Obesity Data Analysis

Team 11
Dharanipriya Ravindran
Yaswanth Kumar Machavarapu
Rishi Kiran Munuswamy

Introduction

- Obesity is a growing public health issue affecting millions worldwide.
- Predicting obesity levels can help healthcare providers take preventive measures.
- Our project aims to classify obesity levels, predict BMI, and group individuals into clusters based on their health and lifestyle factors.

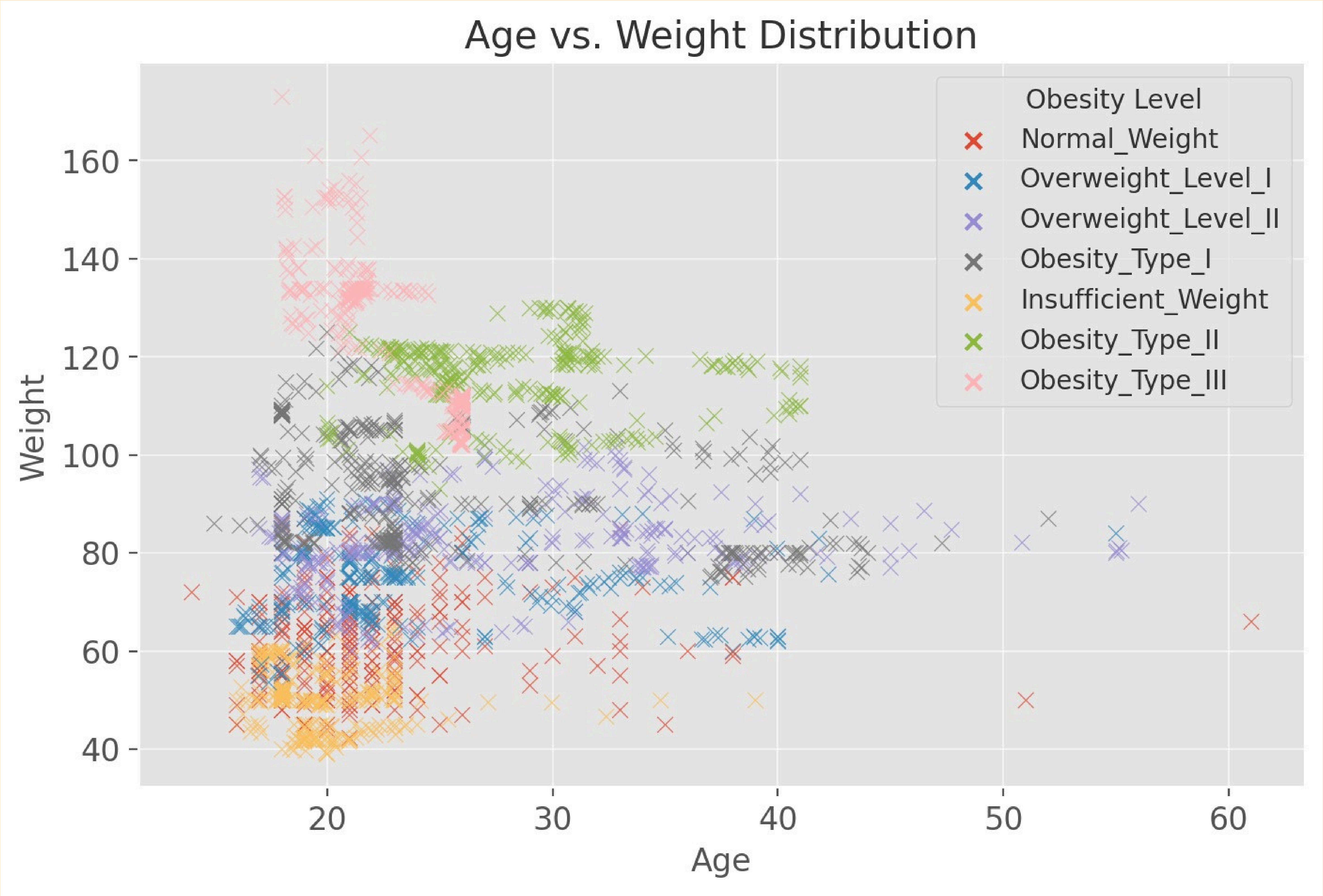
Data Preprocessing & Feature Selection

- Dataset Name: Estimation of Obesity levels
- Source: Public health dataset
- Features: Age, Height, Weight, Food Consumption, Physical Activity, Water Intake, Transportation Mode.

Data Preprocessing & Feature Selection

Data Cleaning:

- Handled missing values.
- Normalized numerical features.
- **Regression Features:** Weight, Age, Height, Physical Activity Frequency (FAF), Number of Meals (NCP).
- **Clustering Features:** Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE.
- **Classification Target:** BMI as the target variable.



Age vs. Weight Distribution
Obesity Level
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III
Weight
160
140
120
100
80
60
40
20
30
40
50
60
Age

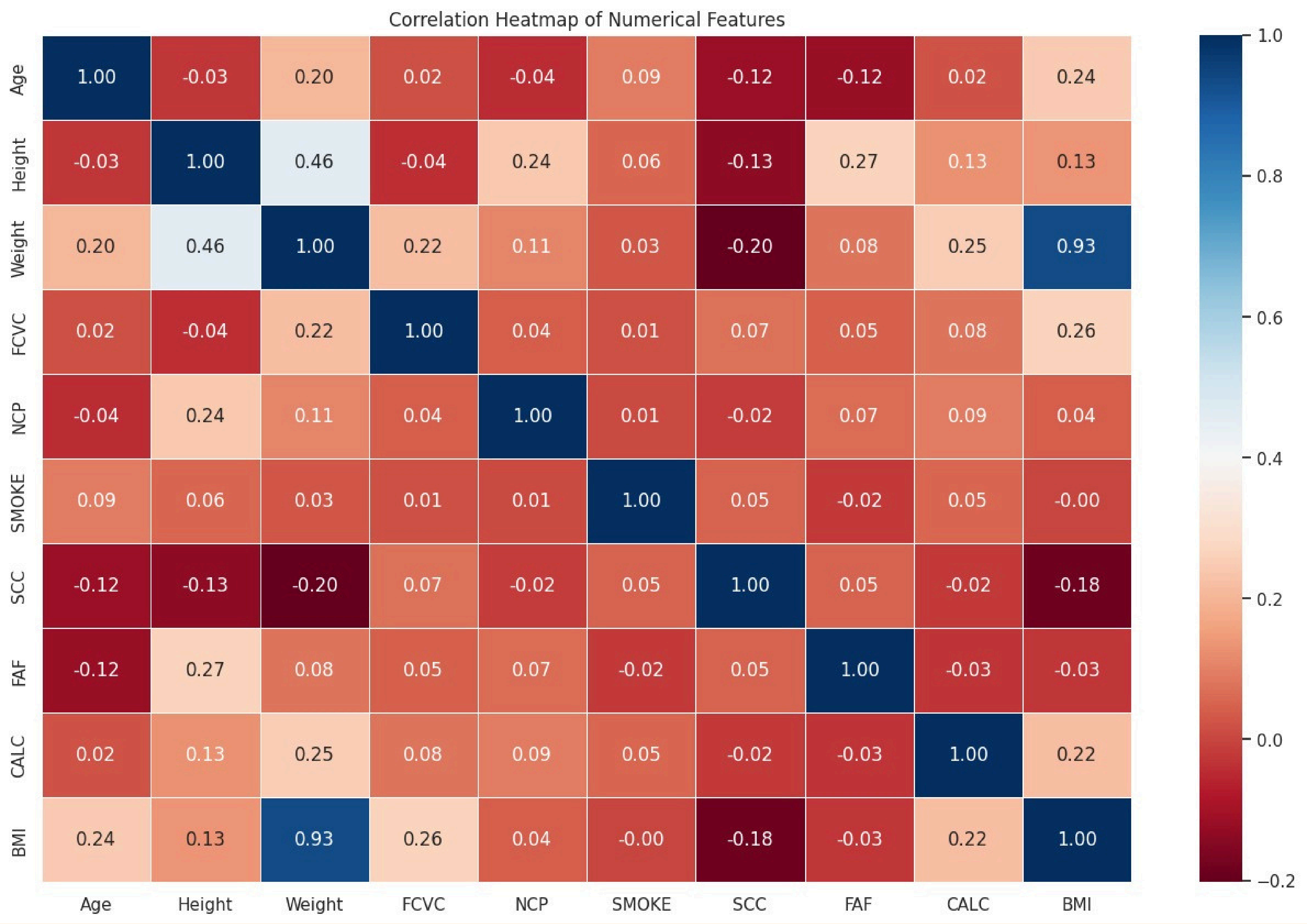
Exploratory Data Analysis (EDA)

Correlation Analysis:

- Weight had the strongest correlation with BMI (0.935).
- Weak correlation for Age and Height.

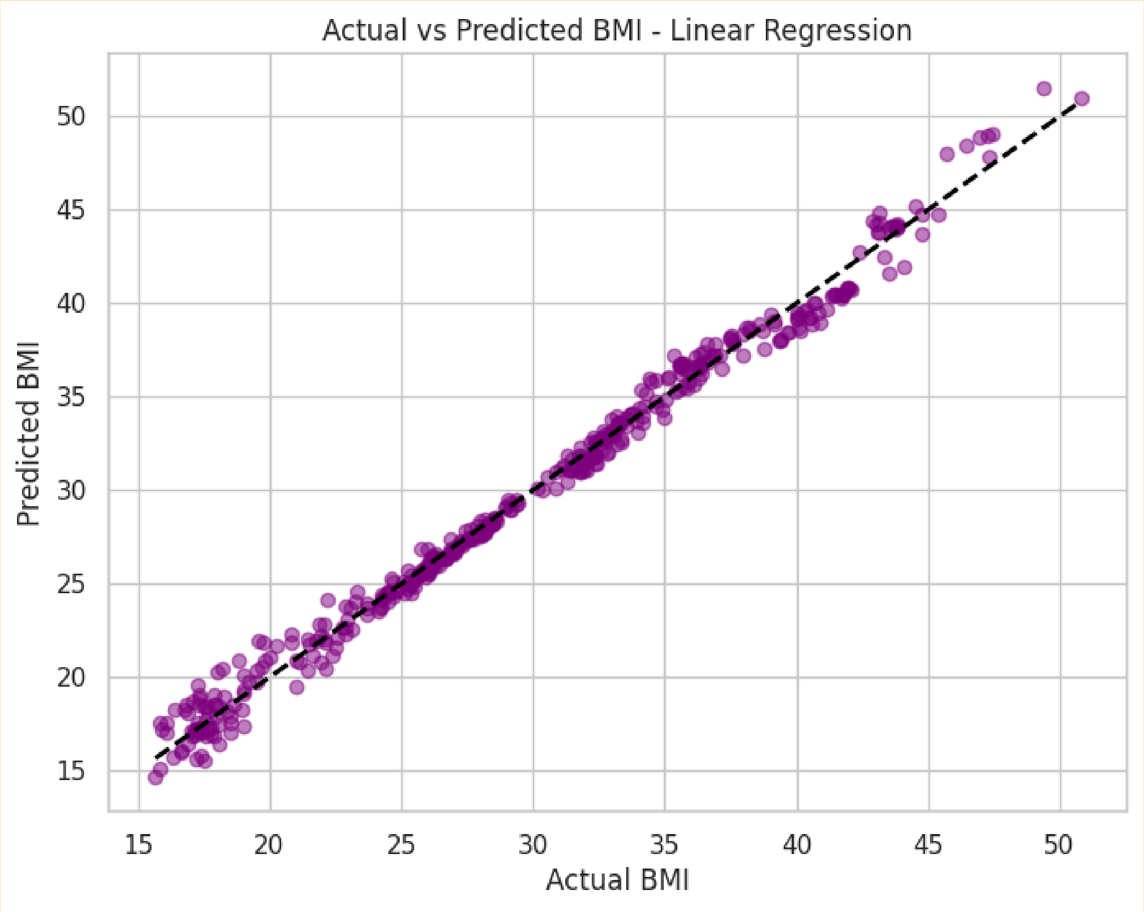
Visualization:

- Histograms, scatter plots, and bar charts to identify obesity trends.


Correlation Heatmap of Numerical Features
Age
Height
Weight
FCVC
NCP
SMOKE
SCC
FAF
CALC
BMI
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Regression

- Objective: Predict BMI based on patient features.
- Models Used:
  - Linear Regression: R^2 Score = 0.76
  - Random Forest Regressor: R^2 Score = 0.89 (Best Model)
- Conclusion: Weight, Height, and Number of Meals were key predictors of BMI.



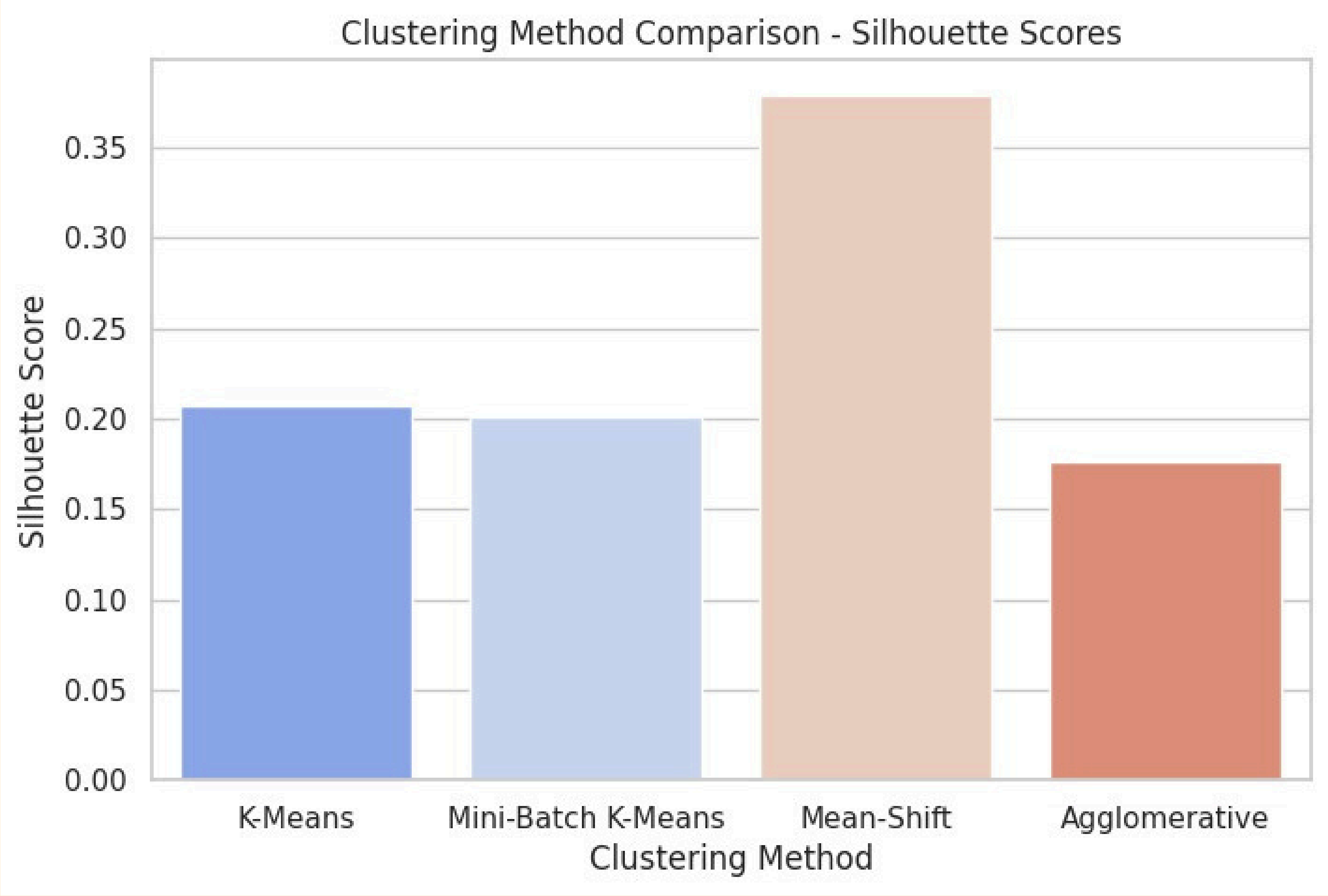
Clustering Analysis

Objective: Identify patterns among obesity groups.

Methods Used:

- K-Means = 0.2068
- Mini-Batch K-Means = 0.2012
- Agglomerative Clustering = 0.1762
- Mean-Shift (Best) = 0.3794

Clustering Analysis



Classification Models and Configurations

Model	Configuration
Logistic Regression	max_iter=1000, random_state=42
K-Nearest Neighbors (KNN)	n_neighbors=5, metric='minkowski', p=2
Decision Tree	max_depth=None, criterion='gini', random_state=42
Random Forest	n_estimators=100, criterion='gini', random_state=42
Support Vector Machine (SVM)	C=1.0, kernel='rbf', gamma='scale', random_state=42
Naive Bayes	GaussianNB()
Neural Networks (MLPClassifier)	hidden_layer_sizes=(100,), max_iter=1000, random_state=42

Classification Analysis

Objective: Classify individuals into obesity categories (Normal Weight, Overweight, Obese, Underweight).

Methods Used:

- Logistic Regression = 82%
- K-Nearest Neighbors (KNN) = 94%
- Decision Trees = 96%
- Support Vector Machine (SVM) = 95%
- Naive Bayes = 90%
- Neural Networks = 94%

Best Model: Decision Trees.

Alternative: Support Naive Bayes.

Key Takeaways

- Regression showed that BMI is influenced by weight, height, and eating habits.
- Mean-Shift Clustering provided the best grouping of obesity risk factors.
- Proper feature selection and preprocessing significantly improved model performance.

Challenges Faced

- Feature Selection: Identifying the most relevant predictors.
- Imbalanced Data: Some obesity categories had fewer samples.
- Model Optimization: Hyperparameter tuning improved results.
- Overfitting: Decision Tree achieved 96% accuracy but required careful validation.

Future Directions

- Improve Feature Engineering: Include additional lifestyle factors.
- Balance Class Distribution: Apply oversampling/undersampling.
- Optimize Model Performance: Implement deep learning techniques.
- Deploy as a Decision Support System for Healthcare Providers.

Thank you!