BRAC
UNIVERSITY

Inspiring Excellence

CSE422

Artificial Intelligence Lab

Project Title

Obesity Level Prediction Using Machine Learning Models

Section No. 28
Group No. 01
Semester: Summer_2025
Submitted Date: 04-09-2025
Submitted to - Mahjabin Chowdhury, Sadman Sakib Alif

Group Members:

Dihan Islam Dhrubo	23301458
Afnan Ul Islam Afif	23301528

Table Of Contents:

Introduction:

Our project focuses on predicting obesity levels based on demographic, dietary, and physical activity features. Obesity is a major global health issue, and early identification of risk factors can guide better health interventions. We use supervised machine learning models to classify individuals into different obesity categories, along with exploratory and unsupervised methods.

Dataset-Description:

- **Total Features**: 16
- **Total Data Points:** 2111
- **Problem Type:**

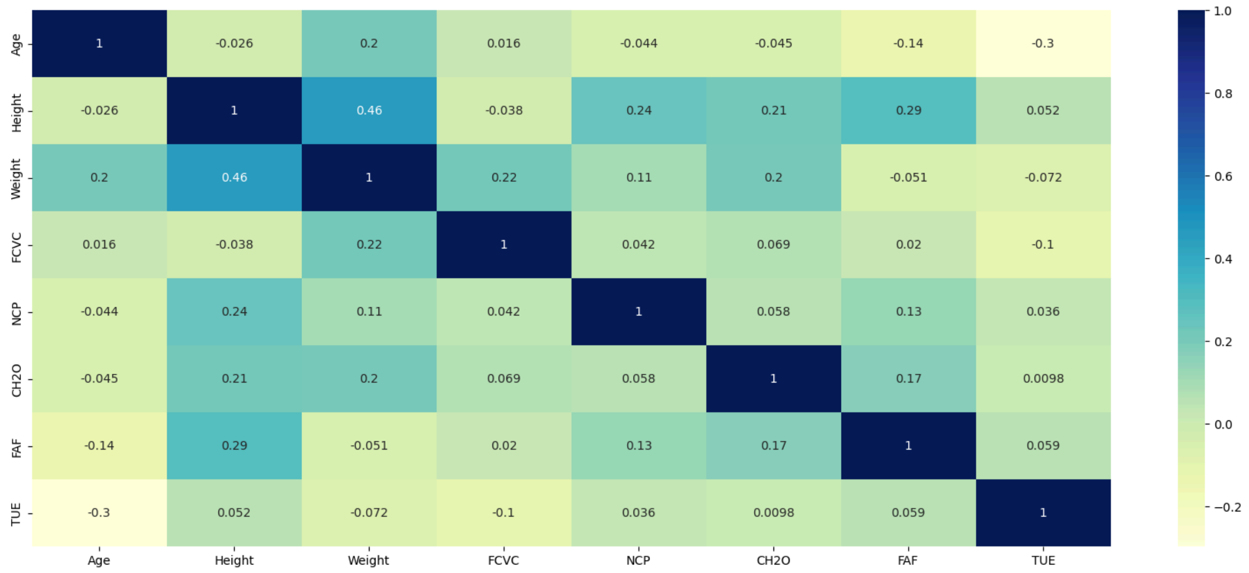
  This is a **Classification problem** because the target variable, NObeyesdad, contains distinct categories. It classifies the result into categories rather than quantitative values.

- **Feature Type:**
  - **Quantitative Feature:**

    Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE.
  - **Categorical Features:**

    Gender, CALC, FAVC, SCC, SMOKE, family_history_with_overweight, CAEC, MTRANS.

Dataset Heatmap before encoding:



Only numeric features (Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE) are included. Categorical features (like Gender, CAEC, CALC, etc.) are still strings, so they are excluded from the correlation matrix because correlation can only be computed on numeric data. Interpretation is limited: you only see relationships among continuous variables.

- ## Encoding Categorical Variables:

  Yes, categorical values have been encoded here.

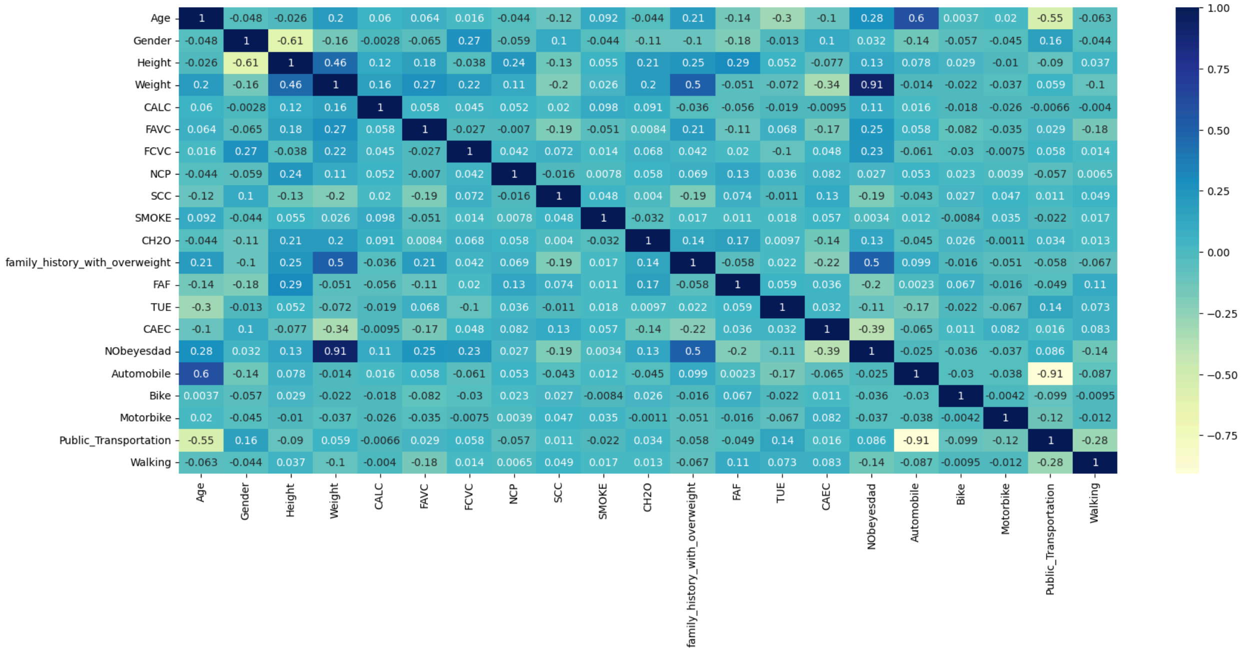
  - **Label Encoding** is used to encode Gender, CALC, FAVC, SCC, SMOKE, family_history_with_overweight, CAEC.
  - **One-Hot Encoding** is used to encode MTRANS.

- ## Correlation Insights:

  - **Weight & Height (0.46 correlation)**
    - Taller individuals generally weigh more, as expected.
    - Both are strong predictors of obesity levels when combined into BMI.

- **Weight & Age (0.20 correlation)**
    - Older individuals show a slight trend of increased weight.
    - Indicates that obesity risk may rise with age.
- **Weight & Dietary/Activity features**
    - Weight has weak positive correlation with vegetable consumption(FCVC, 0.22) and water intake (CH2O, 0.20).
    - Suggests that even with better diet habits, higher weight persists in some individuals, highlighting complex interactions.
- **Height & Lifestyle Factors**
    - Height correlates moderately with NCP (number of main meals, 0.24), CH2O (water consumption, 0.21), and FAF (physical activity, 0.29).
    - Taller individuals tend to report slightly healthier lifestyle habits.
- **Negative Correlations**
    - TUE (time using technology) shows a negative correlation with Age (-0.30) and Weight (-0.07).
    - Indicates younger individuals spend more time on devices, potentially replacing active habits.
    - FAF (physical activity) correlates negatively with Age (-0.14), suggesting reduced activity in older participants.

Correlations heatmap of All Features:



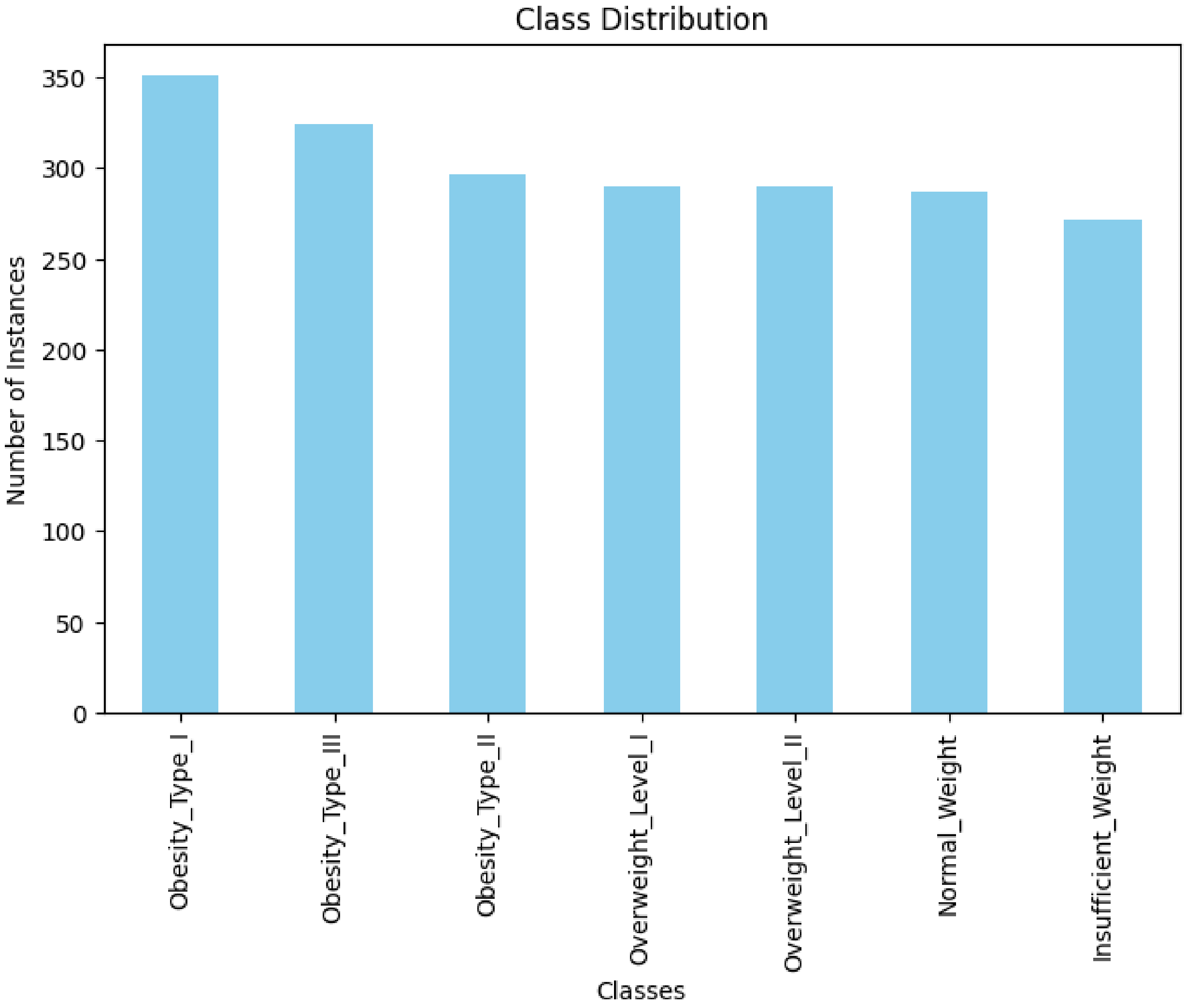
- ## Interpretation of Correlation Test:

  Weight and Height are the most influential features in relation to obesity classification. Lifestyle factors (diet, exercise, screen time) show weaker but meaningful correlations, reinforcing that obesity is multi-factorial. The heatmap suggests that while physical measurements dominate prediction, behavioral factors refine class separation.

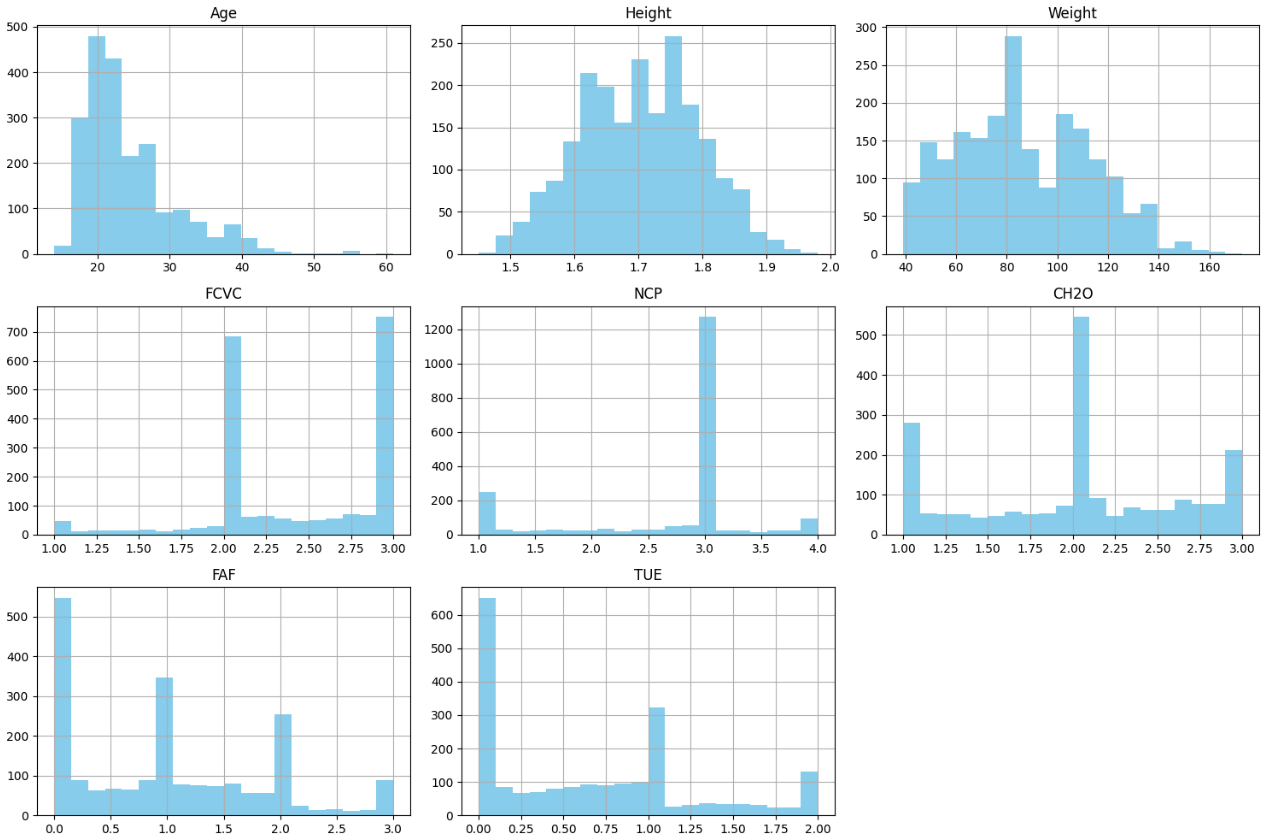
Imbalanced Dataset

The dataset is imbalanced:

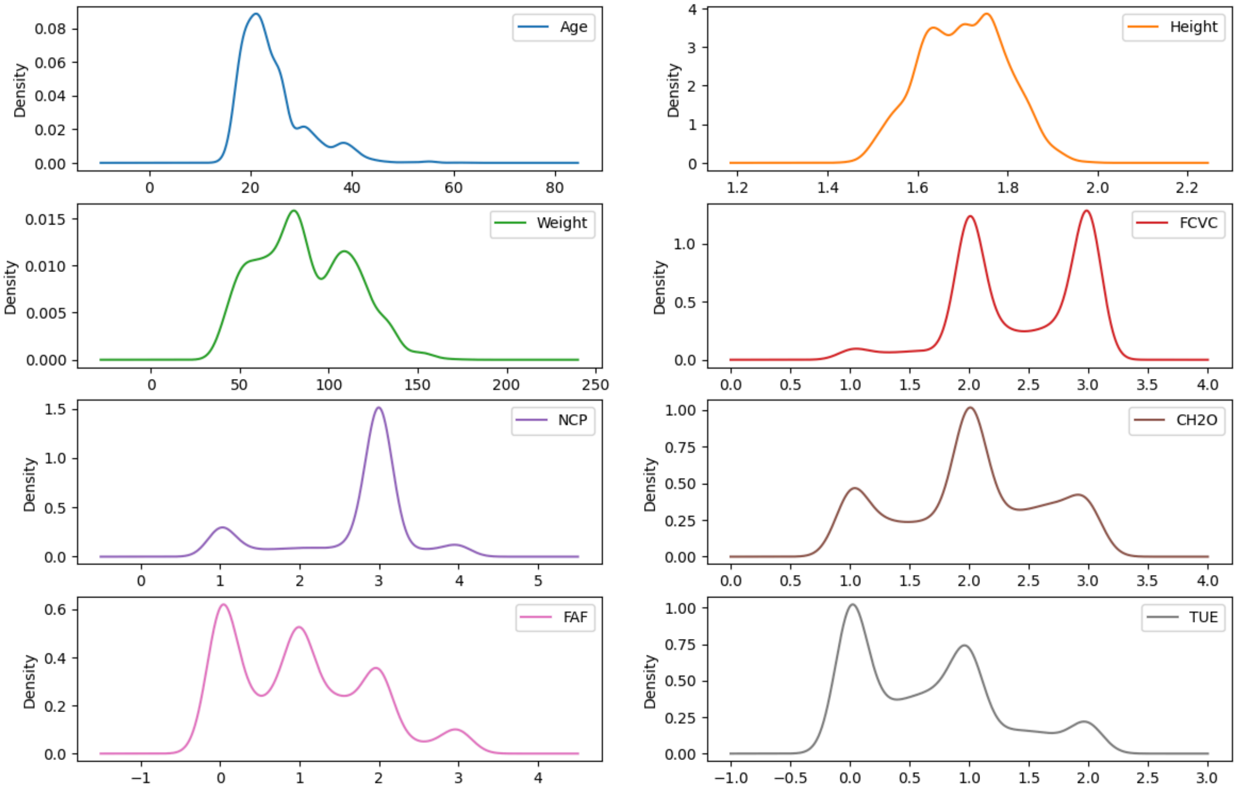
not all unique classes of the target feature “NObeyesdad” have an equal number of instances. Some obesity categories (e.g., Obesity_Type_I with 351 samples) are more frequent, while others (e.g., Insufficient_Weight with 272 samples) are less represented.

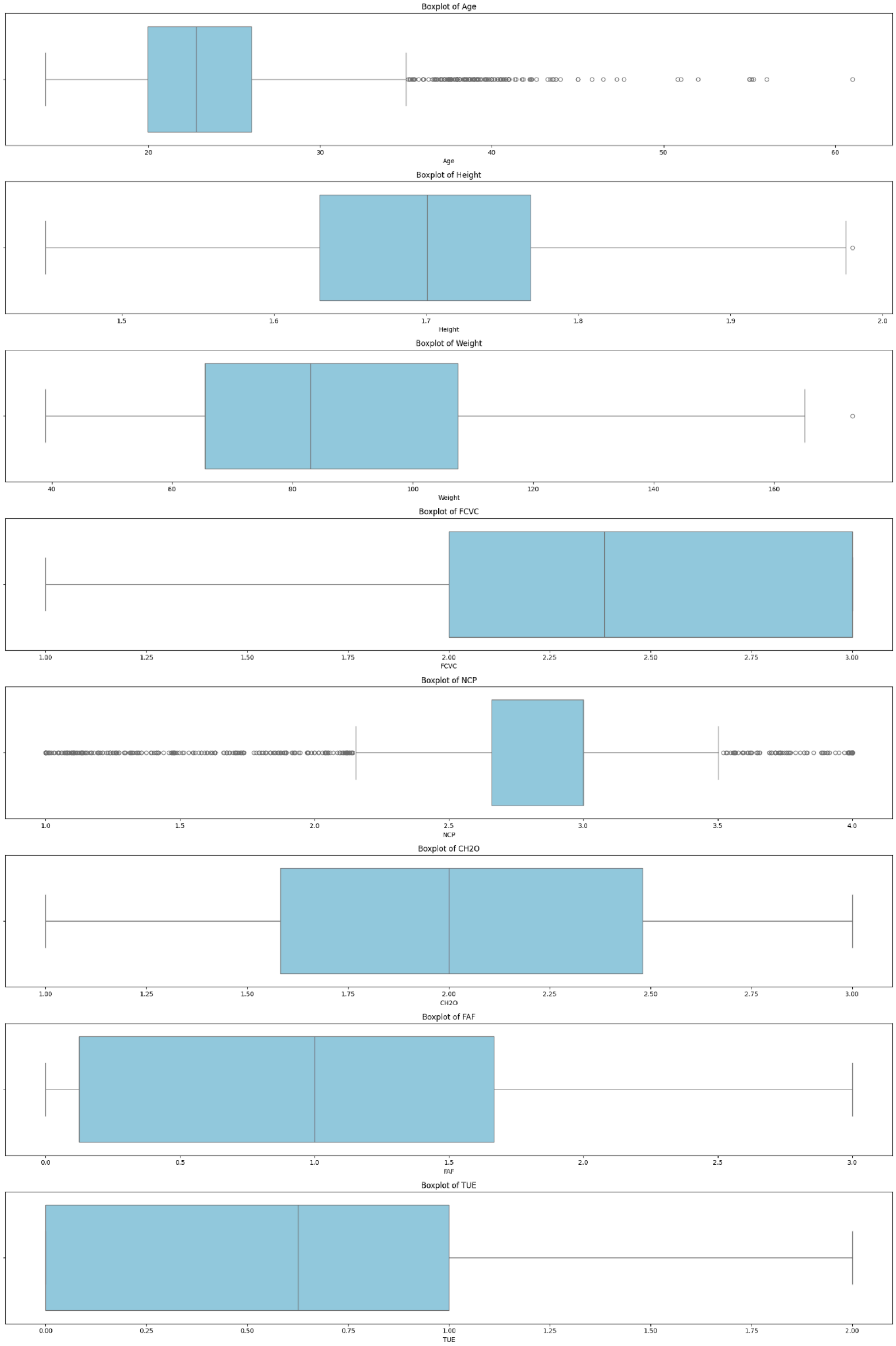


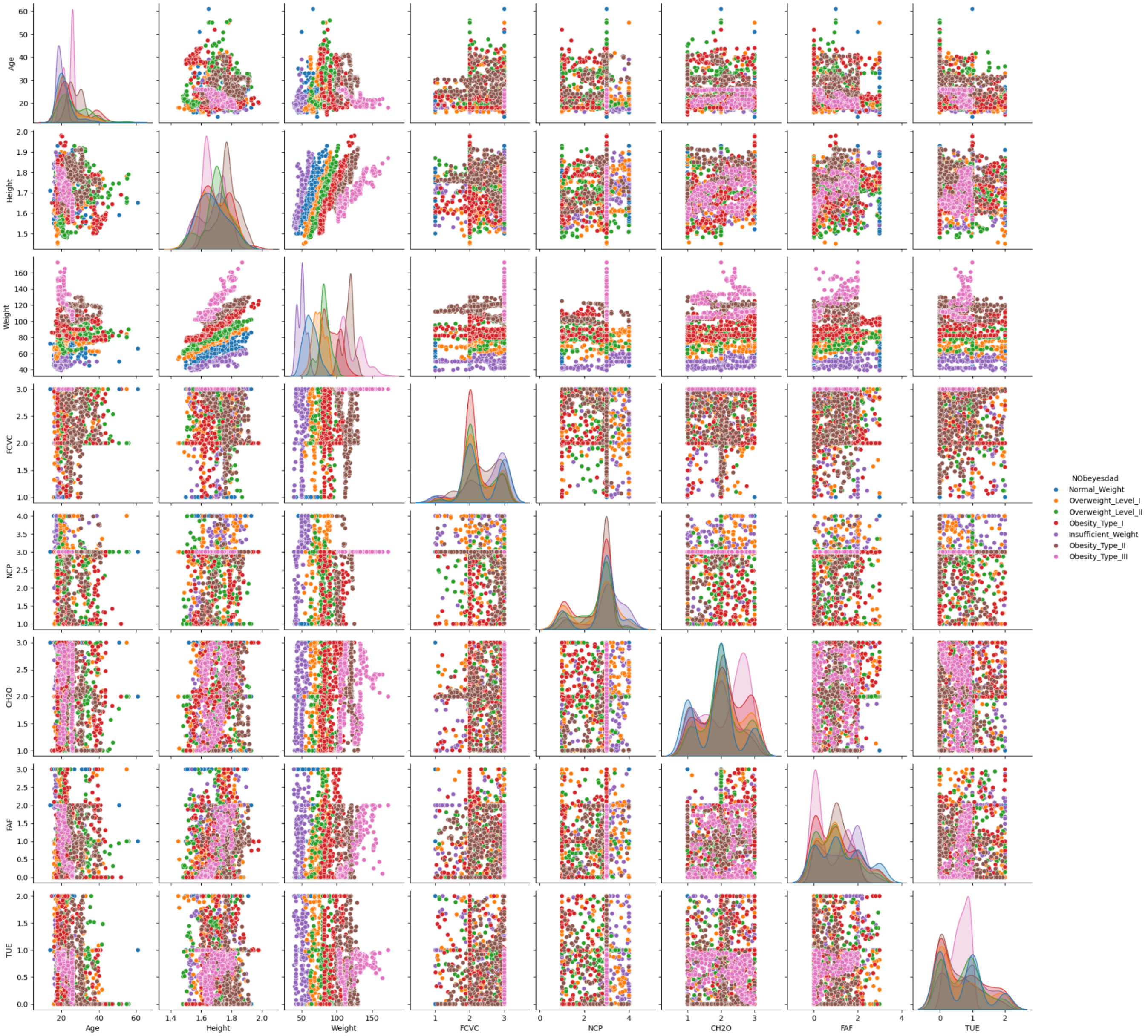
Exploratory Data Analysis (EDA)

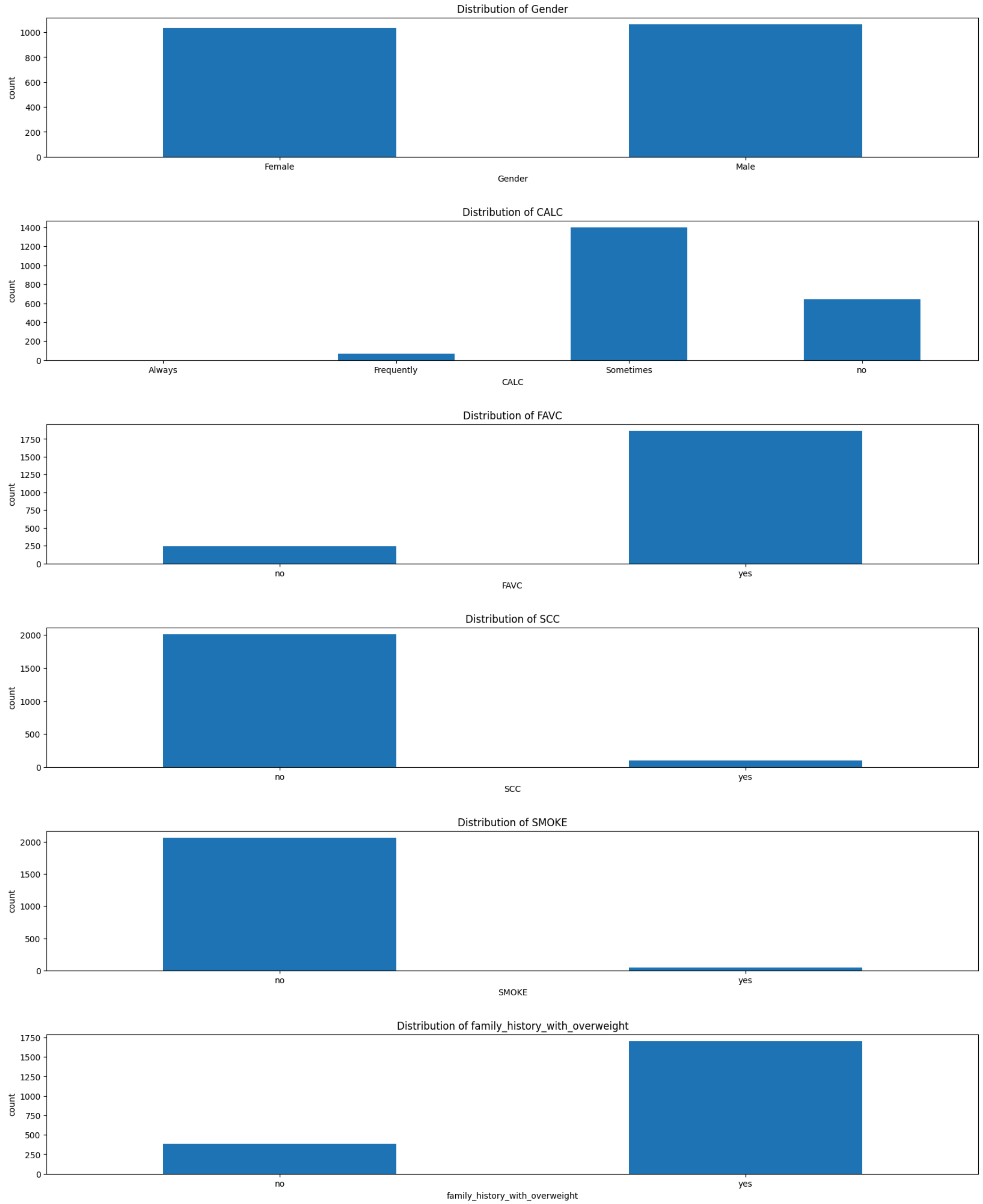


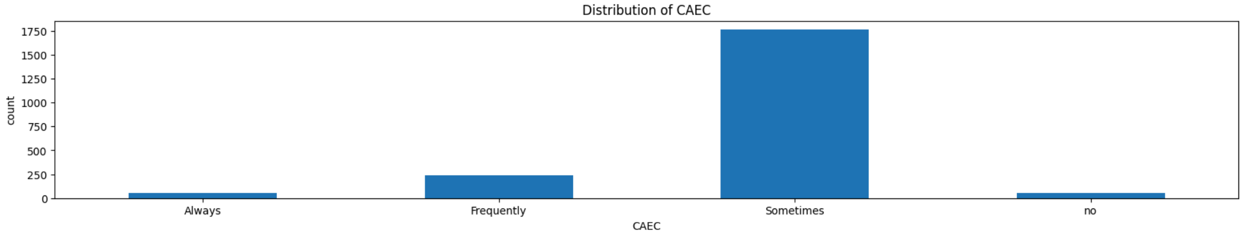
Density plot of Numerical features

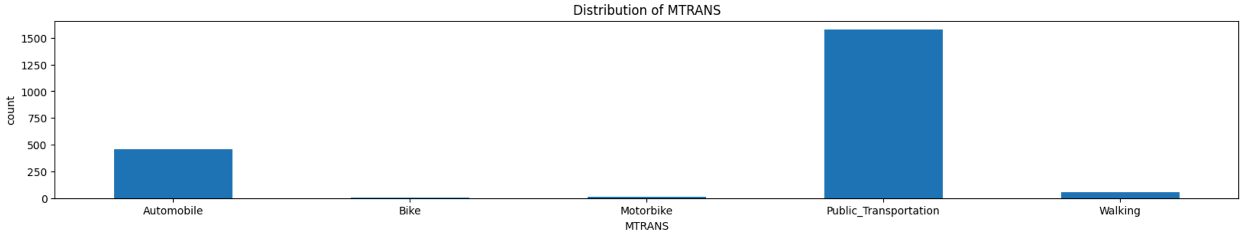


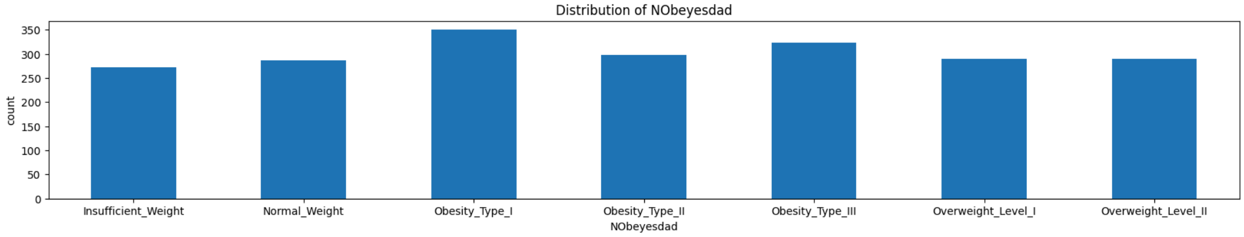

Boxplot of Age
20
30
40
50
60
Age
Boxplot of Height
1.5
1.6
1.7
1.8
1.9
2.0
Height
Boxplot of Weight
40
60
80
100
120
140
160
Weight
Boxplot of FCVC
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
FCVC
Boxplot of NCP
1.0
1.5
2.0
2.5
3.0
3.5
4.0
NCP
Boxplot of CH2O
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
CH2O
Boxplot of FAF
0.0
0.5
1.0
1.5
2.0
2.5
3.0
FAF
Boxplot of TUE
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
TUE


Age
Height
Weight
FCVC
NCP
CH2O
FAF
TUE
NObeyesdad
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III


Distribution of Gender
count
0
200
400
600
800
1000
Female
Male
Gender
Distribution of CALC
count
0
200
400
600
800
1000
1200
1400
Always
Frequently
Sometimes
no
CALC
Distribution of FAVC
count
0
250
500
750
1000
1250
1500
1750
no
yes
FAVC
Distribution of SCC
count
0
500
1000
1500
2000
no
yes
SCC
Distribution of SMOKE
count
0
500
1000
1500
2000
no
yes
SMOKE
Distribution of family_history_with_overweight
count
0
250
500
750
1000
1250
1500
1750
no
yes
family_history_with_overweight


Distribution of CAEC
count
1750
1500
1250
1000
750
500
250
0
Always
Frequently
Sometimes
no
CAEC

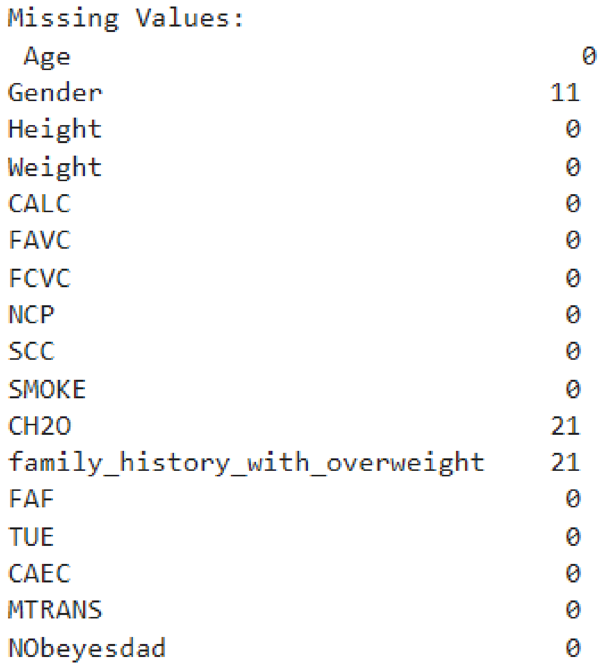

Distribution of MTRANS
count
1500
1250
1000
750
500
250
0
Automobile
Bike
Motorbike
Public_Transportation
Walking
MTRANS


Distribution of NObeyesdad
count
350
300
250
200
150
100
50
0
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II
NObeyesdad

Dataset Pre-Processing

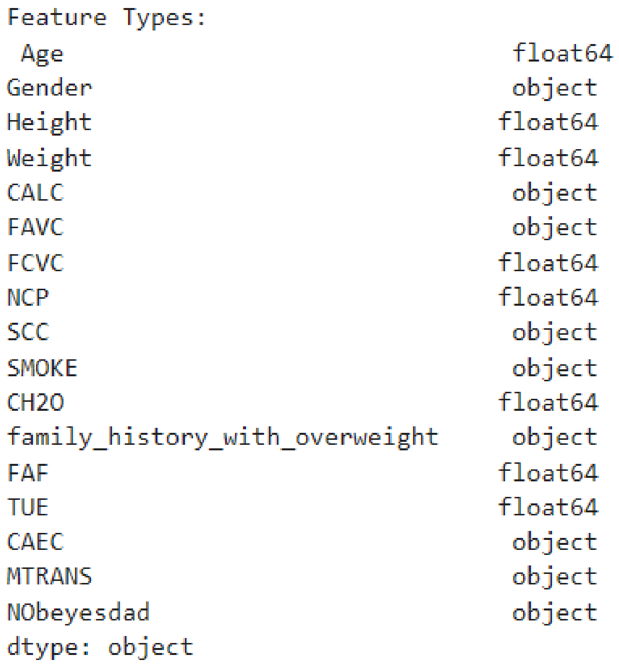
- **Faults:**

Null Values:



```
Missing Values:
 Age                               0
Gender                            11
Height                             0
Weight                             0
CALC                               0
FAVC                               0
FCVC                               0
NCP                                0
SCC                                0
SMOKE                              0
CH2O                              21
family_history_with_overweight    21
FAF                                0
TUE                                0
CAEC                               0
MTRANS                             0
NObeyesdad                         0
```

Categorical Values:



```
Feature Types:
 Age                               float64
Gender                             object
Height                            float64
Weight                            float64
CALC                               object
FAVC                               object
FCVC                              float64
NCP                               float64
SCC                                object
SMOKE                              object
CH2O                              float64
family_history_with_overweight     object
FAF                               float64
TUE                               float64
CAEC                               object
MTRANS                             object
NObeyesdad                         object
dtype: object
```

Feature Scaling

1. **Robust Scaler:** Values with too many outliers.
2. **Standard Scaler:** Values with lower number of outliers.

- ## Solution:
  - **Imputing Values:**
    - **Mode Imputation:** The null values in the features with less outliers are suitable for using mode imputation. Because these values don't have a numeric average. So the most frequent values are the ones chosen as a substitute for the null values. E.g.: Gender, CALC , CAEC , FAVC , SMOKE, SCC , MTRANS, family_history_with_overweight.
    - **Median Imputation:** Median is robust to outliers, so it was applied to skewed numerical features. E.g.: Age, Weight, Height, FCVC, CH2O, FAF, TUE.

  - **Encoding:**
    - **One-Hot Encoding:** The MTRANS feature (mode of transportation) was handled using One-Hot Encoding, since it is a categorical feature with no intrinsic ordering.
    - **Label-Encoding:** Gender, CALC, Sometimes, FAVC, SCC, SMOKE, family_history_with_overweight, CAEC, these features are with no intrinsic values and mostly True False. That is why label encoding is used.
  - **Scaling:**

    Feature scaling is used different scalers depending on the distribution of the data
    - **StandardScaler:** For normally distributed features such as Height, Weight, FCVC, CH2O, FAF, TUE.
    - **RobustScaler:** For skewed or outlier-prone features such asAge, NCP, CALC, CAEC.

Dataset Splitting

To evaluate model performance fairly, the dataset is divided into training and testing subsets.

- **Stratified Split**:
  Since the target variable (NObeyesdad) has 7 classes with imbalanced distribution, we used a stratified split. This ensured that each subset maintained the same class proportion as the original dataset, avoiding bias toward majority classes.
- **Train/Test Ratio**:
  The dataset was split into:
  - 70% Training set – used to train the models.
  - 30% Testing set – used to evaluate the trained models on unseen data.
- **Validation**:
  For the **Neural Network**, we also enabled **early stopping**, which internally uses a portion of the training set as a validation set to prevent overfitting.

This splitting strategy ensured that the models learned from sufficient data while being tested on a representative and unbiased subset.

Model Training and Testing

We trained and evaluated several supervised learning models, along with an unsupervised approach, on the preprocessed obesity dataset. Each model was implemented using scikit-learn.

1. K-Nearest Neighbors (KNN)

- **Implementation:** Tested for values of k from 1 to 30.
- **Best k:** Selected based on maximum accuracy on the test set.
- **Observation:** Performance depended strongly on the choice of k.
- **Result:** Best value of K is 1. Accuracy of the model for the value of K mentioned above: 86.12%

2. Logistic Regression

- **Implementation:** Solver = “lbfgs”, max_iter = 1000.
- **Strengths:** Stable linear baseline.
- **Limitation:** Limited when classes are not linearly separable.
- **Result:** Logistic Regression prediction accuracy: 88.01%

3. Naive Bayes (GaussianNB)

- **Implementation:** Trained on scaled features.
- **Strengths:** Simple and fast.
- **Limitation:** Assumes feature independence, which is not fully valid in this dataset, leading to lower accuracy.
- **Result:** Naive Bayes Prediction accuracy 54.42%

4. Decision Tree

- **Implementation:** Criterion = “entropy”, random_state = 42.
- **Strengths:** Interpretable, captures non-linear relationships.
- **Limitation:** Without pruning, risk of overfitting.
- **Result:** Decision Tree prediction accuracy: 94.01%

5. Neural Network (MLPClassifier)

- **Architecture:** Hidden layers (128, 64), activation = ReLU.
- **Training:** Solver = Adam, learning rate = adaptive, early stopping enabled.
- **Strengths:** Achieved the **highest accuracy**, captured complex feature interactions.
- **Limitation:** Requires more computation and careful tuning.
- **Result:** Test accuracy is 95.27%

6. KMeans Clustering (Unsupervised)

- **Implementation:** Applied KMeans to uncover natural groupings in the dataset.
- **Evaluation:** Measured with silhouette score.

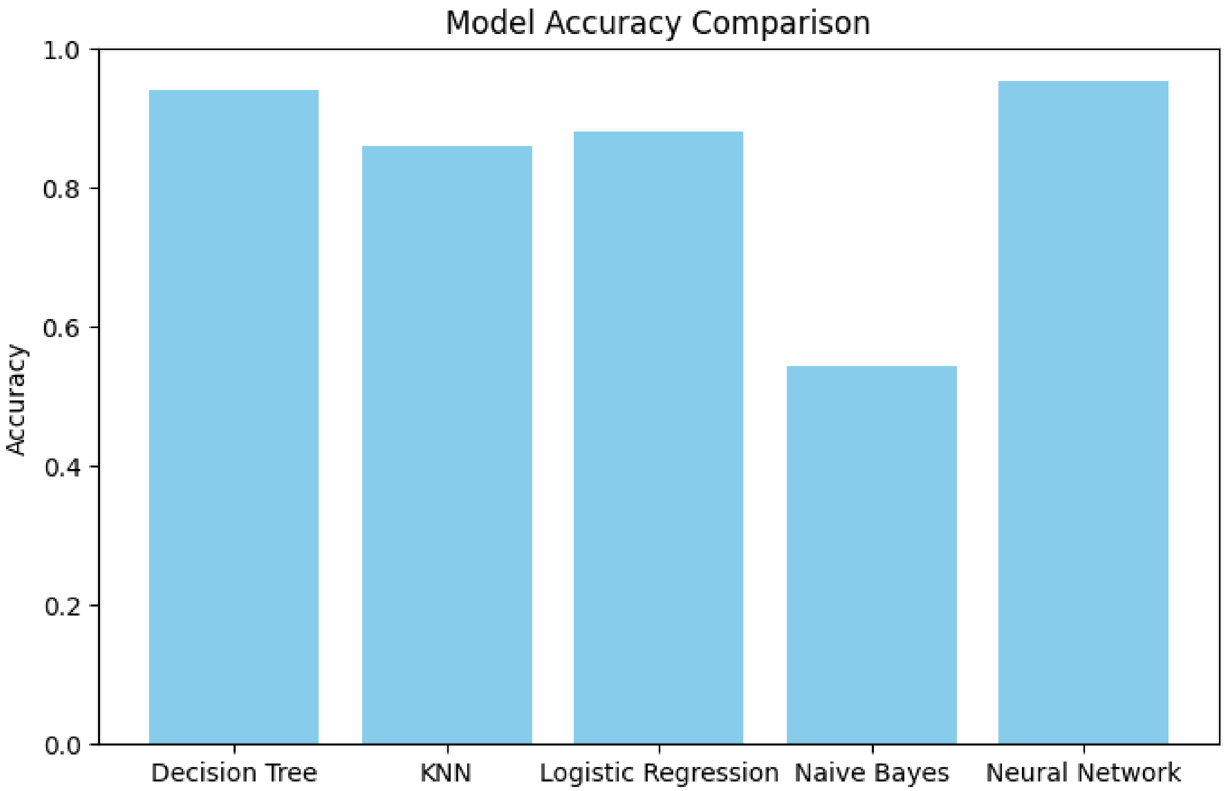
- **Observation:** Some clusters aligned with obesity levels, though unsupervised performance was weaker than supervised methods.

Model Selection / Comparison Analysis

To evaluate the performance of all models, we used accuracy, precision, recall, confusion matrix, and ROC-AUC curves. Since this is a classification problem, regression metrics (R^2, loss) were not applicable.

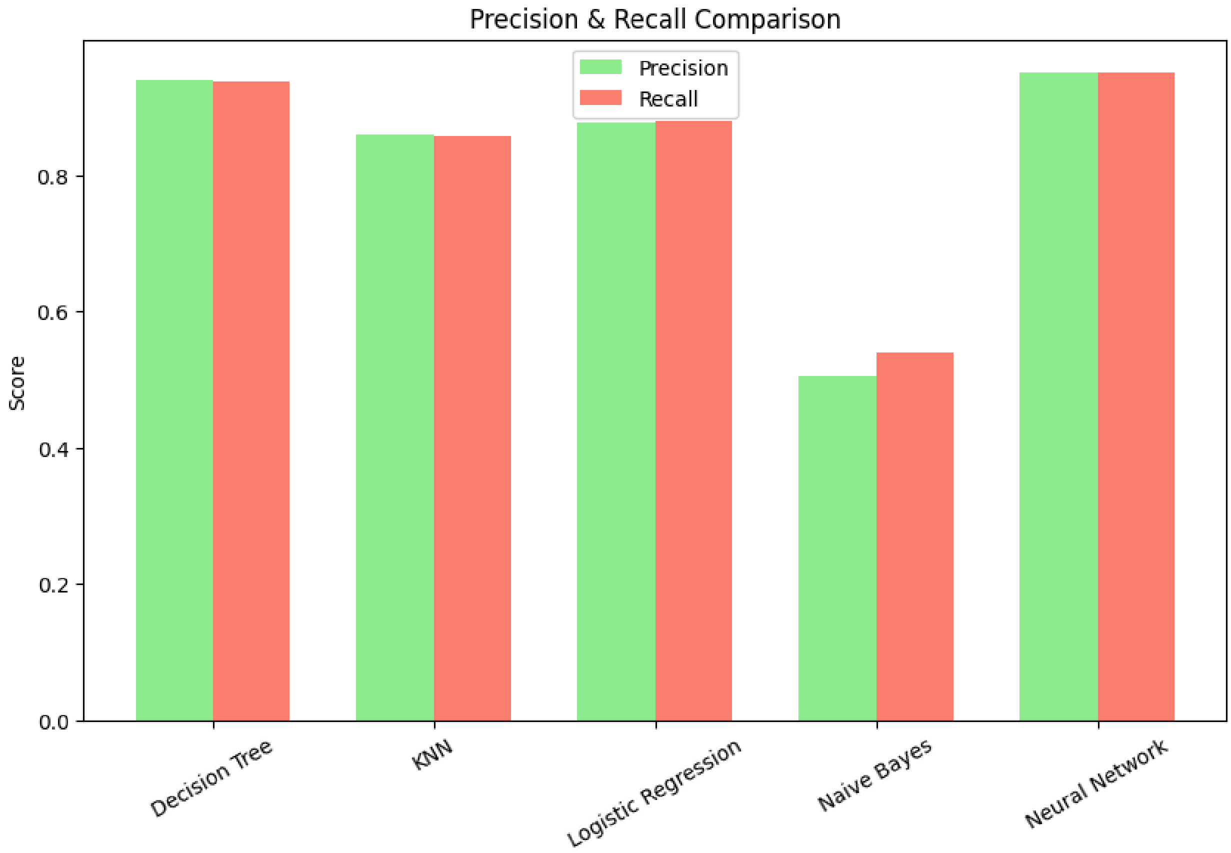
Accuracy Comparison:

A bar chart was plotted comparing the accuracy of each supervised model.



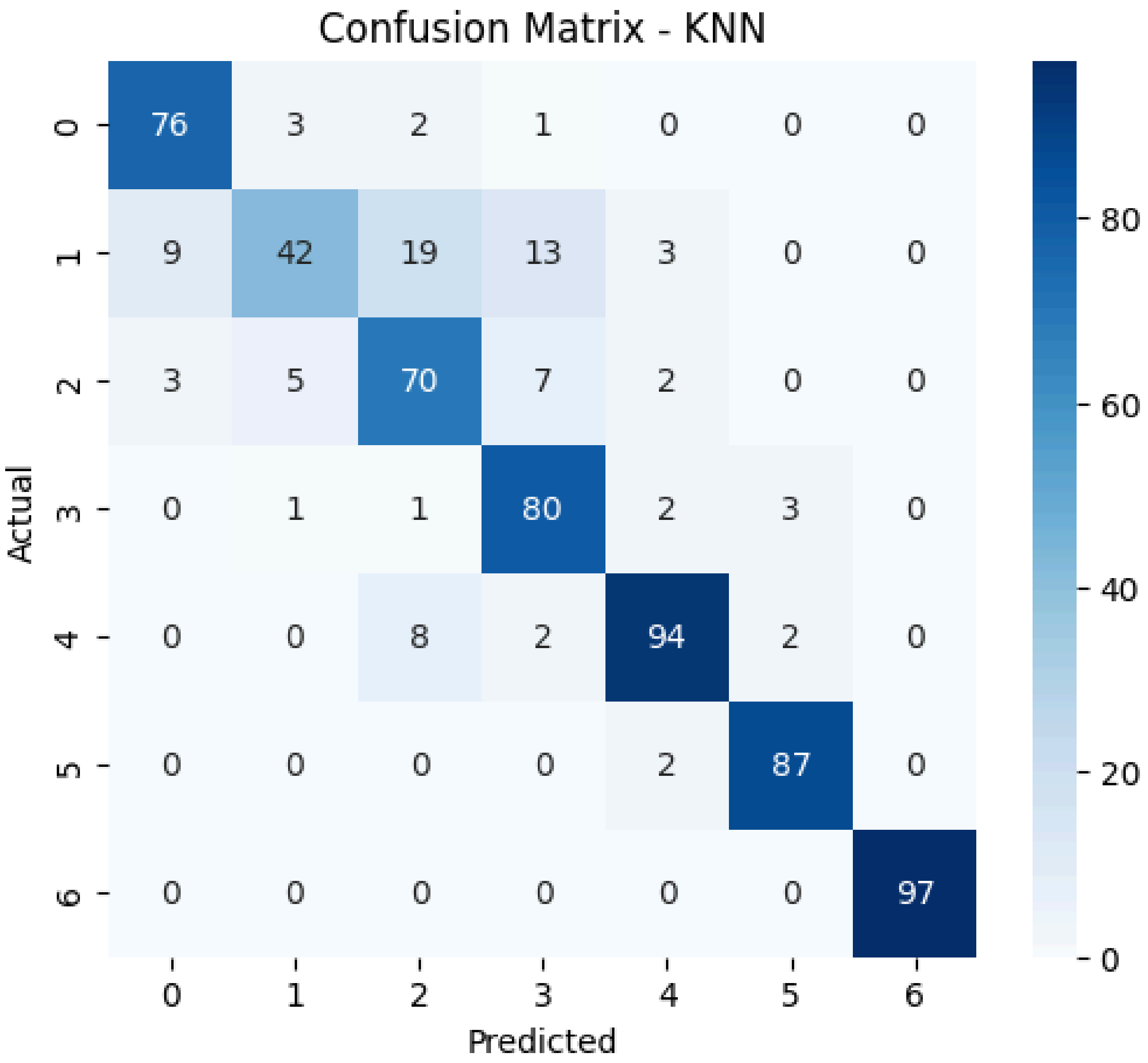
- **Neural Network (MLP)** achieved the highest accuracy, due to its ability to learn complex non-linear patterns.
- **KNN** and Logistic Regression performed moderately well, serving as reliable baselines.
- **Decision Tree** captured non-linear relationships but tended to overfit, resulting in unstable performance.
- **Naive Bayes** was the weakest performer, as it assumes independence among features, which is unrealistic for this dataset.

Precision & Recall:

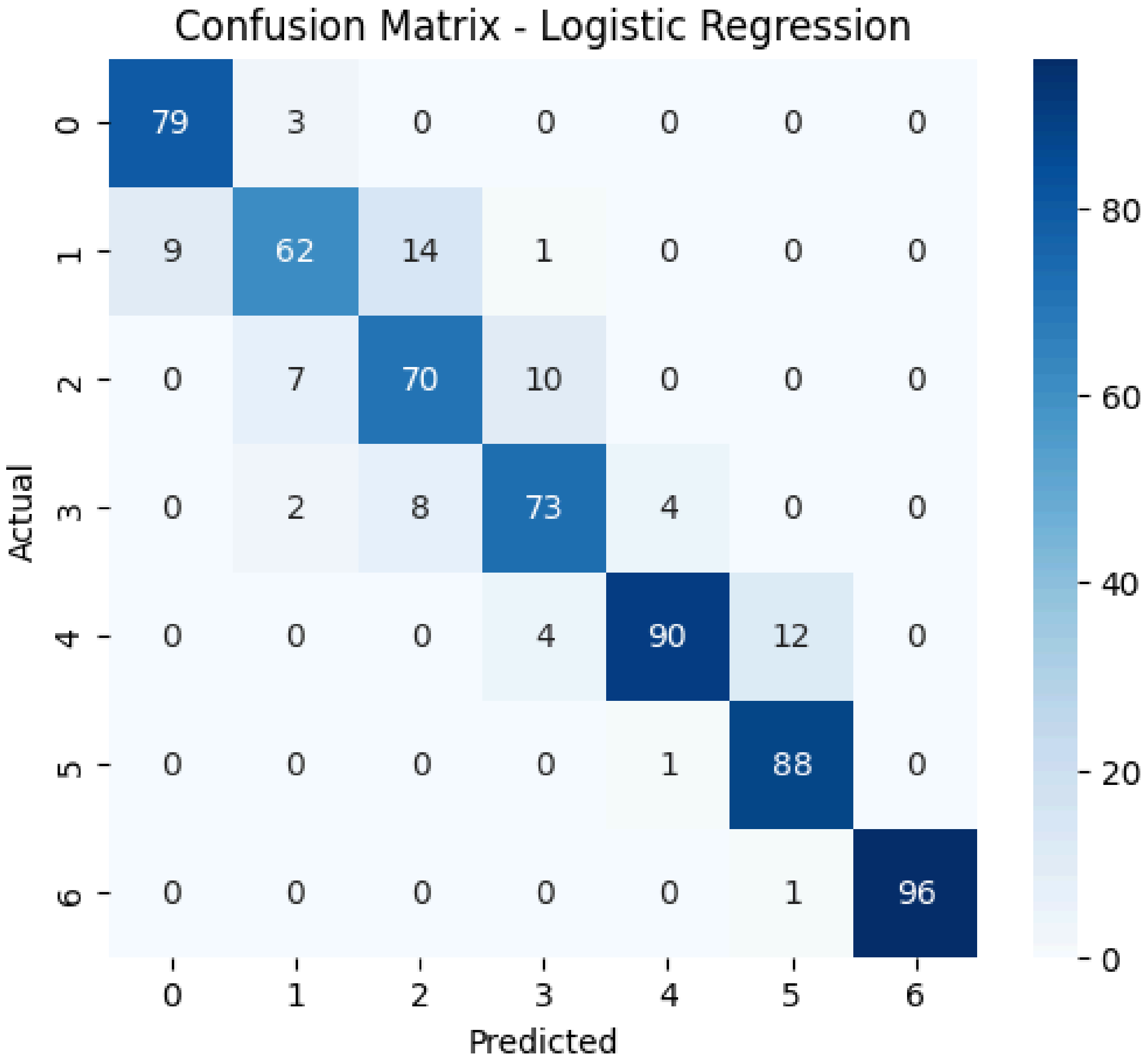


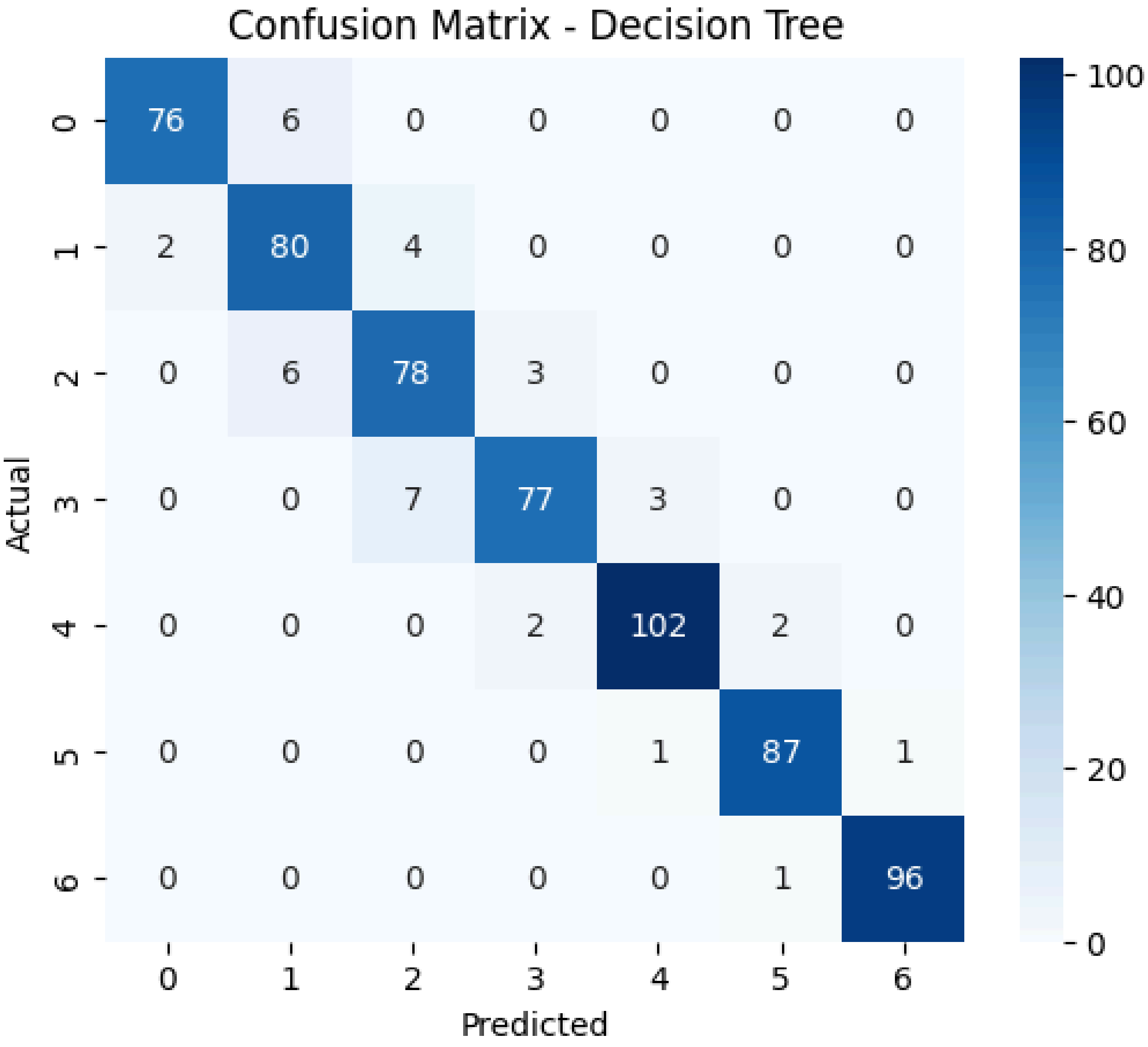
- **Neural Network** achieved strong precision and recall across most obesity classes.
- **KNN** and Logistic Regression showed moderate balance between precision and recall.
- **Naive Bayes** suffered from low recall, especially for minority classes, misclassifying many instances.

Confusion Matrix:



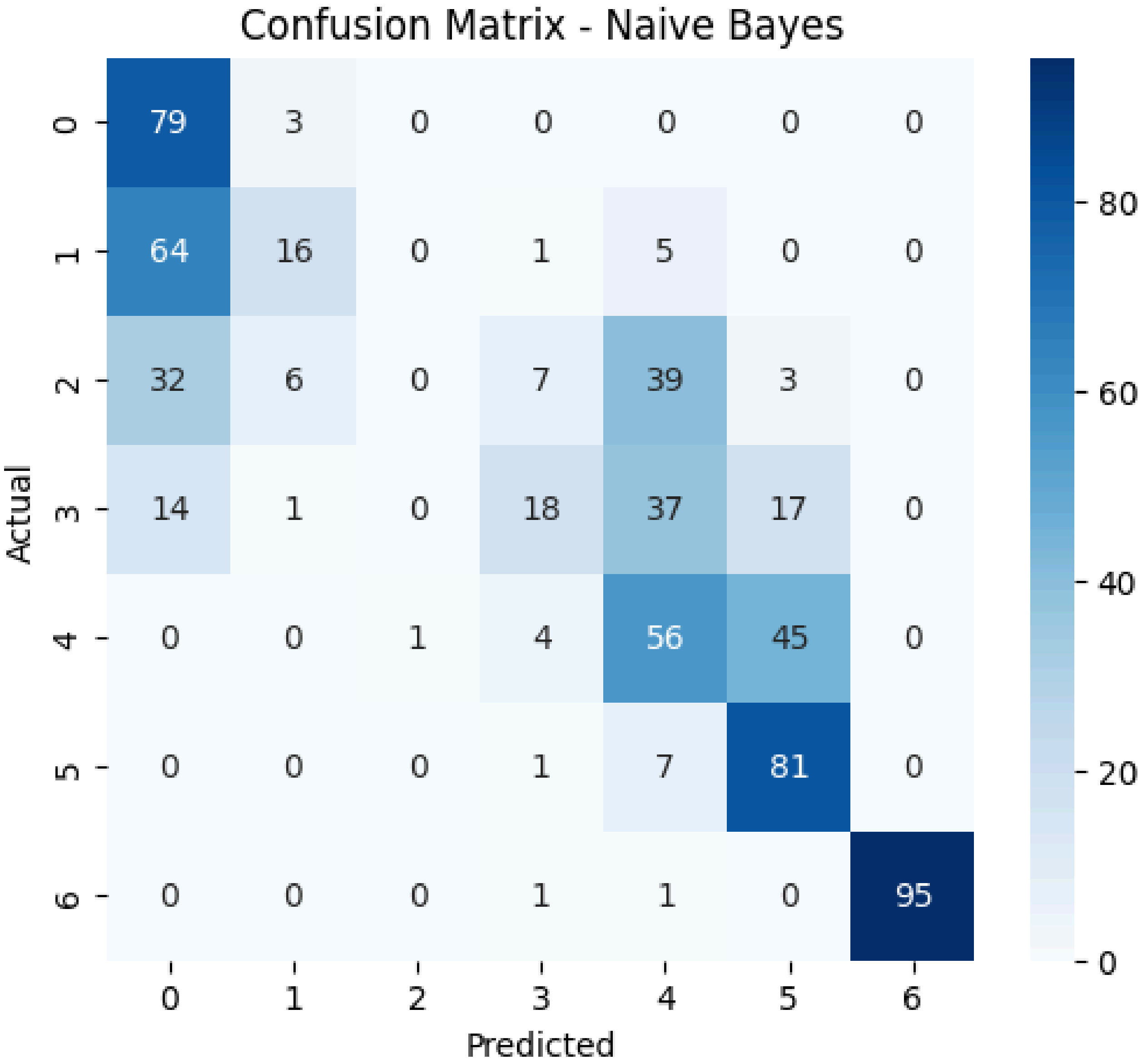
	0	1	2	3	4	5	6
0	76	3	2	1	0	0	0
1	9	42	19	13	3	0	0
2	3	5	70	7	2	0	0
3	0	1	1	80	2	3	0
4	0	0	8	2	94	2	0
5	0	0	0	0	2	87	0
6	0	0	0	0	0	0	97


Confusion Matrix - Logistic Regression
Actual
Predicted
0
1
2
3
4
5
6
79 3 0 0 0 0 0
9 62 14 1 0 0 0
0 7 70 10 0 0 0
0 2 8 73 4 0 0
0 0 0 4 90 12 0
0 0 0 0 1 88 0
0 0 0 0 0 1 96
0
1
2
3
4
5
6
80
60
40
20
0

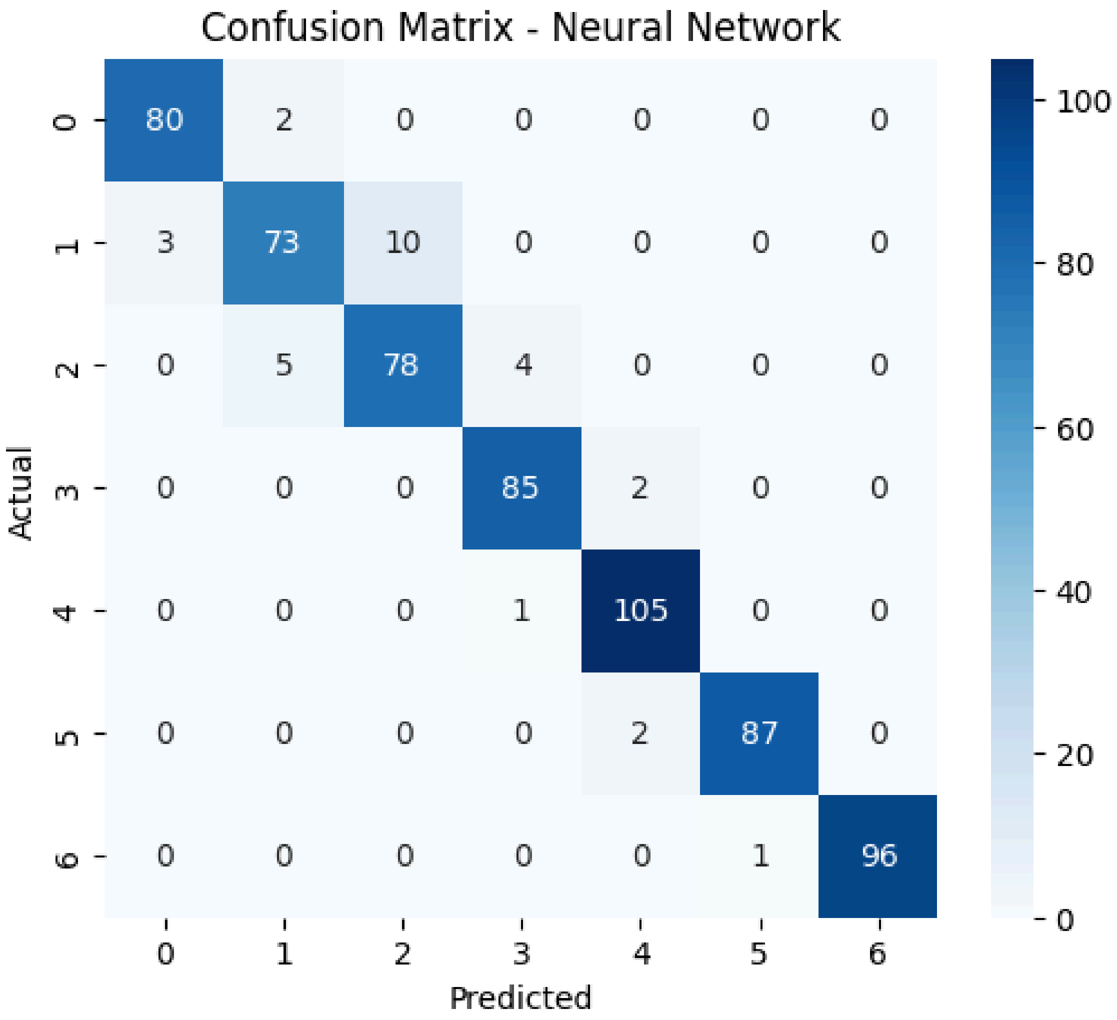


Confusion Matrix - Decision Tree

Actual \ Predicted	0	1	2	3	4	5	6
0	76	6	0	0	0	0	0
1	2	80	4	0	0	0	0
2	0	6	78	3	0	0	0
3	0	0	7	77	3	0	0
4	0	0	0	2	102	2	0
5	0	0	0	0	1	87	1
6	0	0	0	0	0	1	96

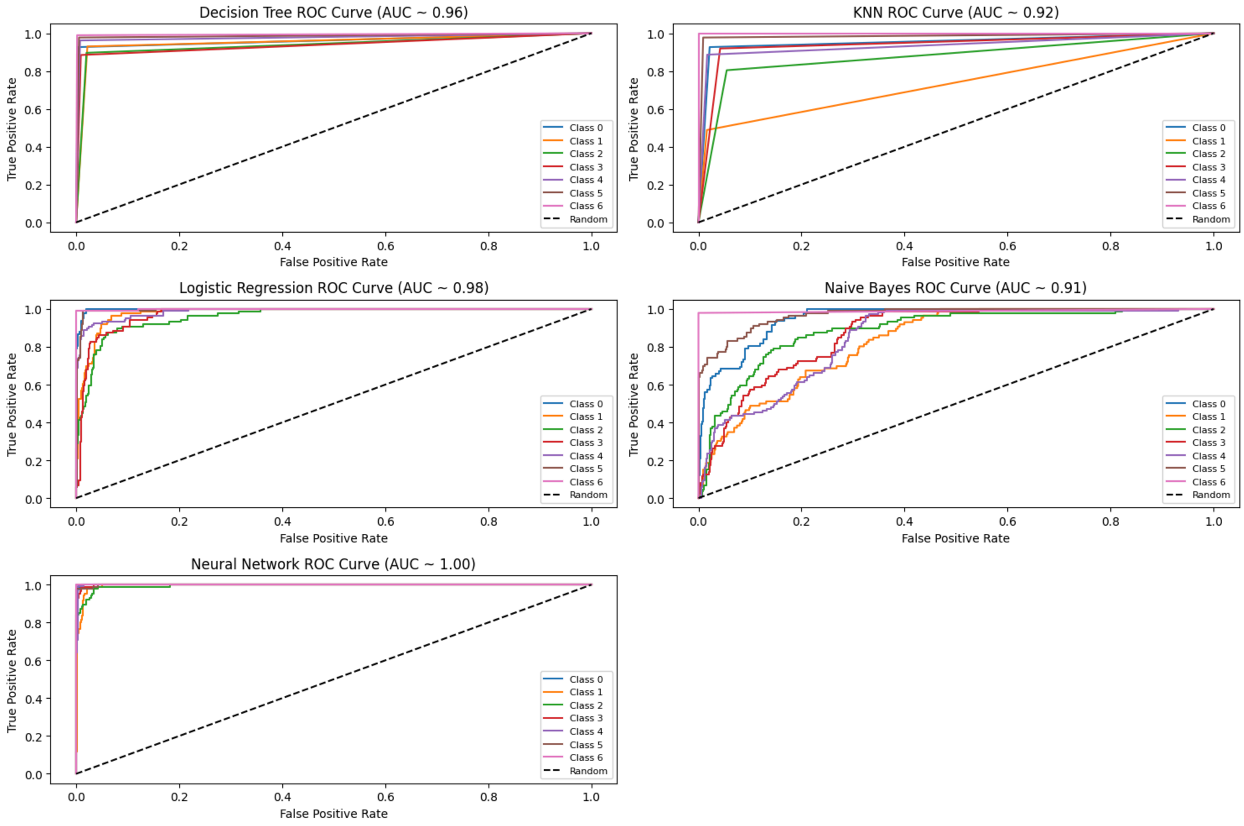


	0	1	2	3	4	5	6
0	79	3	0	0	0	0	0
1	64	16	0	1	5	0	0
2	32	6	0	7	39	3	0
3	14	1	0	18	37	17	0
4	0	0	1	4	56	45	0
5	0	0	0	1	7	81	0
6	0	0	0	1	1	0	95



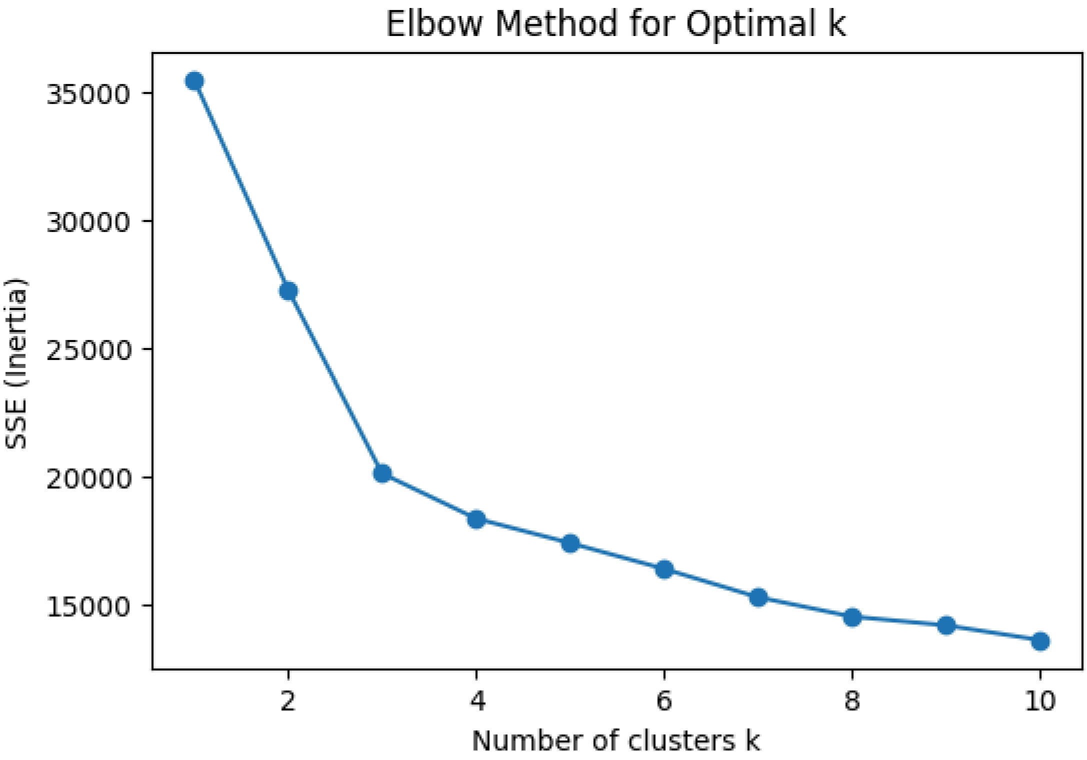
- Neural Network confusion matrix showed the least misclassification, with better class balance.
- KNN misclassified some minority classes into nearby categories (e.g., “Obesity I” confused with “Overweight”).
- Decision Tree overfitted some classes, leading to many correct predictions but also severe misclassifications in other classes.

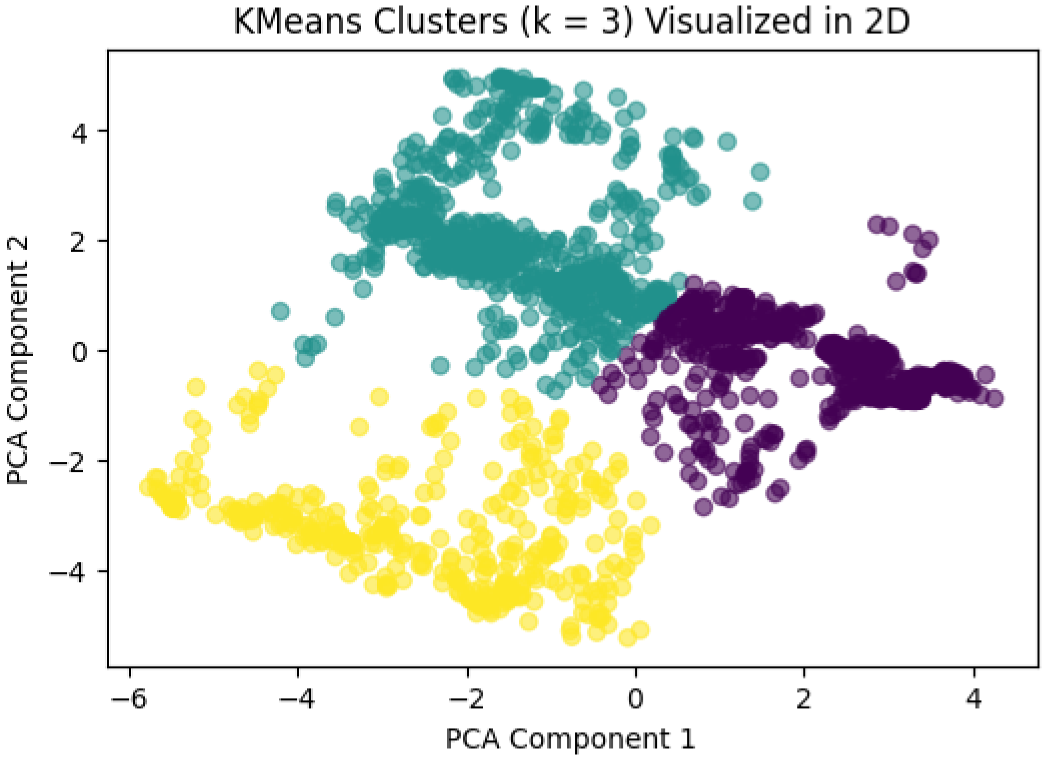
ROC & AUC Score:



- Neural Network had the highest ROC-AUC, showing it is the most discriminative model.
- Logistic Regression also had stable ROC curves.
- Naive Bayes produced weaker ROC curves, reflecting lower predictive power.

KMeans (Unsupervised):





We also applied KMeans clustering to treat the problem in an unsupervised setting.

- **Evaluation:** Silhouette score used to measure cluster quality.
- **Observation:** Some clusters partially aligned with obesity categories, but separation was not as strong as in supervised methods.

Conclusion

From the experiments, the following insights were obtained:

- **Best Model:** The Neural Network (MLPClassifier) consistently outperformed other models, showing robustness across all evaluation metrics.
- **Reliable Alternatives:** KNN and Logistic Regression also gave good performance and can serve as interpretable baseline models.
- **Overfitting Risks:** Decision Tree performed well on training data but overfitted without pruning.
- **Weak Performer:** Naive Bayes was not well-suited due to correlated features and class imbalance.
- **Unsupervised Learning:** KMeans clustering confirmed some grouping patterns but was less effective compared to supervised methods.

Why these results?

- Neural Networks capture complex, non-linear relationships and adapt well with proper scaling and early stopping.
- KNN benefited from feature scaling but struggled with imbalanced classes.
- Logistic Regression is limited by linearity assumptions.
- Naive Bayes underperformed because independence assumptions don’t hold in real-world obesity features.
- Decision Tree overfitted due to unlimited growth without maximum depth or pruning, capturing noise in the dataset.

Challenges faced:

- Handling imbalanced class distribution across 7 obesity categories.
- Choosing correct scalers for different feature distributions.
- Avoiding overfitting in models like Decision Trees and Neural Networks.
- Computational cost of Neural Network training.