

Multi-Class Prediction of Obesity Risk

By

Venkata Ramana Reddy Duggempudi (vduggemp)

Pranavi Sriya Vajha (pvajha)

Venkata Sai Vikas Katuru (vkaturu)

Problem Statement:

To develop a multi-class prediction model to assess obesity risk, focusing on its association with cardiovascular disease. The model Utilizes diverse demographic, clinical, and lifestyle factors for accurate risk classification. The aim of model is to categorize individuals into distinct obesity risk classes, enabling personalized intervention strategies. Data pre-processing, feature selection/engineering, and model development are key steps. Employing machine learning algorithms that are suitable for multi-class classification, ensuring robust performance. Evaluating model performance using metrics like accuracy, precision, and AUC-ROC. Interpretability and validation of predictions are crucial for real-world applicability. The goal is to help spot people who are at risk of obesity related heart problems by early identification, which helps in taking steps to prevent those complications. Ultimately, the model seeks to improve public health by better targeting prevention and intervention efforts.

Background of the problem:

Obesity is one of the significant public health concerns across the globe, Obesity cases have been rising steadily over time. Which is making it as a complicated issue affected by many factors like genetics, environment, and lifestyle choices. Obesity is not only a risk factor for several chronic diseases but also significantly contributes to the development and progression of cardiovascular diseases

(CVD), including coronary artery disease, hypertension, and stroke. So, Early identification of individuals at risk of obesity and related cardiovascular complications is important for taking preventive measures and interventions.

Contribution of the Project towards Problem:

This project significantly addresses the problem of obesity and related cardiovascular health in several ways. They are:

- *Early Identification* – The model helps in early identification of individuals who are at risk of obesity-related cardiovascular complications. Which helps in taking preventive measures resulting in reducing the severity of health outcomes.
- *Personalized Risk Assessment* – The model offers a personalized approach in assessing the obesity risk. By considering various factors like diverse clinical, demographic and lifestyle, will provide a comprehensive evaluation to each individual characteristics, which enhances the accuracy of risk assessment.
- *Targeted Interventions* – The model facilitates targeted interventions strategies by categorizing individuals. Which makes health professionals prioritize resources for individuals who are at higher risk, that optimizes the use of healthcare resources and preventive measures.
- *Public Health Impact* – If the prevention and Intervention efforts related to obesity and its related issues were improved then the public health outcomes will be enhanced. It helps in reducing the burden of obesity and related complications on healthcare systems and society.
- *Data Driven Insights* – Upon analyzing data and interpreting the model's results, the project unfolds important insights into what factors contribute to obesity risk and how it's linked to heart health. This information can guide the people to make policies in making decisions, helping in shaping public health campaigns, and guide future research efforts aimed at fighting obesity and its associated health problems.

Data Sources:

For this project we have collected the data from Kaggle website. The dataset we are working on was introduced in Jan 2024 which is very recent, and the name of the problem and its link are:

Name: Multi – Class Prediction of Obesity Risk

Link: <https://www.kaggle.com/competitions/playground-series-s4e2/overview>

The dataset comprises of total 18 columns(features). In which 'NObeyesdad' column feature is the target variable. All the other columns are considered as Input features. The dataset consists of 20758 entries. The dataset consists of columns namely: id, Gender, Age, Height, family_history_with_overweight, FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC, MTRANS, NObeyesed.

Features Info:

id – Person Number

Gender - Person gender

Age – Age of the person

Height – Height of the person

Family_history_with_overweight – Is there any person in family with over weight

FAVC – Food and Vegetable Consumption

FCVC – Fruit and Vegetable Consumption

NCP – Nutritional Counseling Program

CAEC – Childhood Adverse Experiences and Childhood Obesity

SMOKE – Smoking

CH2O – Water Consumption

SCC – Sedentary Lifestyle and Central Obesity

FAF – Frequency of Fast-Food Consumption

TUE – Television Viewing and Obesity

CALC – Caloric Intake

MTRANS – Mode of Transportation

NObeyesed – It tells whether the person is obese or not and its types. And type of obese.

Data Cleaning:

Checking missing values: In this project, we worked with a dataset containing records of people with obesity conditions. We have checked the number of missing values in each column of the dataset using 'df.isnull().sum()'.

Removing Null values rows: To make the data more manageable and accurate, we have eliminated rows in which 'NCP', 'NObeyesed' values are missing.

Replacing null values of numerical data: In this dataset 'Age', 'Height', 'Weight', 'FCVC', 'CH2O', 'FAF', 'TUE' are the numerical attributes that are having the null values. Using 'mean' as replacing technique to replace the null values with the mean of that feature.

Replacing null values of Categorical data: In this dataset 'family_history_with_overweight', 'Gender', 'FAVC', 'CAEC', 'SMOKE', 'CALC', 'MTRANS' are the categorical features that are having the null values. Using 'Mode' as replacing technique we have replaced the null values with most repeated value.

Remove duplicates: We have removed all the duplicates rows in dataset.

Dropping Columns: We have removed the unnecessary feature in the dataset i.e., 'id' feature. After one hot encoding we have also dropped the "Gender" Column.

Renaming Columns: We have replaced some of the column names with other names. Such that they can be easily understandable. The columns names replaced are: 'CH2O' as 'Water Consumption', 'TUE' as 'Television viewing and Obesity'.

One hot encoding: Using this Technique we have converted the 'Gender' column into Two separate columns with 'male' and 'female' separately. Similarly, we have done one hot encoding for 'MTRANS' column also.

Categorical data encoding: Using this Technique we have converted the column named.

'family_history_with_overweight' values with 'yes' as '1' and 'no' as '0'. Similarly, for columns namely: 'FAVC', 'SMOKE', 'SCC' have applied same Technique.

Splitting Data: We have splitted the data into features and label where the label is 'NObedad' and all the other columns are considered as features.

Normalizing: Applied standard scalar Technique to all the features in the dataset. Which helps in improving the accuracy.

Explanation and Analysis

Algorithms applied for the dataset are: Logistic Regression, K-nearest neighbor, Naïve Bayes, SVM, Decision Tree Classifier and Neural Network.

Need of choosing specific algorithms: We're currently focused on a supervised learning task that involves classification. To improve our

results, we're exploring algorithms that are known to perform effectively for classification tasks.

1- Logistic Regression:

Justification: It is an algorithm that is widely used for binary classification tasks but can be extended for multi-class classification task. The reason for choosing this algorithm is, it is particularly suitable when the relationship between the independent variables and the dependent variable is linear. In the context of predicting obesity risk, logistic regression can be effective if the relationship between the features (e.g., BMI, diet habits, physical activity) and the risk categories (e.g., underweight, normal weight, overweight, obese) is approximately linear.

Work to do to tune/train the model: During phase-1 we have done tasks like data-preprocessing, feature selection and Exploratory Data Analysis (EDA). Now, in phase-2 before applying the logistic regression we have split the data (obtained after phase-1) into train set and test set.

Effectiveness of algorithm and observation:

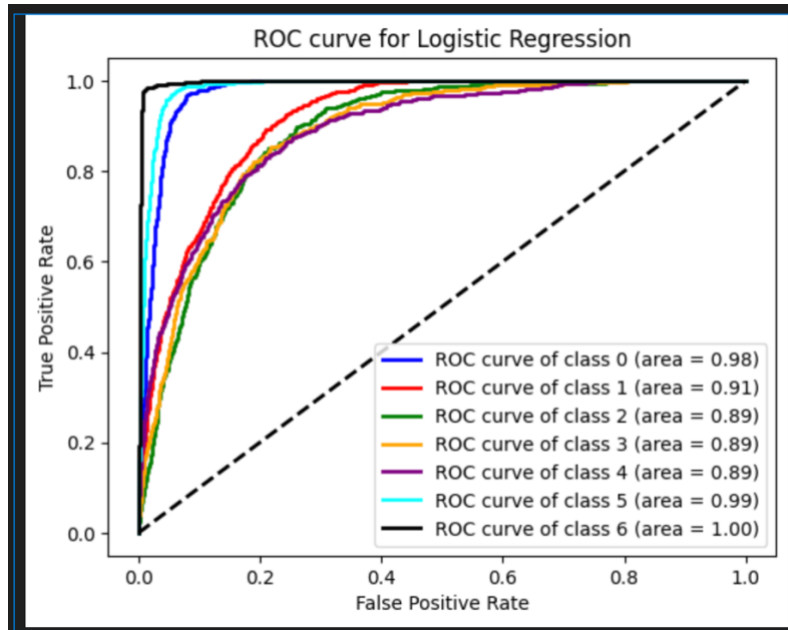
- *Results of logistic regression (metrics):*

```
Confusion Matrix with logistic regression:
[[489 149  4  0  0  0  1]
 [150 485 93 33 12  1  5]
 [  6 151 250 85 61 26 20]
 [  1  37 122 261 149 47 21]
 [  0  4  42  99 377 139 51]
 [  0  0  0  7  34 778  1]
 [  1  0  0  0 12  1 985]]
Accuracy Score with logistic regression: 0.6984585741811176
Precision Score with logistic regression: 0.6833907279326522
Recall Score with logistic regression: 0.6984585741811176
F1 Score with logistic regression: 0.6877128705740079
Classification Report score with logistic regression:
      precision    recall  f1-score   support

0         0.76         0.76         0.76         643
1         0.59         0.62         0.60         779
2         0.49         0.42         0.45         599
3         0.54         0.41         0.46         638
4         0.58         0.53         0.56         712
5         0.78         0.95         0.86         820
6         0.91         0.99         0.95         999

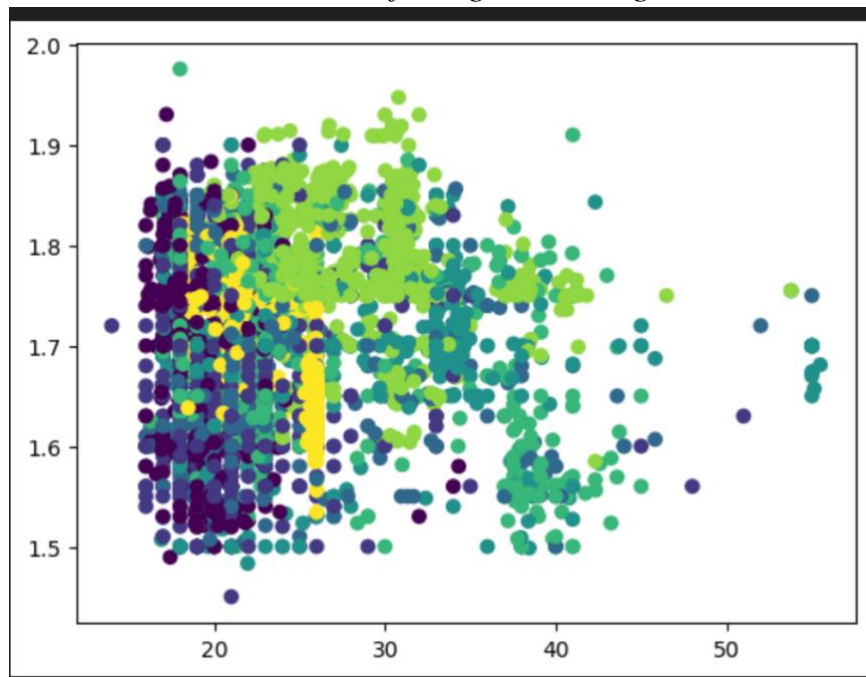
 accuracy         0.70         5190
 macro avg         0.66         0.67         0.66         5190
weighted avg         0.68         0.70         0.69         5190
```

ROC curve for Logistic Regression:

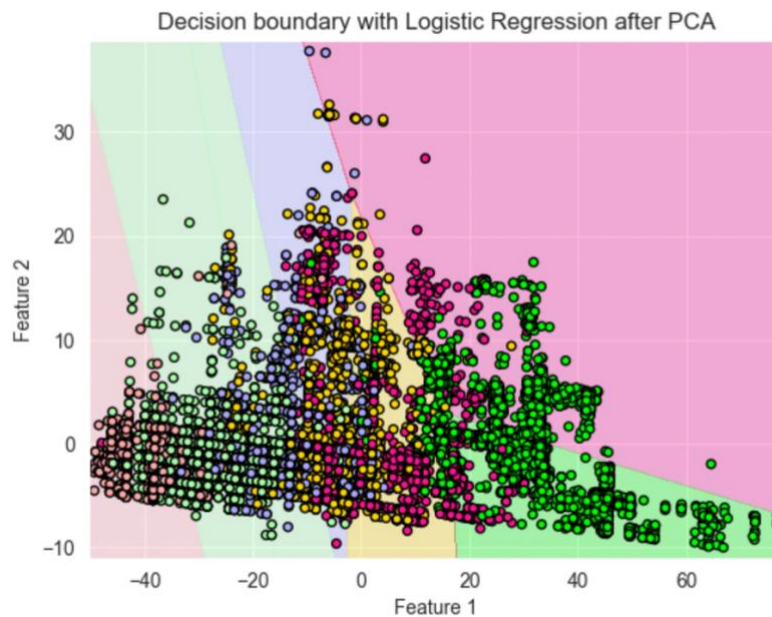


- Visualization of dataset (plots):
 - PCA done for dataset features to get plots as we applied Logistic Regression.

Scatter Plot for Age and Height



- Classification plot of Logistic Regression After PCA:



- *Learnings and observations:*

- From the results we can observe that accuracy-0.698, precision-0.683, recall-0.698, F1-score-0.687. These numbers gives an idea that how the algorithm is performing for the considered dataset and how good the model is in identifying the intricate patterns in the data.
- Precision and Recall values being close to each other indicate that the model maintains a good balance between minimizing false positives and false negatives.
- The area under the ROC curve (AUC) is the measure of the classifier's ability to distinguish between the classes. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents a no-skill classifier, equivalent to random guessing. We can see from the ROC curve that all the values are close to 1.0 which tells that the curve is almost a perfect classifier.
- The model's performance highlights the importance of these features in accurately categorizing individuals into different risk levels, indicating that interventions targeting these factors could be effective in mitigating obesity risk.
- Overall, the results demonstrate that the logistic regression algorithm is a promising approach (but not the best model among all) for predicting obesity risk categories. Even in the plots we can observe that the classification of data points it good but not

best, but it provides valuable insights for addressing obesity-related challenges and informing targeted interventions or policies to a good extent. But, Cannot be much useful because other models are performing better than this in classifying.

2- K-nearest neighbor(KNN):

Justification: It is simple and effective algorithm for classification tasks, especially it performs well when there is no clear boundary between different classes. In the case of obesity risk prediction, individuals with similar characteristics (e.g., similar weight & height patterns, similar dietary patterns) are likely to fall into similar risk categories. KNN works well in such scenarios as it classifies instances based on their similarity to neighboring instances in the feature space.

Work to do to tune/train the model: During phase-1 we have done tasks like data-preprocessing, feature selection and Exploratory Data Analysis (EDA). Now, in phase-2 before applying the KNN algorithm we have split the data (obtained after phase-1) into train set and test set.

Effectiveness of algorithm and observation:

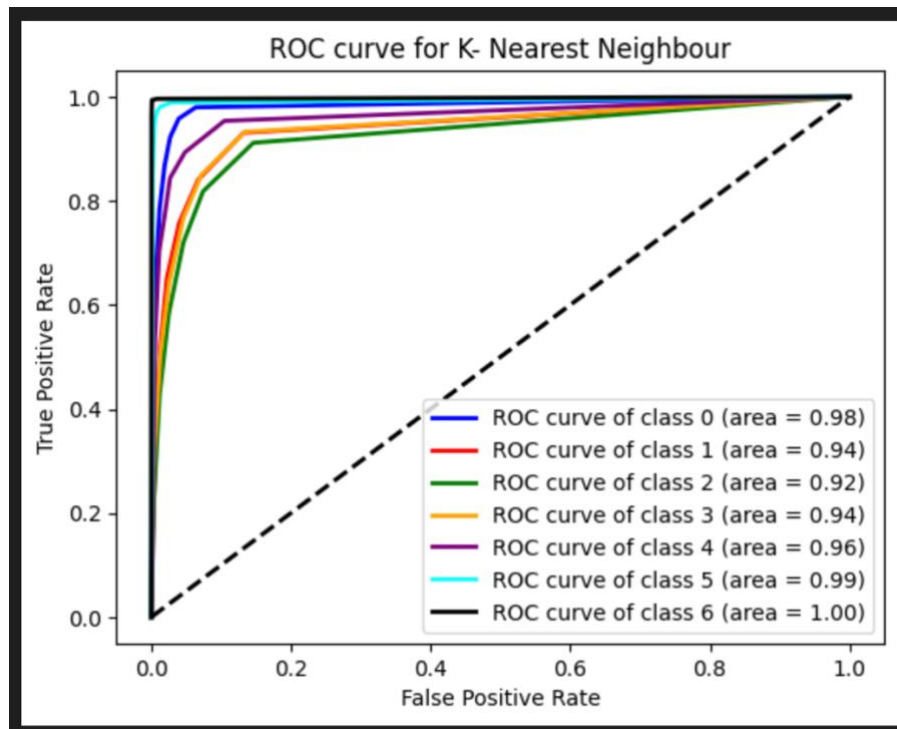
- *Results of KNN algorithm (metrics):*

```
Confusion Matrix with K-Nearest Neighbors:
[[593  50   0   0   0   0   0]
 [120 559  79  20   1   0   0]
 [  1  81 415  85  17   0   0]
 [  0  22  89 466  58   3   0]
 [  1   0  23  83 577  23   5]
 [  0   0   0   4  25 791   0]
 [  1   0   0   0   2   4 992]]
Accuracy Score with K-Nearest Neighbors: 0.8464354527938343
Precision Score with K-Nearest Neighbors: 0.8466983852912797
Recall Score with K-Nearest Neighbors: 0.8464354527938343
F1 Score with K-Nearest Neighbors: 0.845942196607488
Classification Report score with K-Nearest Neighbors:
      precision    recall  f1-score   support

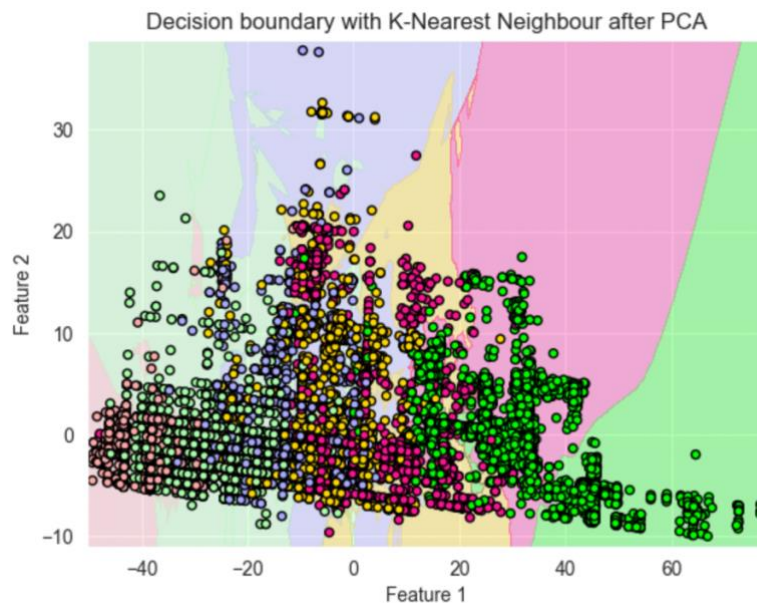
0           0.83       0.92       0.87         643
1           0.79       0.72       0.75         779
2           0.68       0.69       0.69         599
3           0.71       0.73       0.72         638
4           0.85       0.81       0.83         712
5           0.96       0.96       0.96         820
6           0.99       0.99       0.99         999

 accuracy          0.85         5190
 macro avg         0.83         0.83         0.83         5190
 weighted avg      0.85         0.85         0.85         5190
```

- *ROC curve for KNN algoirthm:*



- *Visualization of dataset (plots):*



- *Learnings and observations:*

- From the results we can observe that accuracy-0.846, precision-0.846, recall-0.846, F1-score-0.845. The high accuracy, precision, recall, and F1-score values suggest that the KNN algorithm is highly effective for predicting obesity risk categories in the dataset.
- The consistency in performance across all metrics indicates that the KNN model reliably classifies individuals into different

obesity risk categories.

- The model's performance highlights the importance of these features in accurately categorizing individuals into different risk levels, indicating that interventions targeting these factors could be effective in mitigating obesity risk.
- The area under the ROC curve (AUC) is the measure of the classifier's ability to distinguish between the classes. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents a no-skill classifier, equivalent to random guessing. We can see from the ROC curve that all the values are close to 1.0 and better than Logistic Regression which tells that the curve is almost a perfect classifier.
- Even in the plots we can observe that the classification of data points is very effective, and we can observe that the classification is not perfectly linear and especially it performs well when there is no clear boundary between different classes as we can observe it in the above plot.
- Overall, KNN algorithm is highly effective for predicting obesity risk categories, providing valuable insights for personalized health interventions or strategies aimed at mitigating obesity-related risks and it provides valuable insights for addressing obesity-related challenges and informing targeted interventions or policies to a good extent.

3- Naïve Bayes:

Justification: It is a probabilistic classifier that assumes independence between features. It's known for its simplicity, speed, and scalability. In the context of obesity risk prediction, if the features are conditionally independent given the class (e.g., height, weight, diet habits, physical activity), Naïve Bayes can provide fast and efficient classification if we consider the features independent.

Work to do to tune/train the model: During phase-1 we have done tasks like data-preprocessing, feature selection and Exploratory Data Analysis (EDA). Now, in phase-2 before applying the Naïve Bayes algorithm we have split the data (obtained after phase-1) into train set and test set.

Effectiveness of algorithm and observation:

- *Results of Naïve Bayes (metrics):*

Confusion Matrix with Naive Bayes:

```
[[489  32   5   1 112   0   4]
 [279 182  51  38 202   3  24]
 [ 42  62 127  44 291  27   6]
 [  9  31  19 128 359  91   1]
 [  1   4  14  31 447 212   3]
 [  0   2   0   9  30 779   0]
 [  1   0   0   2   0   1 995]]
```

Accuracy Score with Naive Bayes: 0.6063583815028901

Precision Score with Naive Bayes: 0.6296643799579701

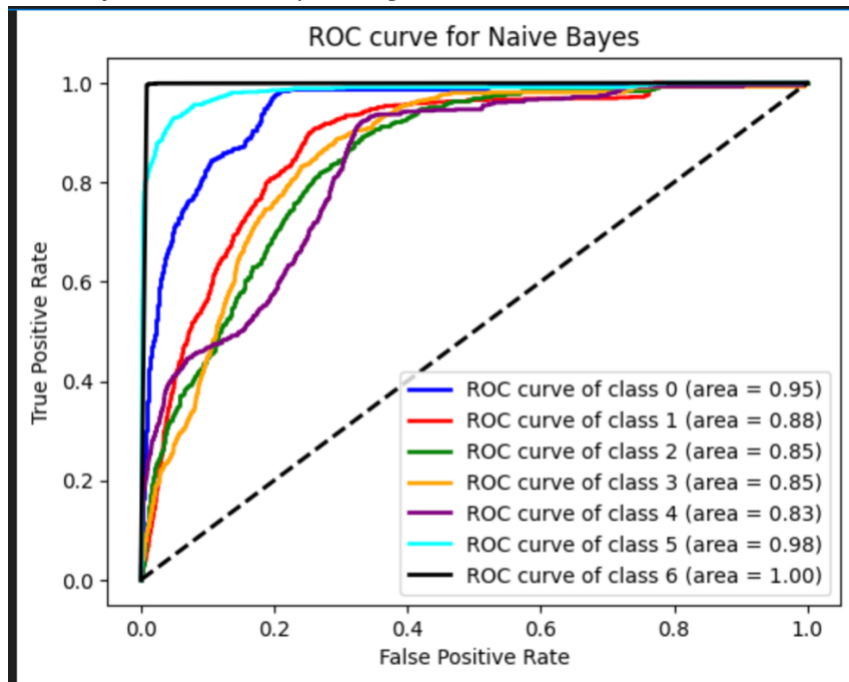
Recall Score with Naive Bayes: 0.6063583815028901

F1 Score with Naive Bayes: 0.5769020866576603

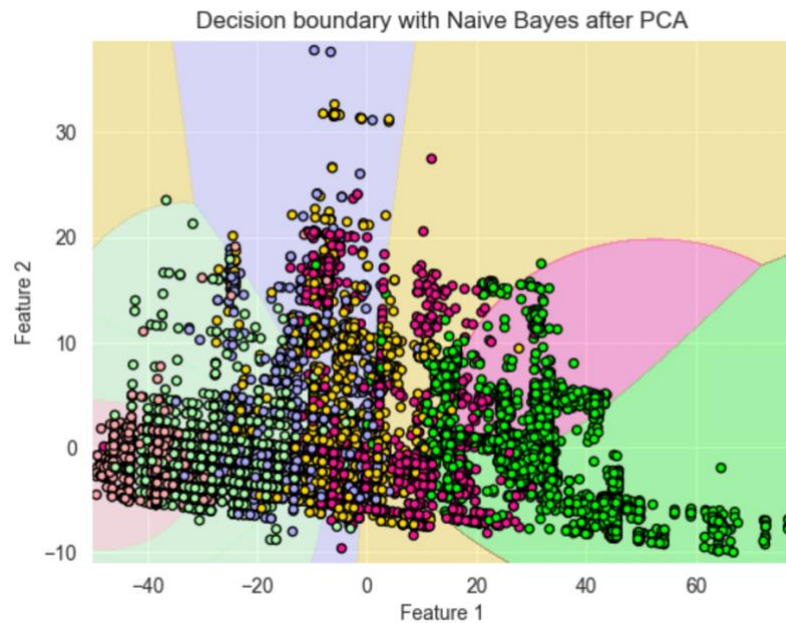
Classification Report score with Naive Bayes:

	precision	recall	f1-score	support
0	0.60	0.76	0.67	643
1	0.58	0.23	0.33	779
2	0.59	0.21	0.31	599
3	0.51	0.20	0.29	638
4	0.31	0.63	0.42	712
5	0.70	0.95	0.81	820
6	0.96	1.00	0.98	999
accuracy			0.61	5190
macro avg	0.61	0.57	0.54	5190
weighted avg	0.63	0.61	0.58	5190

- *ROC curve for Naïve Bayes algorithm:*



- *Visualization of dataset (plots):*



- *Learnings and observations:*
 - From the results we can observe that accuracy-0.606, precision-0.629, recall-0.606, F1-score-0.576. The moderate accuracy, precision, recall, and F1-score values suggest that the Naïve Bayes algorithm performs moderately well for predicting obesity risk categories in the dataset.
 - While the model's performance is not as high as some other algorithms, Naïve Bayes is known for its simplicity and speed, making it a computationally efficient choice for classification tasks.
 - These results suggest that while Naïve Bayes may not be the most accurate algorithm for this specific dataset, it still provides some predictive power and can be considered as an alternative in situations where computational resources are limited or when interpretability is crucial.
 - The area under the ROC curve (AUC) is the measure of the classifier's ability to distinguish between the classes. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents a no-skill classifier, equivalent to random guessing. We can see from the ROC curve that few of the values are close to 1.0 and few are in between 0.5 to 1 which tells that the curve is neither perfect nor no skill classifier.
 - Even in the plots we can observe that the datapoints are not classified accurately, and we can observe that the classification is not perfectly linear.

- Overall, while the Naïve Bayes algorithm may not achieve the highest performance metrics, it still offers a valuable approach for predicting obesity risk categories, particularly in scenarios where computational efficiency and simplicity are prioritized.

4- Support Vector Machine (SVM):

Justification: It is the powerful and efficient algorithm for classification tasks, especially when dealing with high-dimensional data or when there is a clear margin of separation between classes. In the context of obesity risk prediction, SVM can effectively classify individuals into different risk categories by finding the hyperplane that maximizes the margin between classes. It's particularly useful when the decision boundary is nonlinear and complex, as SVM can use different kernel functions to map the data into a higher-dimensional space where separation is easier.

Work to do to tune/train the model: During phase-1 we have done tasks like data-preprocessing, feature selection and Exploratory Data Analysis (EDA). Now, in phase-2 before applying the SVM algorithm we have split the data (obtained after phase-1) into train set and test set.

Effectiveness of algorithm and observation:

- *Results of Support Vector Machine (metrics):*

Confusion Matrix with SVM with linear kernel:

```
[[615 28 0 0 0 0 0]
 [ 73 628 72 3 3 0 0]
 [ 1 58 424 102 14 0 0]
 [ 0 16 93 437 87 5 0]
 [ 1 1 6 77 583 43 1]
 [ 0 0 0 4 17 799 0]
 [ 1 0 0 0 1 1 996]]
```

Accuracy Score with SVM with linear kernel: 0.8635838150289017

Precision Score with SVM with linear kernel: 0.8622914504747091

Recall Score with SVM with linear kernel: 0.8635838150289017

F1 Score with SVM with linear kernel: 0.8626046673286105

Classification Report score with SVM with linear kernel:

	precision	recall	f1-score	support
0	0.89	0.96	0.92	643
1	0.86	0.81	0.83	779
2	0.71	0.71	0.71	599
3	0.70	0.68	0.69	638
4	0.83	0.82	0.82	712
5	0.94	0.97	0.96	820
6	1.00	1.00	1.00	999
accuracy			0.86	5190
macro avg	0.85	0.85	0.85	5190
weighted avg	0.86	0.86	0.86	5190

Confusion Matrix with SVM with rbf kernel:

```
[[542 101 0 0 0 0 0]
 [180 454 127 17 1 0 0]
 [ 5 116 276 187 15 0 0]
 [ 1 38 42 459 86 12 0]
 [ 1 0 6 293 354 34 24]
 [ 0 0 0 2 47 733 38]
 [ 1 0 0 0 2 0 996]]
```

Accuracy Score with SVM with rbf kernel: 0.7348747591522158

Precision Score with SVM with rbf kernel: 0.7436653537986903

Recall Score with SVM with rbf kernel: 0.7348747591522158

F1 Score with SVM with rbf kernel: 0.7318546411126785

Classification Report score with SVM with rbf kernel:

	precision	recall	f1-score	support
0	0.74	0.84	0.79	643
1	0.64	0.58	0.61	779
2	0.61	0.46	0.53	599
3	0.48	0.72	0.58	638
4	0.70	0.50	0.58	712
5	0.94	0.89	0.92	820
6	0.94	1.00	0.97	999
accuracy			0.73	5190
macro avg	0.72	0.71	0.71	5190
weighted avg	0.74	0.73	0.73	5190

Confusion Matrix with SVM with poly kernel:

```
[[545 98 0 0 0 0 0]
 [199 449 115 15 1 0 0]
 [ 7 145 265 170 12 0 0]
 [ 2 41 55 451 79 10 0]
 [ 1 0 8 290 369 41 3]
 [ 0 0 0 2 45 772 1]
 [ 1 0 0 0 2 0 996]]
```

Accuracy Score with SVM with poly kernel: 0.7412331406551059

Precision Score with SVM with poly kernel: 0.7497266788505963

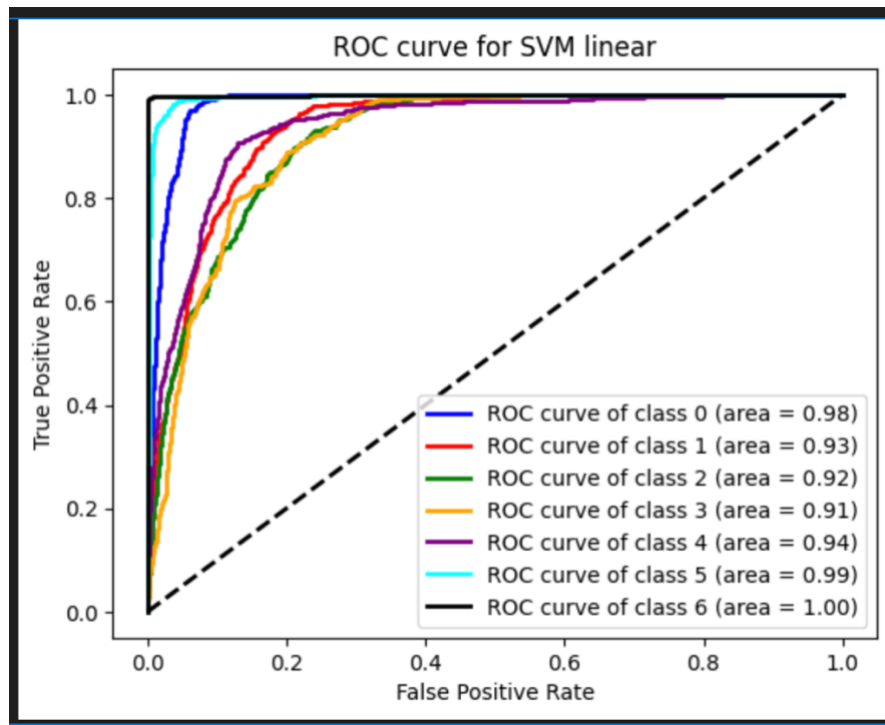
Recall Score with SVM with poly kernel: 0.7412331406551059

F1 Score with SVM with poly kernel: 0.7385253637089724

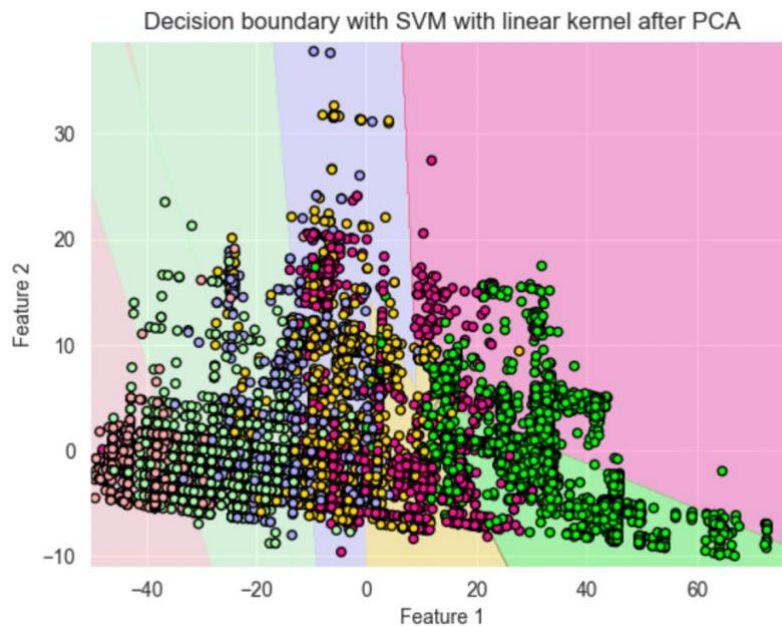
Classification Report score with SVM with poly kernel:

	precision	recall	f1-score	support
0	0.72	0.85	0.78	643
1	0.61	0.58	0.59	779
2	0.60	0.44	0.51	599
3	0.49	0.71	0.58	638
4	0.73	0.52	0.60	712
5	0.94	0.94	0.94	820
6	1.00	1.00	1.00	999
accuracy			0.74	5190
macro avg	0.73	0.72	0.71	5190
weighted avg	0.75	0.74	0.74	5190

- *ROC curve for SVM with linear kernel:*



- *Visualization of dataset (plots):*



- *Learnings and observations:*

- From the results we can observe that accuracy-0.863, precision-0.862, recall-0.863, F1-score-0.862. The high accuracy, precision, recall, and F1-score values suggest that the SVM model with a linear classifier performs very well for predicting obesity risk categories in the dataset.
- Compared to other SVM kernels like radial basis function

(RBF) or polynomial, the linear SVM model appears to provide better performance in this dataset, indicating that the data might be better suited for linear separation.

- Linear SVM classifiers are particularly effective when the data can be separated well by a linear boundary, and in this case, it seems that the obesity risk categories can be effectively separated by a linear decision boundary.
- The area under the ROC curve (AUC) is the measure of the classifier's ability to distinguish between the classes. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents a no-skill classifier, equivalent to random guessing. We can see from the ROC curve that all the values are close to 1.0 which tells that the curve is almost a perfect classifier.
- Even in the plots we can observe that the datapoints are classified in a better manner, and we can observe that the data is classified linearly and SVM performs especially well when dealing with high-dimensional data or when there is a clear margin of separation between classes.
- Overall, the results demonstrate that the SVM model with a linear classifier is highly effective for predicting obesity risk categories, providing valuable insights for personalized health interventions or strategies aimed at mitigating obesity-related risks.

5- Decision Tree Classifier:

Justification: Decision trees are used for both classification and regression task. Decision trees are basically intuitive and easy to interpret, which makes them suitable for tasks where understanding the reasoning behind predictions is important. In the case of obesity risk prediction, decision trees can provide insights into which features are most influential in determining the risk categories (e.g., dietary factors, physical activity level). Decision trees can handle both numerical and categorical data and are robust to outliers and irrelevant features.

Work to do to tune/train the model: During phase-1 we have done tasks like data-preprocessing, feature selection and Exploratory Data Analysis (EDA). Now, in phase-2 before applying the Decision Tree Classifier we have split the data (obtained after phase-1) into train set and test set.

Effectiveness of algorithm and observation:

- Results of Decision Tree Classifier (metrics):

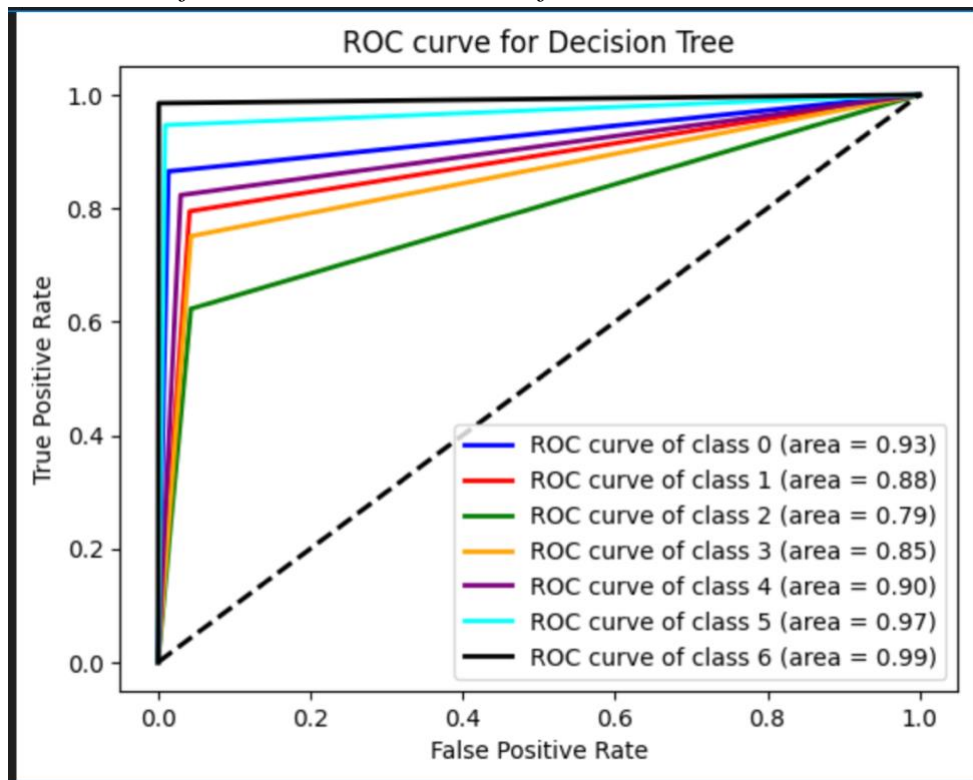
Confusion Matrix with Decision Tree:

```
[[556 83 4 0 0 0 0]
 [ 51 619 87 17 4 0 1]
 [ 7 81 373 112 24 2 0]
 [ 0 15 80 479 57 7 0]
 [ 2 4 26 61 586 32 1]
 [ 0 0 1 7 34 776 2]
 [ 1 0 0 0 14 0 984]]
```

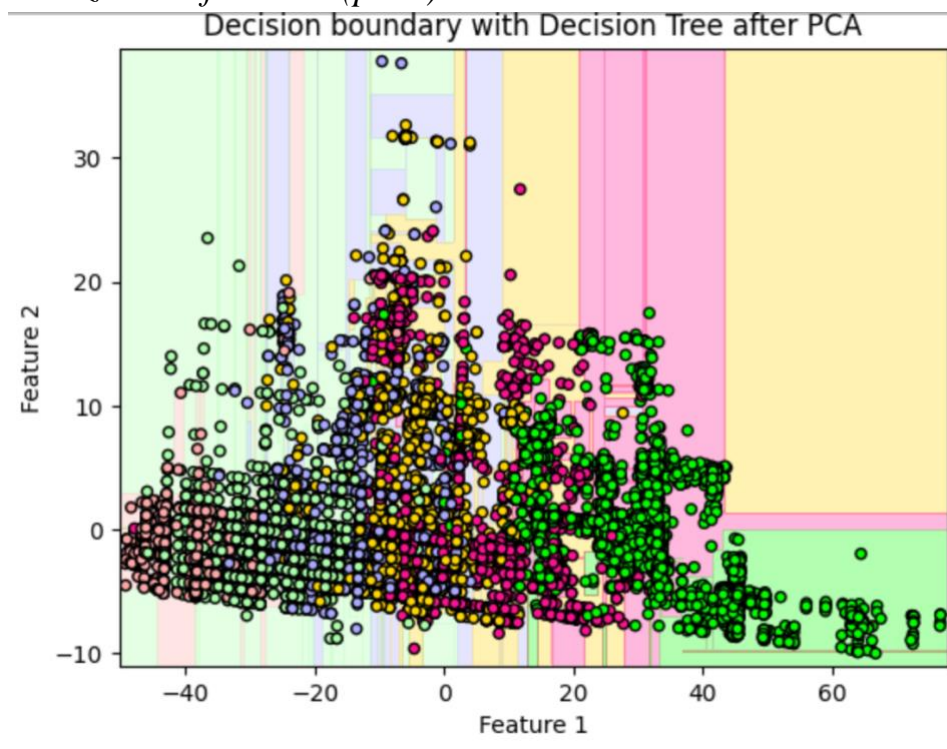
Accuracy Score with Decision Tree: 0.8425818882466282
Precision Score with Decision Tree: 0.8435726548238058
Recall Score with Decision Tree: 0.8425818882466282
F1 Score with Decision Tree: 0.8428795929043639
Classification Report score with Decision Tree:

	precision	recall	f1-score	support
0	0.90	0.86	0.88	643
1	0.77	0.79	0.78	779
2	0.65	0.62	0.64	599
3	0.71	0.75	0.73	638
4	0.82	0.82	0.82	712
5	0.95	0.95	0.95	820
6	1.00	0.98	0.99	999
accuracy			0.84	5190
macro avg	0.83	0.83	0.83	5190
weighted avg	0.84	0.84	0.84	5190

- ROC curve for Decision Tree classifier:



- *Visualization of dataset (plots):*



- *Learnings and observations:*
 - From the results we can observe that accuracy-0.846, precision-0.847, recall-0.846, F1-score-0.847. The high accuracy, precision, recall, and F1-score values suggest that the Decision Tree Classifier model performs very well for predicting obesity risk categories in the dataset.
 - Decision trees are known for their simplicity and interpretability, making them useful for gaining insights into the factors influencing obesity risk prediction.
 - Decision trees can capture complex relationships between features and the target variable, which can provide valuable insights into the decision-making process.
 - The area under the ROC curve (AUC) is the measure of the classifier's ability to distinguish between the classes. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents a no-skill classifier, equivalent to random guessing. We can see from the ROC curve that few of the values are close to 1.0 and few are in between 0.5 to 1 which tells that the curve is neither perfect nor no skill classifier.

- Even in the plots we can observe that the datapoints are classified well, suggest that the Decision Tree Classifier is effective at capturing the underlying patterns in the data and can provide valuable insights into the relationships between different features and obesity risk.
- Overall, the results demonstrate that the Decision Tree Classifier is highly effective for predicting obesity risk categories, providing valuable insights for personalized health interventions or strategies aimed at mitigating obesity-related risks.

6- Neural Network:

Justification: Neural networks, particularly deep learning models, are known for their ability to capture complex patterns and relationships in data. In the context of obesity risk prediction, neural networks can potentially uncover intricate associations between various features and the risk categories. They excel at learning hierarchical representations of data, which might be beneficial when dealing with diverse and multi-dimensional feature sets. However, neural networks typically require a large amount of data and computational resources for training, and they can be less interpretable compared to simpler models like decision trees or logistic regression.

Work to do to tune/train the model: During phase-1 we have done tasks like data-preprocessing, feature selection and Exploratory Data Analysis (EDA). Now, in phase-2 before applying the Neural Network we have split the data (obtained after phase-1) into train set and test set.

Effectiveness of algorithm and observation:

→ Test results:

```
# Evaluate the model on the test data and obtain loss and accuracy
loss, accuracy = model.evaluate(X_test, y_test, verbose=0)

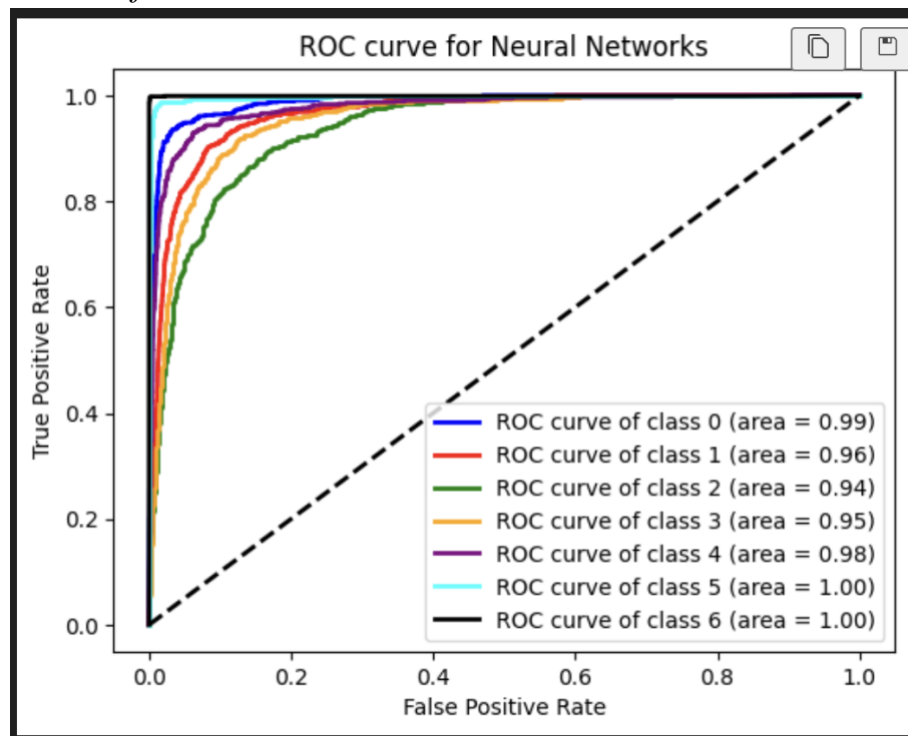
# Predict class labels for test data
y_pred = model.predict(X_test)

print('Accuracy Score with Neural Networks:', accuracy)
```

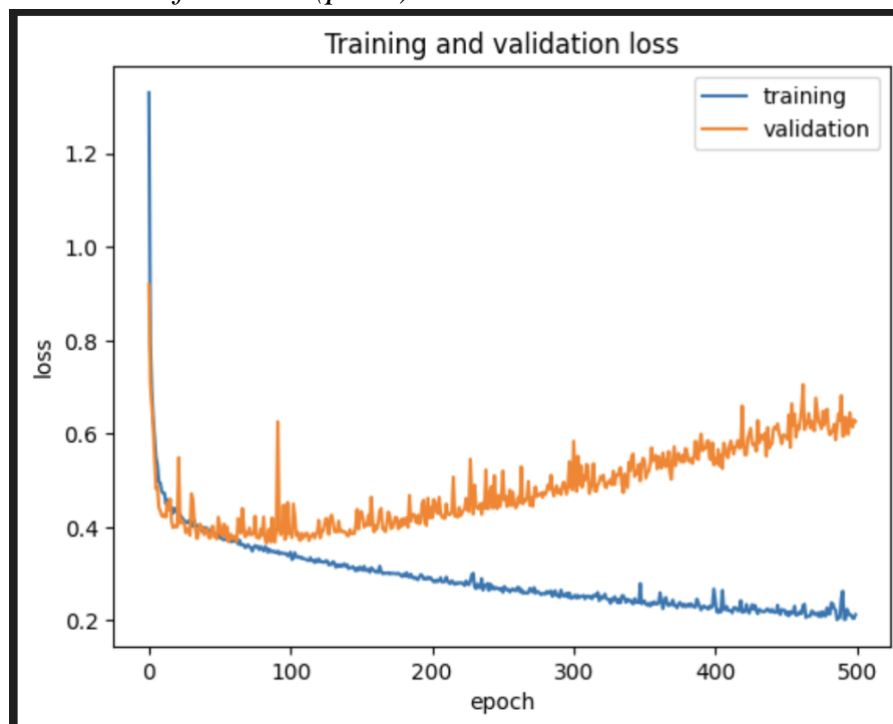
Python

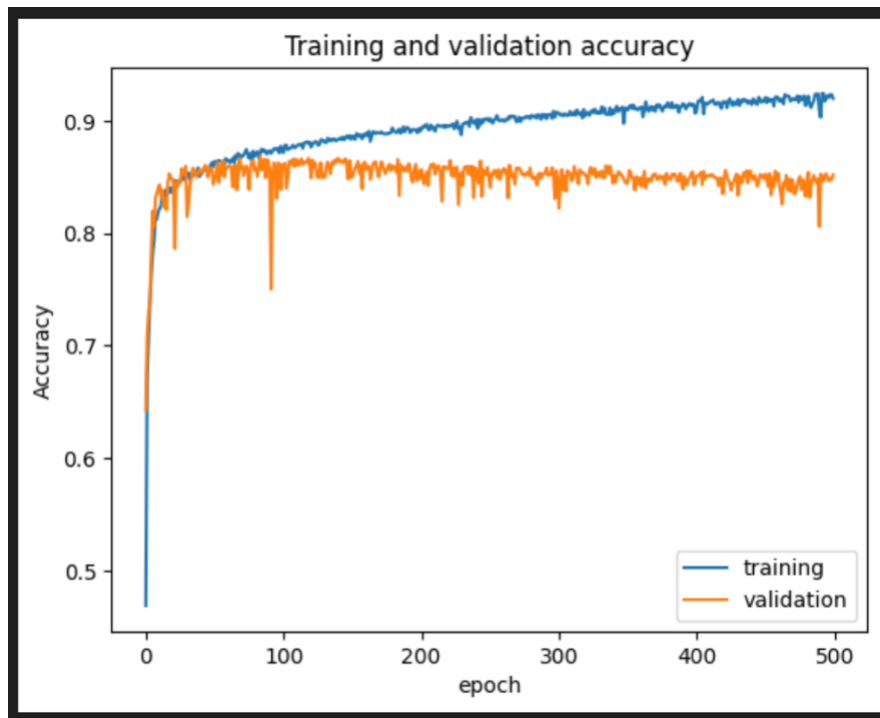
163/163 ————— 0s 286us/step
Accuracy Score with Neural Networks: 0.8518304228782654

- *ROC curve for Neural Network:*



- *Visualization of dataset (plots):*





- *Learnings and observations:*
 - Here, the neural network model is trained for 500 epochs. And then it is subjected to test data which resulted in an accuracy of 0.851.
 - The high accuracy suggests that the neural network effectively captures the underlying patterns in the data and can make reliable predictions about obesity risk categories. However, it's important to consider other performance metrics such as precision, recall, and F1-score to get a comprehensive understanding of the model's effectiveness.
 - Neural Networks can capture complex relationships between features and the target variable, which can provide valuable insights into the model's performance.
 - The area under the ROC curve (AUC) is the measure of the classifier's ability to distinguish between the classes. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents a no-skill classifier, equivalent to random guessing. We can see from the ROC curve that all the values are close to 1.0 which tells that the curve is almost a perfect classifier.
 - Even in the plots we can observe that the datapoints are classified well, and the model is performing well for the test data as the network is trained for 50 epochs, resulting it to learn the intricate patterns from it. We can also say that there

are chances of getting better performance if we train the model for more epochs.

- Overall, achieving an accuracy of 0.851 after training the neural network for 50 epochs indicates that the model has learned to effectively classify individuals into different obesity risk categories and demonstrates promising performance in predicting obesity-related risks.

- Sample predicted results for neural networks:

```
highest_labels_mapped=prediction(X, model)
highest_labels_mapped

433/433 ————— 0s 207us/step

['Obesity_Type_II',
 'Overweight_Level_I',
 'Obesity_Type_III',
 'Obesity_Type_I',
 'Obesity_Type_III',
 'Insufficient_Weight',
 'Insufficient_Weight',
 'Normal_Weight',
 'Overweight_Level_II',
 'Normal_Weight',
 'Insufficient_Weight',
 'Obesity_Type_III',
 'Obesity_Type_III',
 'Obesity_Type_I',
 'Obesity_Type_III',
 'Overweight_Level_II',
 'Obesity_Type_I',
 'Obesity_Type_I',
 'Obesity_Type_III',
 'Obesity_Type_I',
 'Normal_Weight',
 'Obesity_Type_III',
 'Obesity_Type_III',
 'Obesity_Type_II',
 'Overweight_Level_I',
 ...
 'Overweight_Level_I',
 'Insufficient_Weight',
 'Insufficient_Weight',
 'Insufficient_Weight',
 ...]
```

Inference from all the above models:

After using different ml methods on the dataset ‘Multi-class Prediction of Obesity Risk’. Based on the results we can conclude that the neural networks are comparatively performing better than other ml techniques. So, we have considered the neural network model for developing the user interface.

USER INTERFACE

The interface we have designed for the project is user friendly and works well. The images of the interface designed by us is provided below:

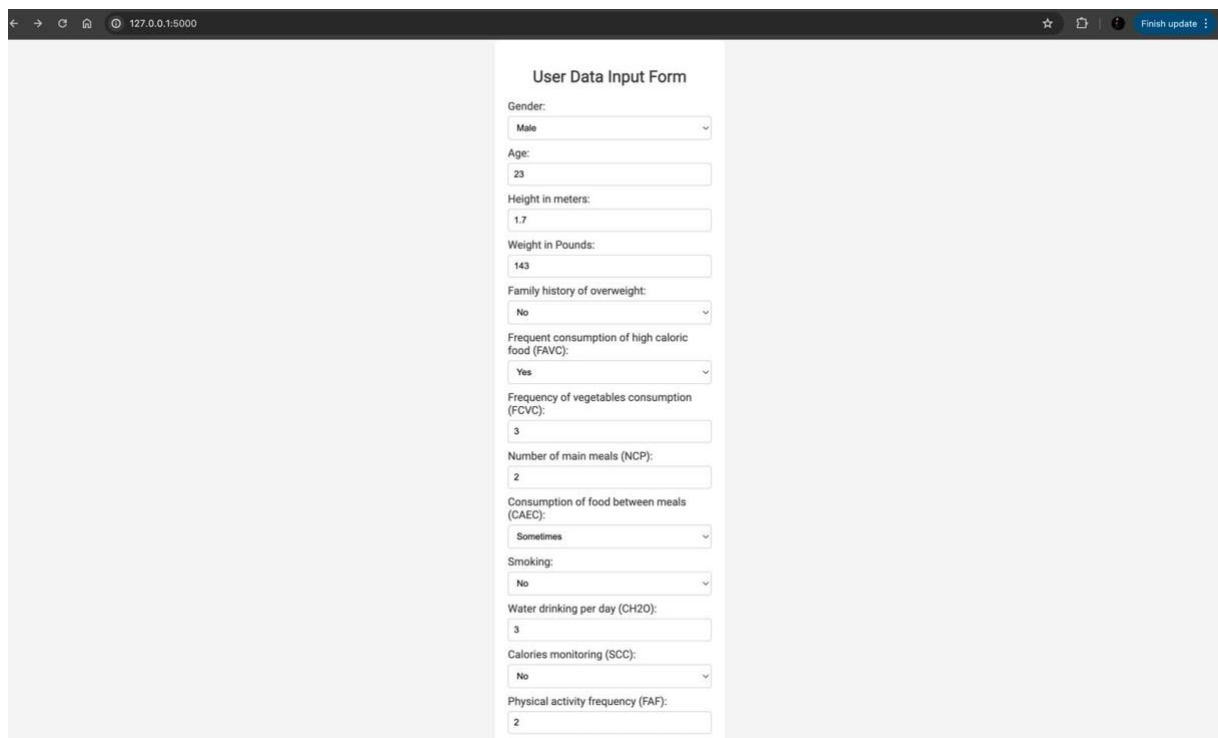
A screenshot of a web browser displaying a 'User Data Input Form'. The browser's address bar shows '127.0.0.1:5000'. The form is titled 'User Data Input Form' and contains the following fields: 'Gender' (dropdown menu with 'Male' selected), 'Age' (text input with '23'), 'Height in meters' (text input with '1.7'), 'Weight in Pounds' (text input with '143'), 'Family history of overweight' (dropdown menu with 'No' selected), 'Frequent consumption of high caloric food (FAVC)' (dropdown menu with 'Yes' selected), 'Frequency of vegetables consumption (FCVC)' (text input with '3'), 'Number of main meals (NCP)' (text input with '2'), 'Consumption of food between meals (CAEC)' (dropdown menu with 'Sometimes' selected), 'Smoking' (dropdown menu with 'No' selected), 'Water drinking per day (CH2O)' (text input with '3'), 'Calories monitoring (SCC)' (dropdown menu with 'No' selected), and 'Physical activity frequency (FAF)' (text input with '2'). The form is centered on a light gray background.

Fig: User Data Input form

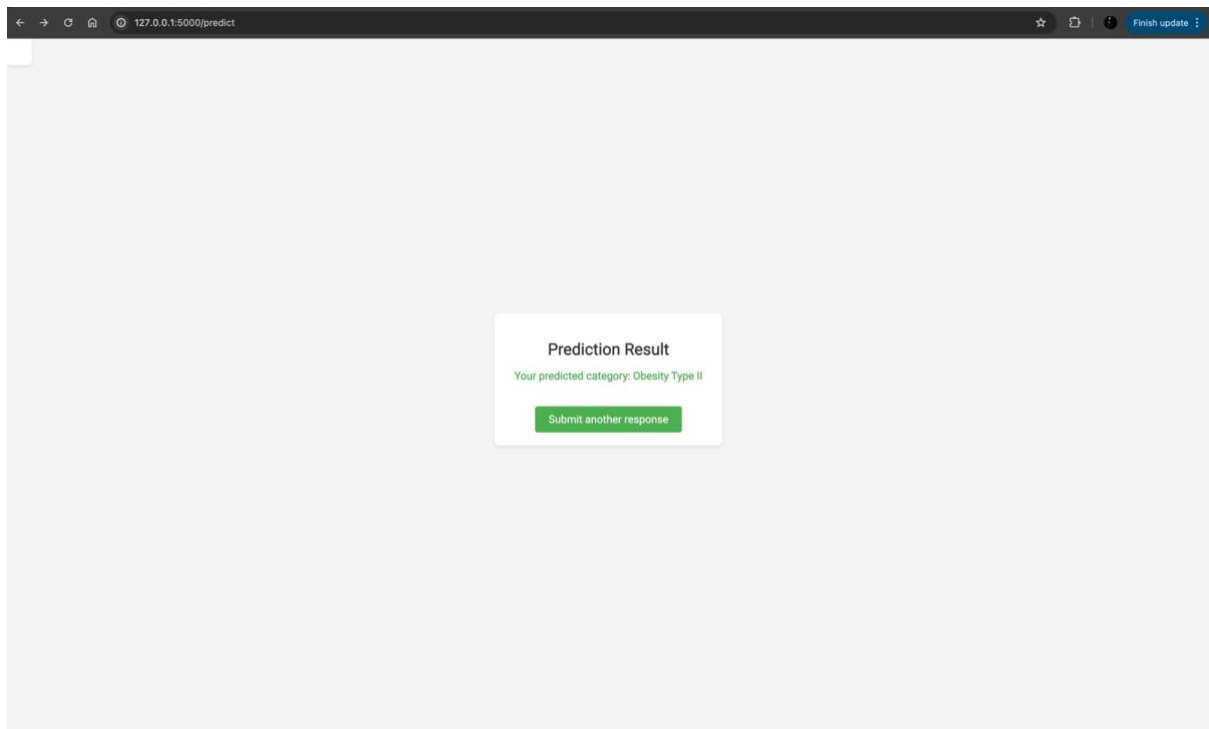


Fig: Prediction Results Page

Description of User Interface:

The Interface has 2 pages namely User Data Input Form and Prediction Results page.

1. User Data Input Form:

- The Fields in the form are neatly organized in a vertical sequence, which will be clear and easy to fill out.
- The form consists of various input fields which is useful in gathering essential data for obesity risk prediction.
- The fields are related to Demographic Information, Dietary Habits, Eating Patterns and Lifestyle of people.
- The fields present in the form are:
 - Gender, Age, Height in meters, Weight in Pounds, family history of overweight, Frequent consumption of high caloric food (FAVC), Frequency of vegetables consumption (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Smoking, Water drinking per day (CH2O), Calories monitoring (SCC), Physical activity frequency (FAF).

- Description of each field in the form:
 - Gender: The gender of the individual (e.g., Male or Female).
 - Age: The age of the individual.
 - Height: The height of the individual in meters.
 - Weight: The weight of the individual in kilograms.
 - family_history_with_overweight: A binary indicator (yes/no) whether there is a family history of being overweight.
 - FAVC: Frequent consumption of high caloric food (yes/no).
 - FCVC: Frequency of consumption of vegetables (on a scale from 1 to 3, where 3 indicates a higher frequency).
 - NCP: Number of main meals (on a scale, typically from 1 to 4).
 - CAEC: Consumption of food between meals (Never, Sometimes, Frequently, Always).
 - SMOKE: Smoking status of the individual (yes/no).
 - CH2O: Consumption of water daily (in litres).
 - SCC: Calories consumption monitoring (yes/no).
 - FAF: Physical activity frequency per week (on a scale from 0 to 3 where 3 indicates high frequency).
 - TUE: Time using technology devices per day (in hours).
 - CALC: Consumption of alcohol (Never, Sometimes, Frequently, Always).
 - MTRANS: Mode of Transportation (e.g., Public Transportation, Automobile).

- Regarding functionality:
 - For all the non-numerical fields we have provided dropdown menus where there will be a limited set of responses, which helps in reducing the risk of errors.
 - Text and number inputs are used where precise data is to be written, by making sure that the data collected is accurate and quantifiable.

- Each field is clearly labelled such that users understand what information is to be filled in each field. The form layout was created in a way to avoid confusion for the users.
- After providing the details in each field and 'submit' then it will be directed to Prediction Results Page.

2. Prediction Results Page:

- The results page was designed in such a way with a clear focus on presenting the prediction result. Used a large, central card to display the prediction which will be visible to users upon page load.
- The predictions obtained in the results page will be one of them:
 - 0: 'Insufficient Weight',
 - 1: 'Normal Weight',
 - 2: 'Overweight Level I',
 - 3: 'Overweight Level II',
 - 4: 'Obesity Type I',
 - 5: 'Obesity Type II',
 - 6: 'Obesity Type III'
- Regarding Functionality: Based on the provided details in the input form the results page displays the 'predicted obesity category' among the 7 provided above. Additionally, 'Submit another response' button is provided which allows new assessments or re-evaluation, such that encouraging user interaction and continuation engagement.

Link of user interface: <http://127.0.0.1:5000/>

Contribution: Everyone contributed equally.

References

1. https://scikit-learn.org/stable/modules/neural_networks_supervised.html
2. <https://scikit-learn.org/stable/modules/tree.html>
3. <https://scikit-learn.org/stable/>
4. <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
5. <https://developer.mozilla.org/en-US/docs/Web/HTML>
6. <https://developer.mozilla.org/en-US/docs/Web/CSS>