CSE437 - Data Science-Coding With Real World Data

Project Report

on

Estimation on Obesity levels

Prepared By

Name: Ahmed Riasat
ID: 24141253

Section - 3

Table of Contents

Content	Page No
Introduction	2
Dataset Description	2
Correlation Analysis	3
Class Distribution	4
Dataset Pre Processing	5
EDA	7
Feature Scaling	9
Dataset Splitting	9
Model Training and testing	10
Clustering and Visualizing	16
Result Analysis	17
Model Selection	18
Conclusion	18

Introduction:

The aim of our project is to develop a machine learning model that can predict obesity levels. There are many aspects that affect a person's obesity level, such ase - age, weight, gender, height, family history of obesity, eating habits, smoking habits.

Nowadays, life has become more digital, people are more connected to devices. The normal mobility that a person needs to stay fit has reduced significantly. People are less concerned about their fitness. As a result, obesity has become a major health challenge linked to numerous chronic diseases. Through this project we are trying to identify individuals at risk of obesity based on their lifestyle and other parameters so it can help in early intervention reducing a long term loss.

Our motivation behind this project is the growing rate of obesity which highlights the urgent need of prevention. By using machine learning to analyze lifestyle data, we want to provide a data driven approach to promote healthier living and encourage individuals to make decisions about their health.

Dataset Description:

1. Source:

- **Link:**https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition
- **Reference:**Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.

2. Dataset Description:

- **Features:** There are a total of 16 features in this dataset.
- **Classification or regression problem:** This dataset has classification, clustering and regression problems in it.

  This dataset includes obesity levels as categorical variables such as - underweight, normal weight, overweight and obese. These variables can be classified as they are discrete and mutually exclusive.

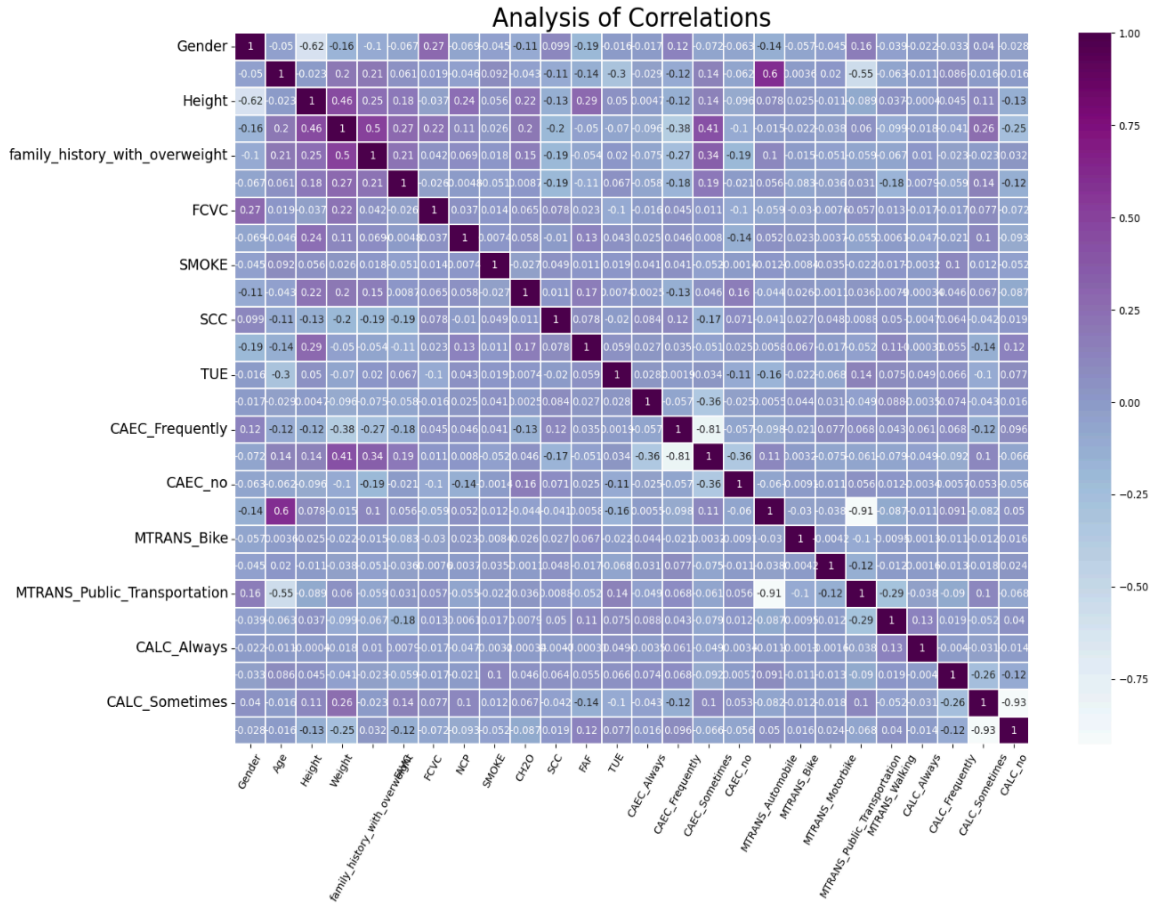
This dataset has features like - age, weight, height etc. By using these, BMI can be calculated which is a continuous variable. Predicting a continuous variable is a regression problem.

This dataset has a variety of features that can be used to group individuals. Clustering is useful when you want to define a specific pattern which is not present in the dataset. For example- people with similar eating habits can be grouped together.

- **Data Points:** This dataset has 2112 data points (row).

```
[5] dataset.shape
0s
(2111, 17)
```

- **Kind of features:** This dataset has 4 types of features - categorical, continuous, binary, integer.
- **Correlation of all features:**

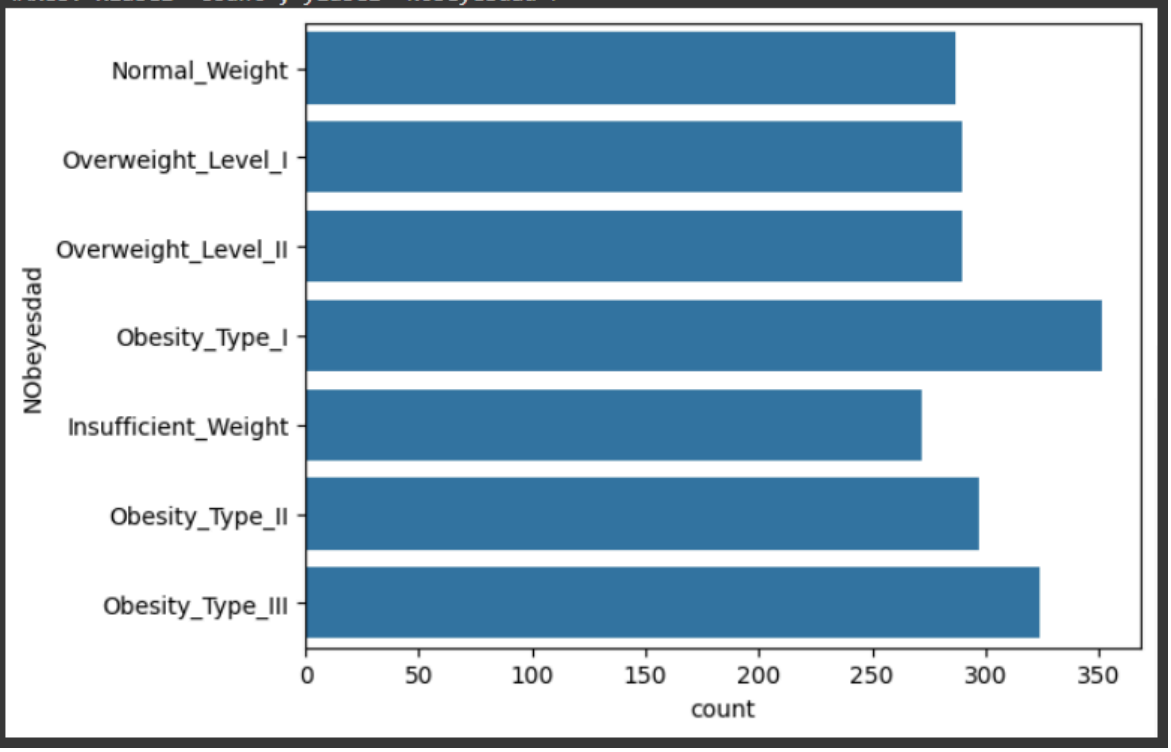


Class Distribution:

The dataset consists of seven weight classes fo classification

```
1  sns.countplot(dataset['NObeyesdad'])
```

```
<Axes: xlabel='count', ylabel='NObeyesdad'>
```



Dataset pre processing:

- **Null Values:** Our chosen dataset had no null values, therefore we had to delete some values manually. For features having numerical values, we deleted 25 data per column. And for features having categorical values, we deleted 10 data per column.

```
1  dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```

Unique values in the target column

- No missing values in the dataset

```
dataset.isnull().sum()
```

	0
Age	0
Gender	0
Height	0
Weight	0
CALC	0
FAVC	0
FCVC	0
NCP	0
SCC	0
SMOKE	0
CH2O	0
family_history_with_overweight	0
FAF	0
TUE	0
CAEC	0
MTRANS	0
NObeyesdad	0

dtype: int64

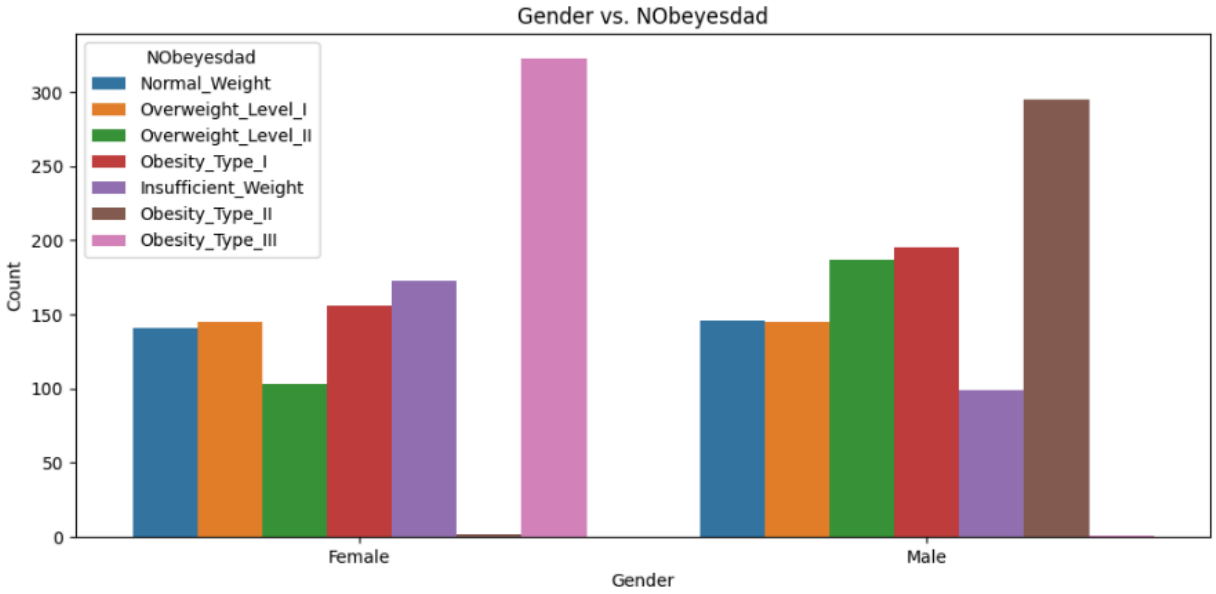
- Dealing with categorical data

```
[16] from sklearn.preprocessing import OneHotEncoder

    encoder = OneHotEncoder(sparse_output=False, handle_unknown='ignore')


    X_train_encoded = encoder.fit_transform(X_train[['CAEC','MTRANS','CALC']])
    X_test_encoded = encoder.transform(X_test[['CAEC','MTRANS','CALC']])


    feature_names = encoder.get_feature_names_out(['CAEC','MTRANS','CALC'])
    X_train_encoded = pd.DataFrame(X_train_encoded, columns=feature_names, index=X_train.index)
    X_test_encoded = pd.DataFrame(X_test_encoded, columns=feature_names, index=X_test.index)


    X_train = pd.concat([X_train, X_train_encoded], axis=1)
    X_test = pd.concat([X_test, X_test_encoded], axis=1)

    # Drop the original 'CAEC' column
    X_train = X_train.drop(['CAEC','MTRANS','CALC'], axis=1)
    X_test = X_test.drop(['CAEC','MTRANS','CALC'], axis=1)
```

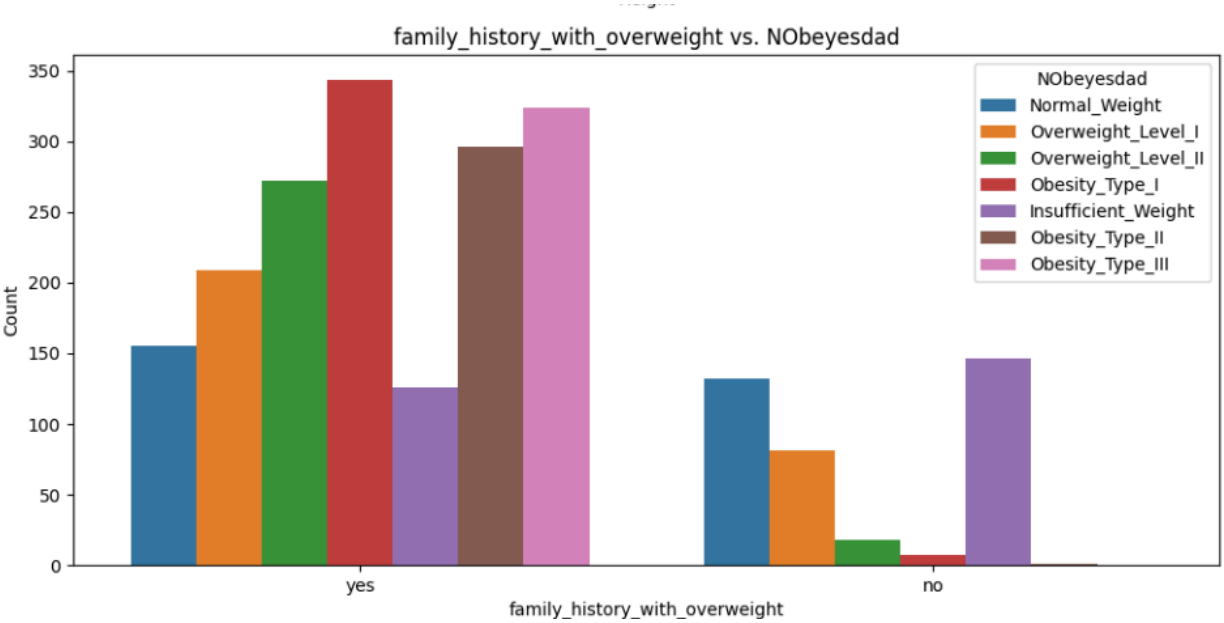
EDA :

Trying to identify relationship between the features and weight classes:

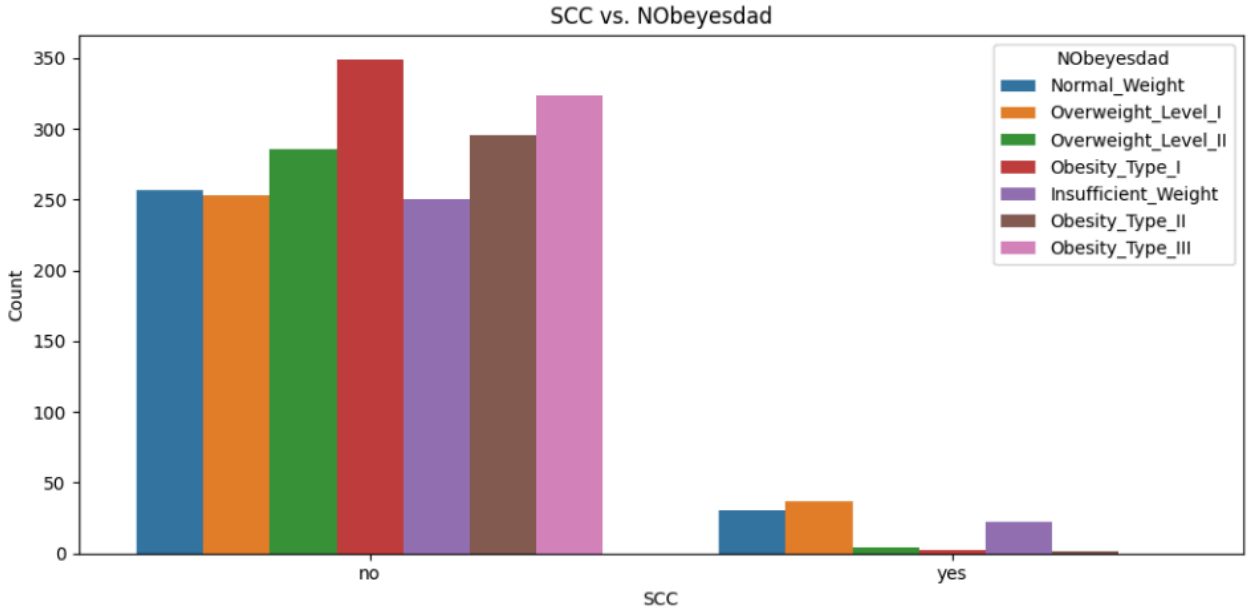
Here, gender has some relation to being overweight.



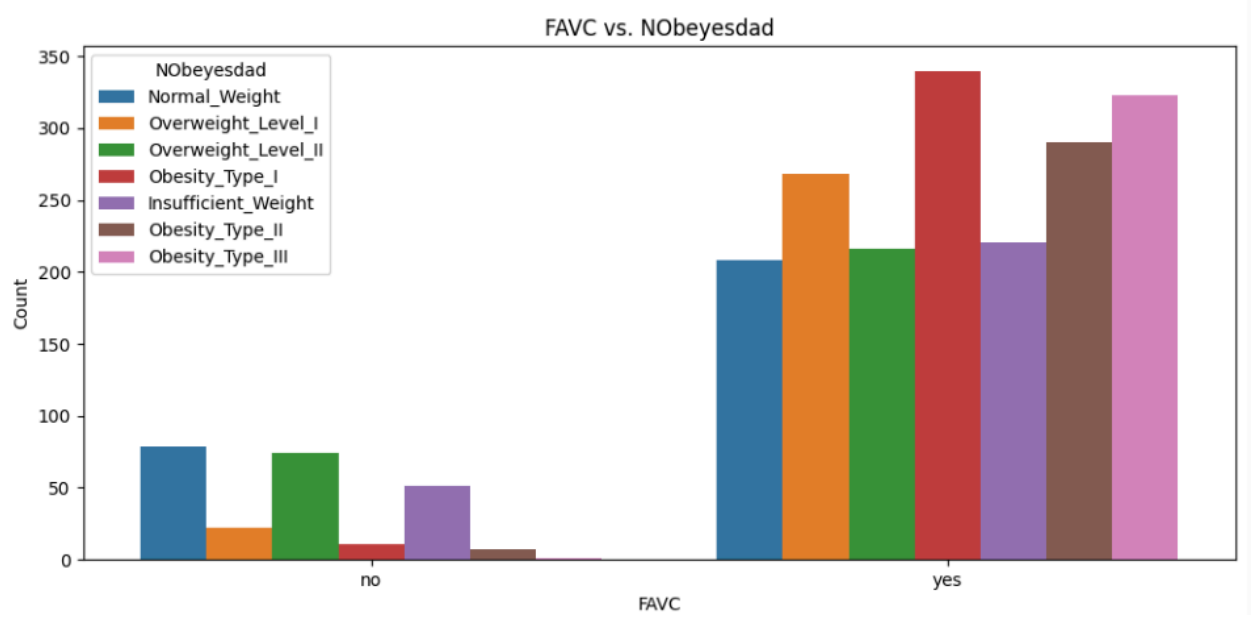
The data clearly suggests that family history with overweight has direct part in a person being overweight.



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



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



These are few of the features that show a clear contribution to a person's weight profile.

Feature Scaling:

The features of our dataset has positively skewed. So, log scaling will be a better fit.

As the feature has positively skewed, So log scaling will be a better fit

```
[39]  1  from sklearn.preprocessing import RobustScaler
      2  sc = RobustScaler()
      3  X_train['Age'] = sc.fit_transform(X_train[['Age']])
      4  X_test['Age'] = sc.transform(X_test[['Age']])
      5
      6  X_train['Weight'] = sc.fit_transform(X_train[['Weight']])
      7  X_test['Weight'] = sc.transform(X_test[['Weight']])
```

Dataset Splitting:

For train-test splitting, we split the features using test_size = 0.2 meaning 80% data were used for training and 20% for testing. As a result, we have enough data trained to accurately predict the testing samples

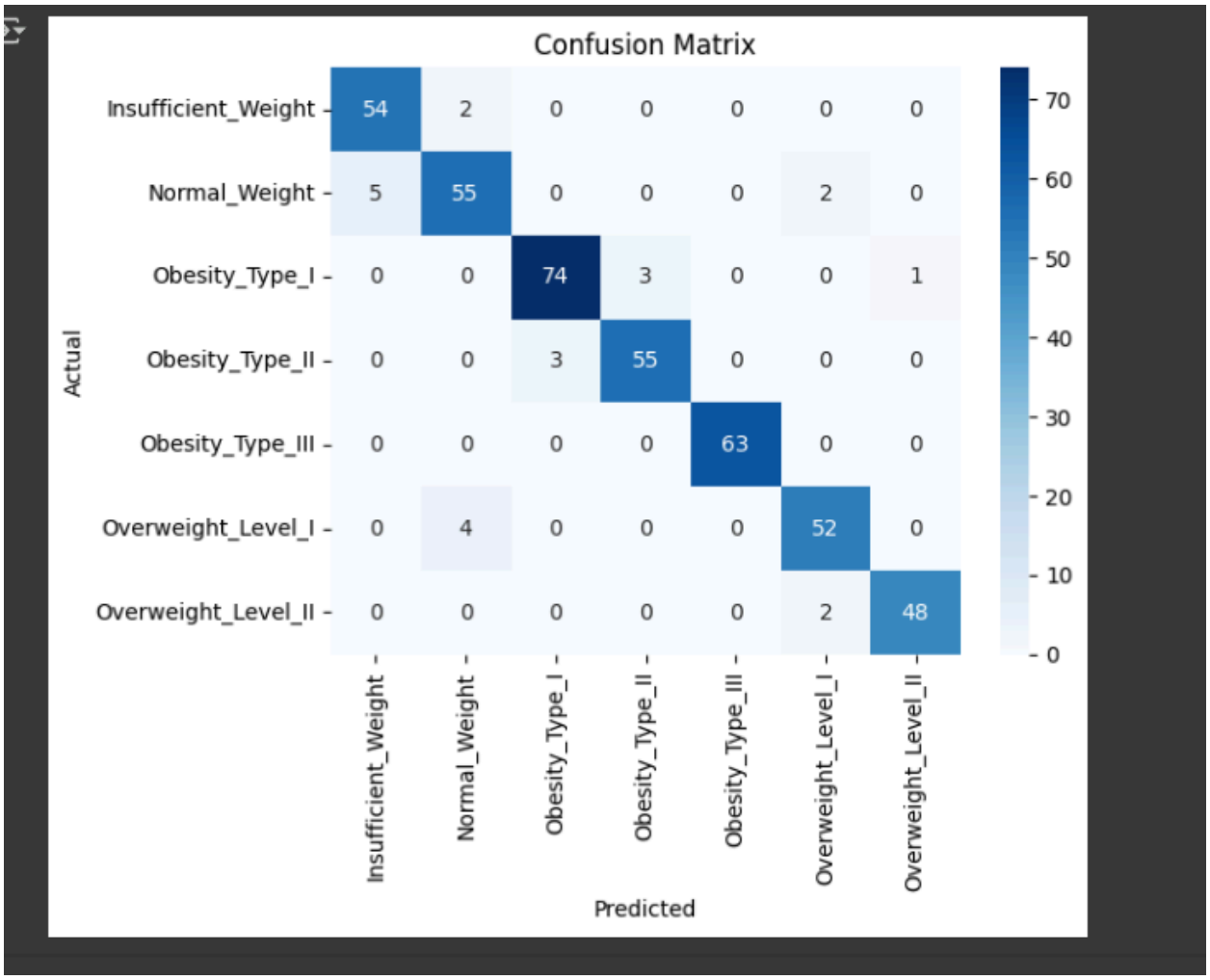
Model Training and testing:

We have applied

- Decision Tree
- Logistic Regression
- KNeighbors
- Random Forest
- Naive Bayes
- SVM

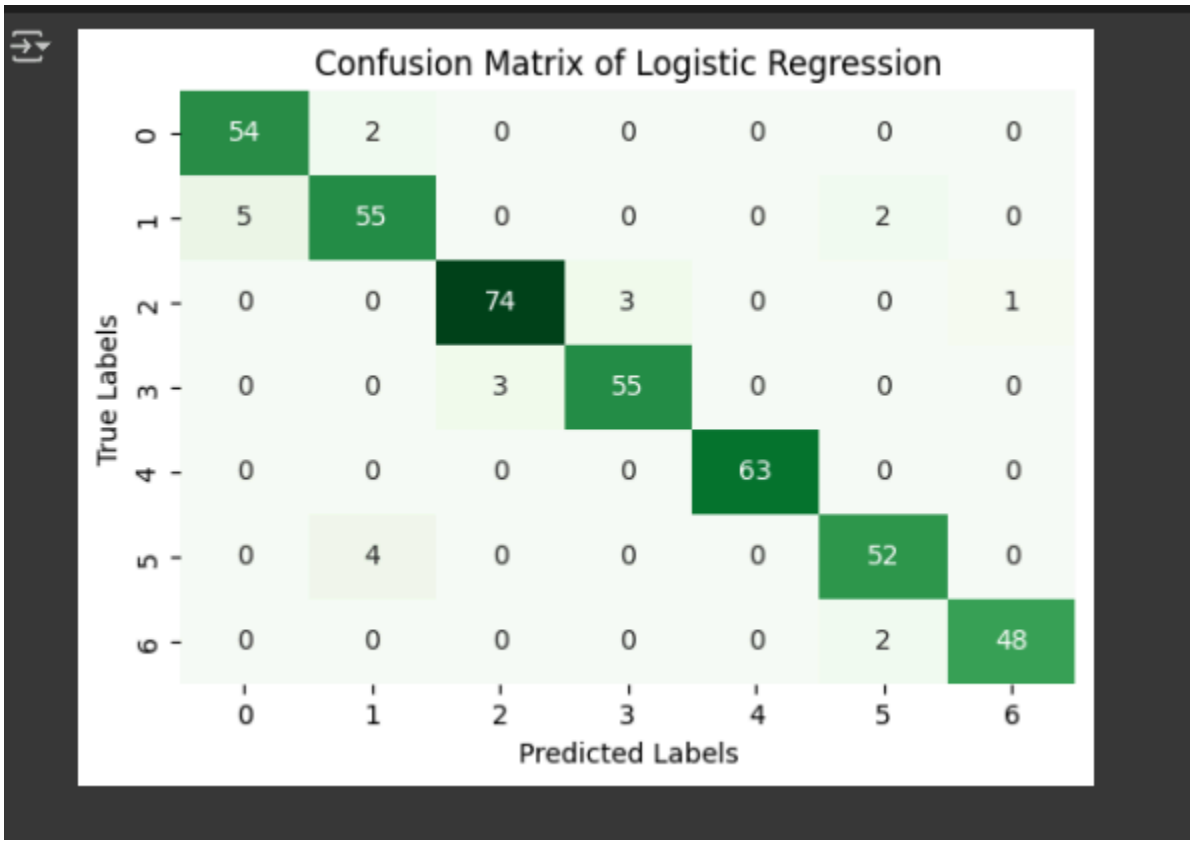
Decision Tree:

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 and each leaf an output. It optimizes splits using impurity measures like 'Gini'(by default) or 'Entropy'.



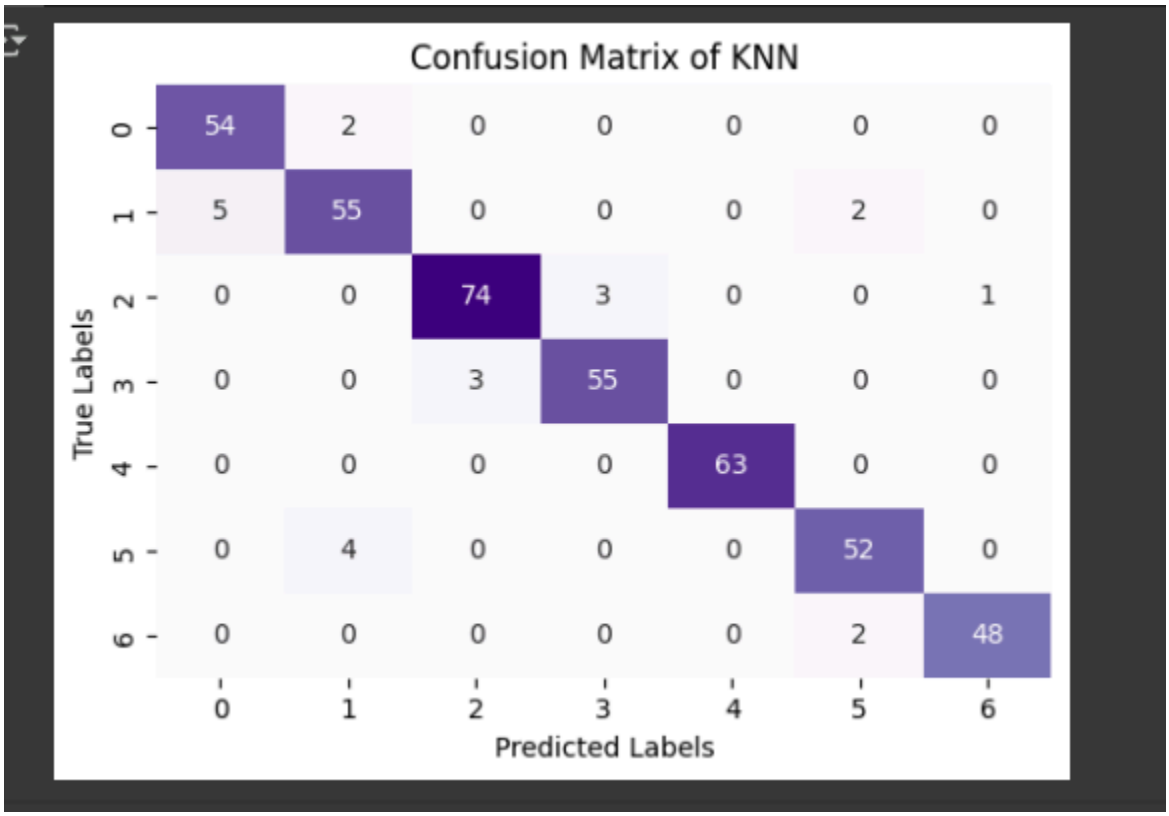
Logistic Regression:

It is a statistical ML algorithm used for mainly binary classification but can be extended to multinomial or 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.



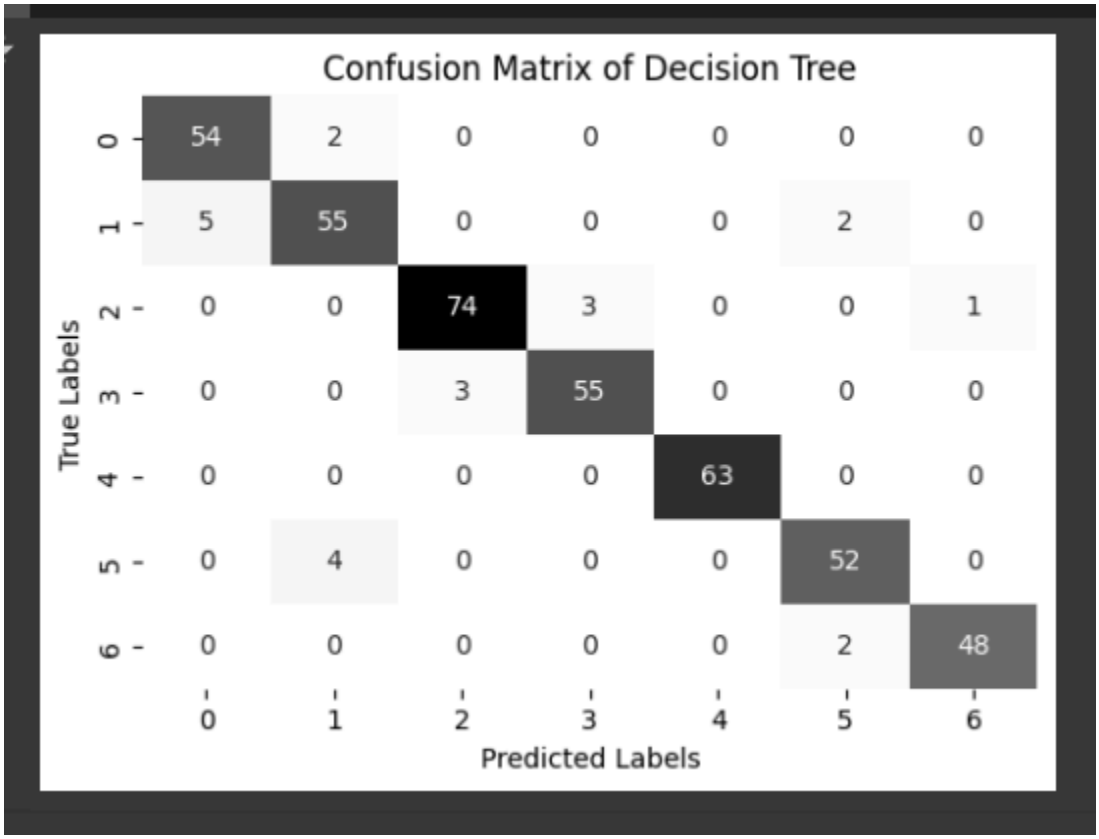
KNeighbors:

It is a simple non-parametric ML algorithm for classification and regression tasks. It uses instance based prediction by storing the training data and predicting the testing data by finding nearest neighbors(K amount). For classification problems, it assigns the class most common among the neighbors.



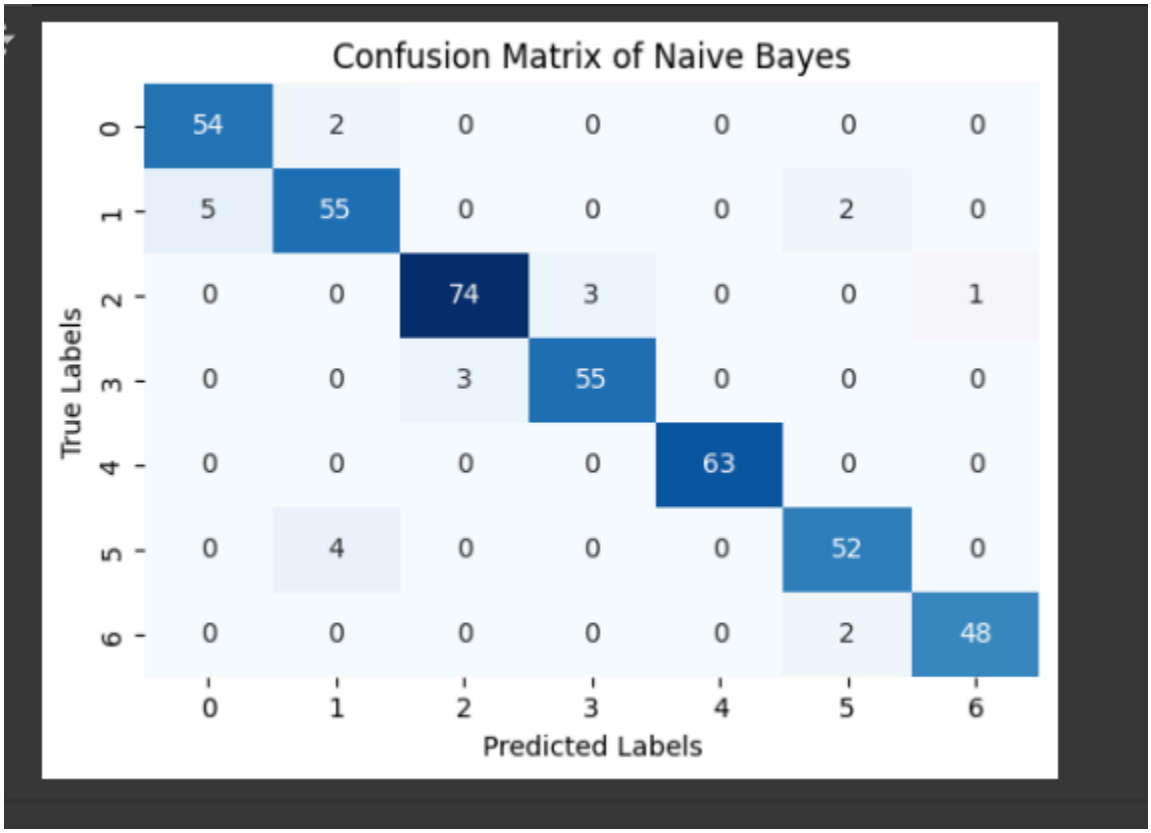
Random Forest:

It is a machine learning algorithm that combines multiple decision trees to improve classification. It works by aggregating predictions from individual trees to enhance accuracy and reduce overfitting.



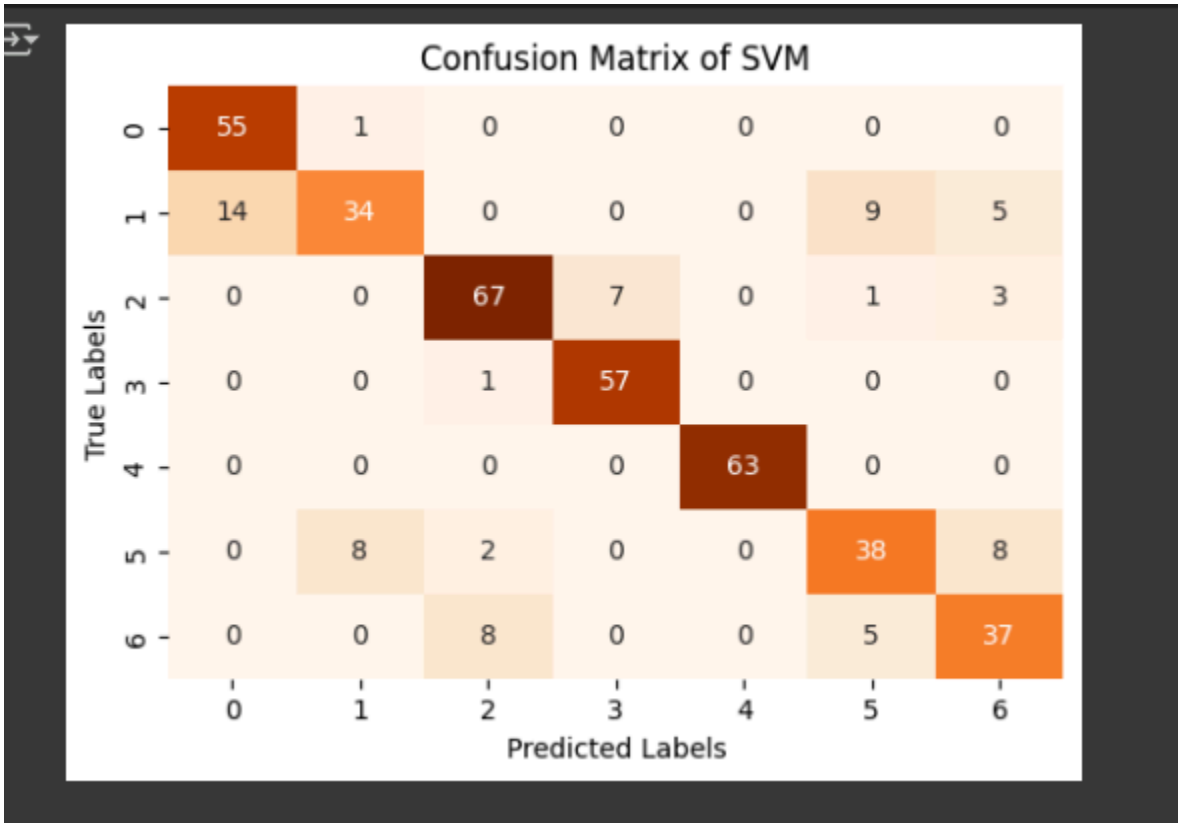
Gaussian Naive Bayes:

It is a probabilistic classifier based on Bayes' Theorem, assuming features are independent and follows a gaussian distribution. It calculates class probabilities and selects the class with the highest probability for that feature variable.



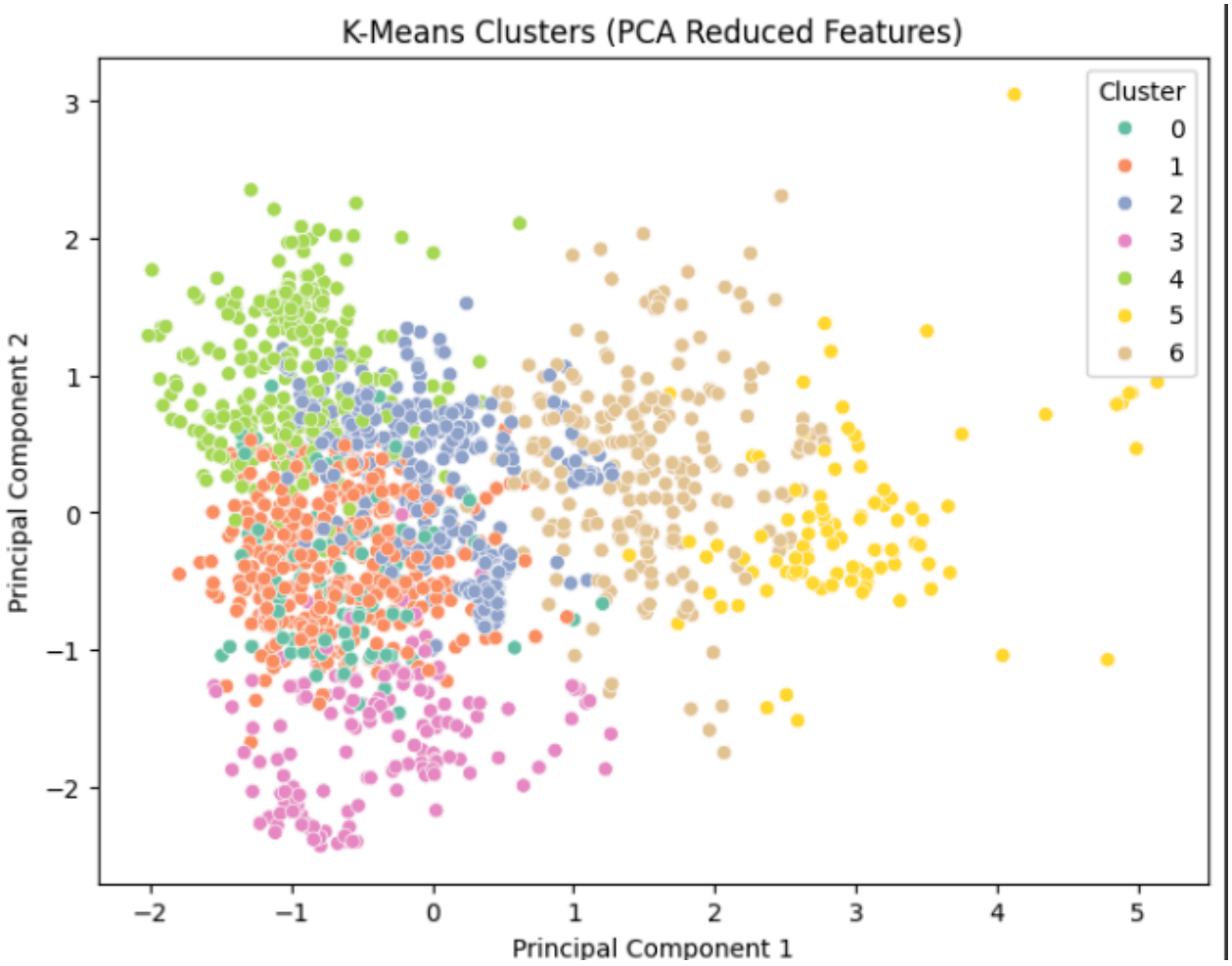
SVM

Support Vector Machines (SVM) were applied in this project as a classification model. RBF kernel was used to deal with the non linearity of the dataset.



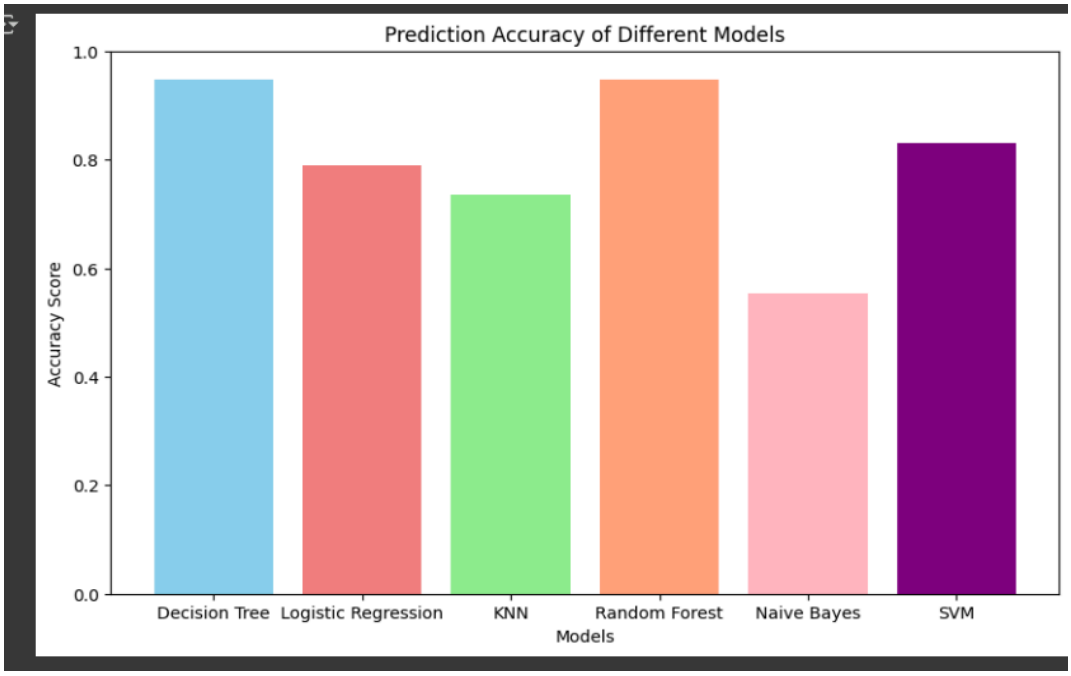
Clustering:

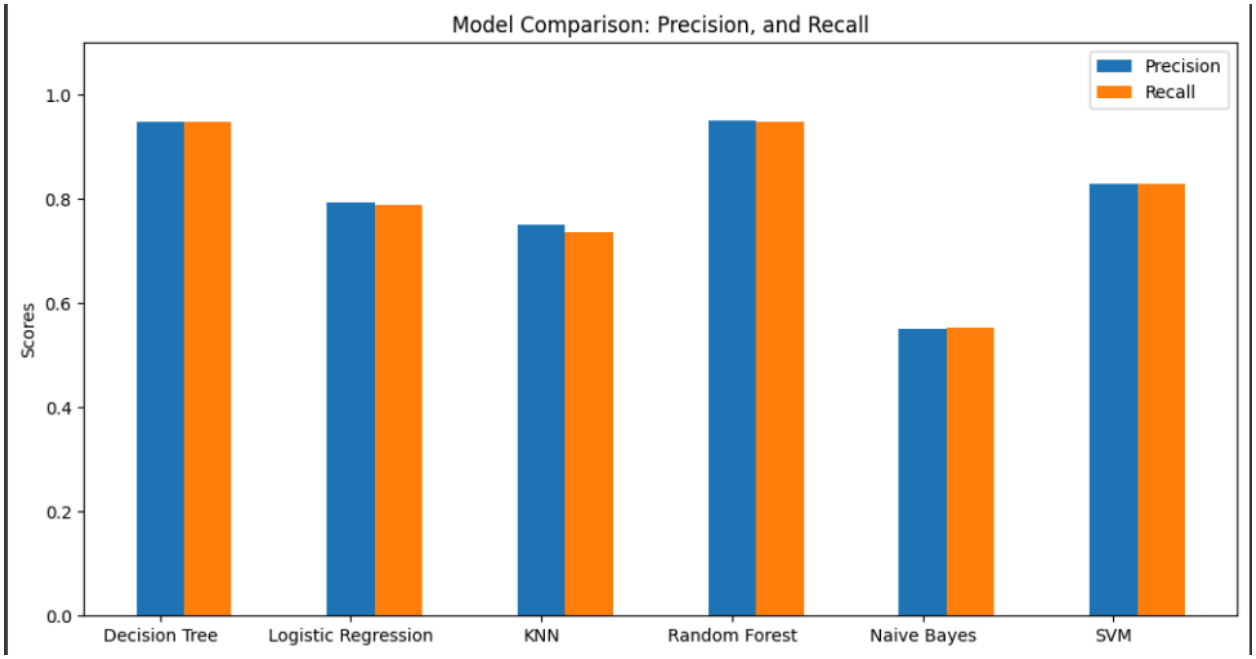
Here, Kmeans clustering algorithm was used in an attempt to divide the dataset into groups. PCA was applied to reduce the dimensionality to help us visualize the clusters on a 2D plane.



Model Result Comparison Analysis:

- Barchart showcasing accuracy of the applied models.





Model Selection:

Based on the results, Decision Tree and Random forest are the two models who score an accuracy of 0.94 each and f1 score of 0.95. So, for classification task on this dataset both Decision Tree or Random Forest will be a good choice.

Conclusion:

This project successfully demonstrates the application of machine learning to predict obesity levels based on eating habits, physical activity, and demographic data. By analyzing the dataset, we explored various features and addressed preprocessing steps like handling missing values, scaling, and encoding categorical data.

Through training and testing multiple models — Decision Tree, Logistic Regression, K-Nearest Neighbors, Random Forest, and Gaussian Naive Bayes — we compared their performance using metrics like accuracy and confusion matrices. These models provide valuable insights into identifying individuals at risk of obesity and highlight the potential of data-driven methods to support early intervention.

The project emphasizes the importance of lifestyle factors in determining obesity levels, advocating for actionable insights to promote healthier living. Our results demonstrate the ability of machine learning to provide accurate and interpretable predictions, showcasing its role in addressing real-world health challenges like obesity.