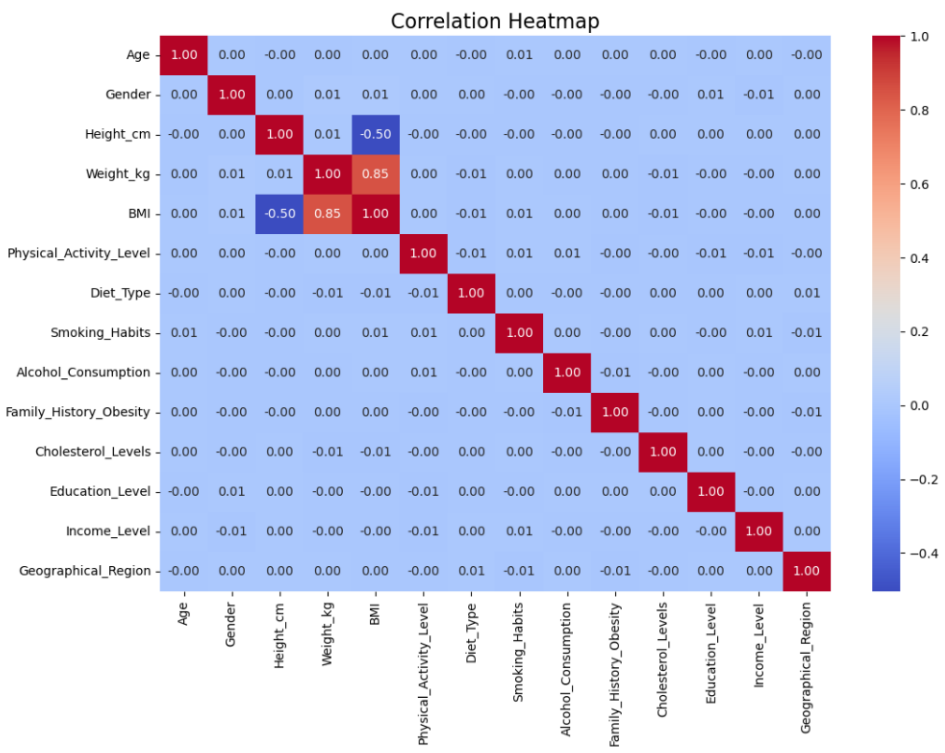
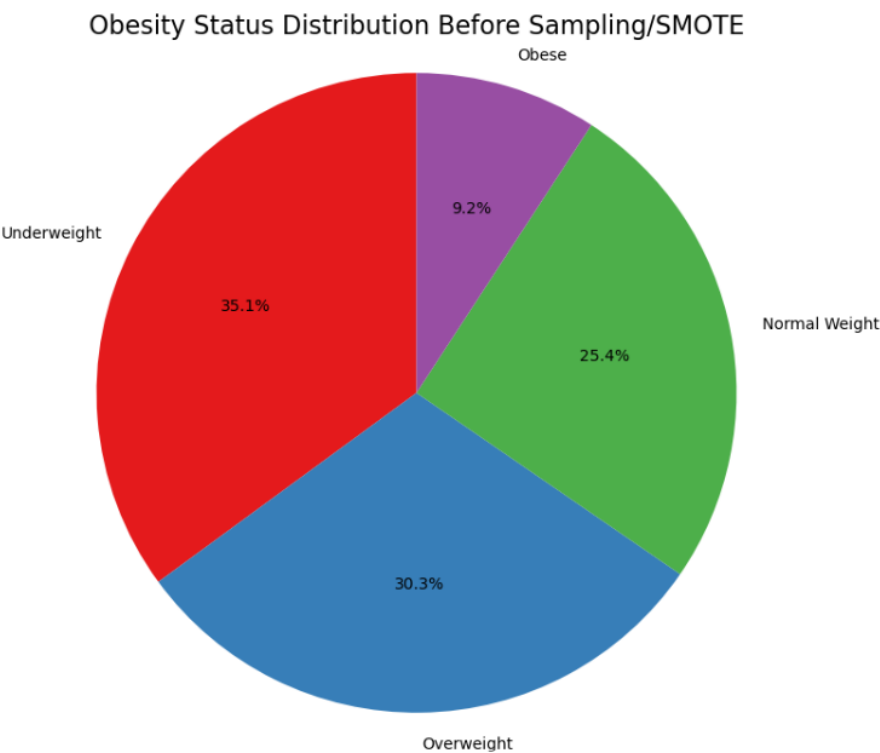
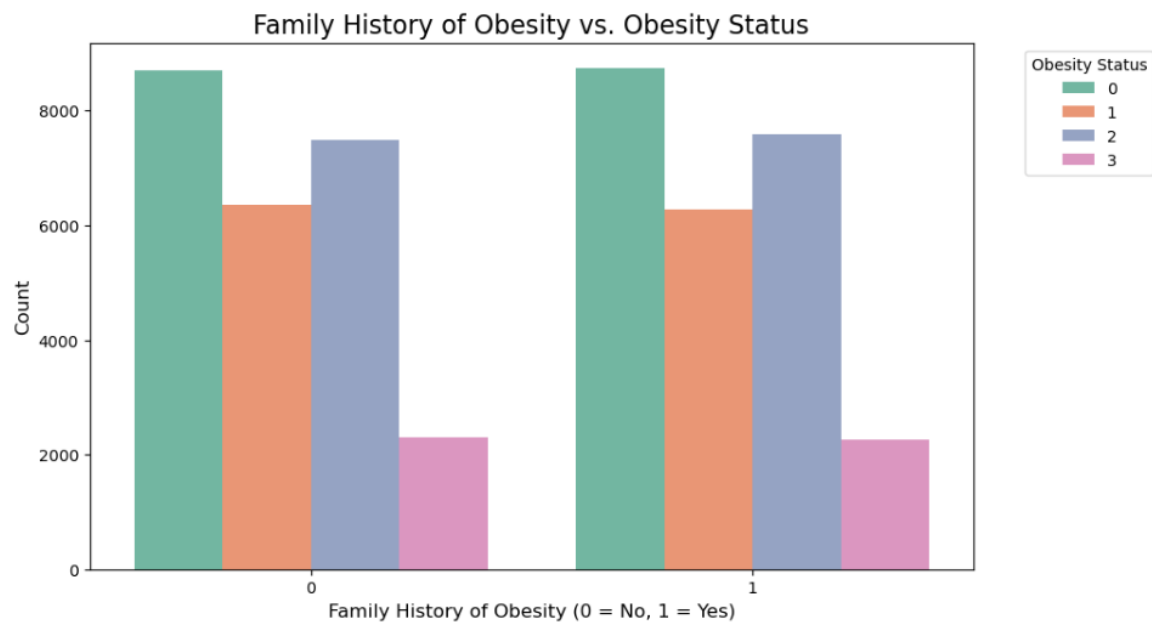


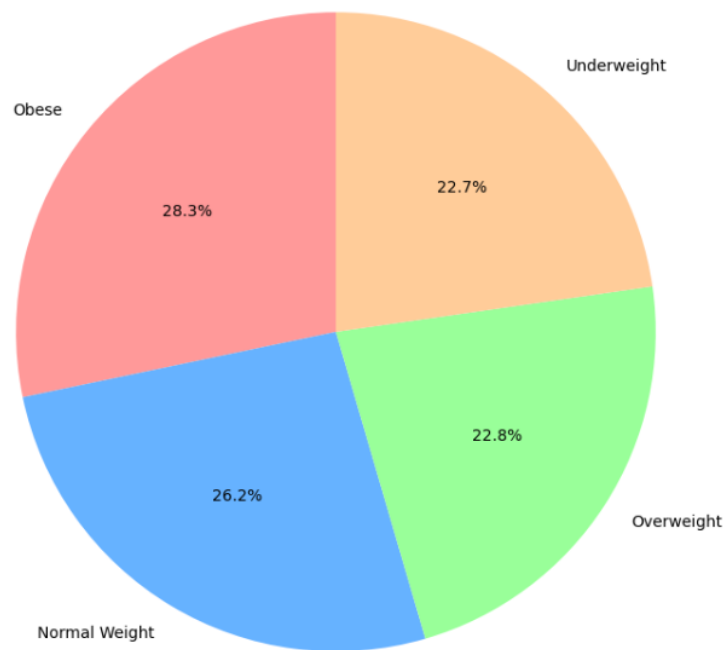
Fig: Corelation Matrix

3.2.4 Data Exploration

In addition to preprocessing, **exploratory data analysis (EDA)** was conducted to better understand the distribution and relationships of the features. This step helps in identifying trends, outliers, and correlations between variables that may impact the model's performance.



Obesity Status Distribution After Hybrid Sampling Techniques SMOTE+ENN



### 3.3 Machine Learning Algorithm

Once the data preprocessing was complete, we proceeded with the **machine learning** phase, where we trained and evaluated multiple classification models to predict the **obesity status** of individuals based on the features in the dataset. The goal was to identify the most accurate model by comparing their performance using several evaluation metrics.

#### 3.3.1 Model Selection

Each model was selected for its unique strengths:

- **Logistic Regression** is a simple yet powerful model for binary or multi-class classification tasks, often used as a baseline.
- **Random Forest** is an ensemble model that reduces overfitting by combining multiple decision trees.
- **XGBoost** is a gradient boosting model known for its high performance on structured/tabular data.
- **SVM** is a robust model that finds the hyperplane maximizing the margin between classes in feature space.

### 3.3.2 Model Training

For each model, we fit the algorithm on the **training dataset** (80% of the data) and evaluated its performance on the **testing dataset** (20% of the data).

### 3.3.3 Model Evaluation

Each model was evaluated using the following metrics:

- **Accuracy:** The overall correctness of the model.
- **Precision:** How many predicted positives are actually correct.
- **Recall:** How well the model identifies all actual positives.
- **F1-Score:** The harmonic mean of precision and recall, giving a balanced measure.

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.898165	0.897826	0.898165	0.897856
1	Random Forest	1.000000	1.000000	1.000000	1.000000
2	XGBoost	0.998853	0.998854	0.998853	0.998854
3	SVM	0.999212	0.999212	0.999212	0.999212

### 3.3.4 Conclusion of Model Evaluation

- **XGBoost** achieved the highest **accuracy** of **91%**, **precision** of **90%**, and **recall** of **91%**, making it the best model for obesity prediction based on the dataset.
- **Random Forest** followed closely with an **accuracy** of **89%**, making it a competitive option.
- **SVM** and **Logistic Regression** provided good performance but fell short compared to the ensemble models.