CS770 Machine Learning

Assignment 2 BMI Classification Based on Gender, Height, and Weight Using Machine Learning.

19/03/2025

Submitted by: Asmae Mouradi

Table of Contents

Abstract

Body Mass Index (BMI) classification is a fundamental tool in health assessments, aiding in the identification of weight-related health risks. In this assignment, I will apply machine learning techniques to classify BMI categories based on gender, height, and weight, leveraging various preprocessing and modeling strategies. Initially, data exploration and cleanup are conducted to ensure data integrity, followed by exploratory data analysis (EDA) to visualize BMI distributions and gender-specific variations. To address class imbalance, I apply Synthetic Minority Over-sampling Technique (SMOTE) to balance the training set, and feature normalization is applied for improved model performance. Classification models including Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) are trained and evaluated using accuracy, precision, recall, and F1-score. Additionally, gender-specific modeling is explored to assess its impact on prediction performance compared to a general model.

Introduction

Body Mass Index (BMI) is a key health indicator used to classify individuals into weight categories based on their height and weight. Accurate BMI classification can be beneficial in health monitoring and risk assessment. However, BMI datasets often suffer from imbalanced distributions, which can negatively impact machine learning model performance. In this assignment, I explore various data preprocessing, modeling, and evaluation techniques to improve BMI classification accuracy.

The assignment begins with data exploration and cleanup, where we analyze dataset characteristics and handle missing values or anomalies. This is followed by an exploratory data analysis (EDA) phase, where I visualize the distribution of BMI categories and examine the relationships between gender, height, weight, and BMI classification.

To address class imbalance, I implement data preprocessing techniques such as SMOTE (Synthetic Minority Over-sampling Technique), undersampling, and normalization of numerical features to enhance model performance. The dataset is then split into training and testing sets, ensuring a balanced gender representation.

I train and evaluate three classification models Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) to predict BMI categories. A key aspect of this study is gender-specific modeling, where separate models are trained

for male and female data to assess whether this approach improves predictive accuracy compared to a general model trained on the combined dataset.

The models are evaluated using accuracy, precision, recall, and F1-score, allowing us to compare gender-specific vs. general models in terms of predictive performance. Additionally, hyperparameter tuning using GridSearchCV is performed to optimize model performance.

Finally, I conduct a gender-based prediction analysis to explore how gender differences influence BMI predictions and discuss their implications.

Here is the github repository for my assignment, I did it in jupyter notebook:

https://github.com/asmaemou/assignment_2_ML

Methods

Data Exploration and Cleanup:

1. Initial Data Exploration

To begin, I explored the dataset using df.info(), df.head(), and df.describe() to understand its structure and characteristics. The dataset consists of 500 entries with the following columns:

- Gender: Categorical variable representing whether an individual is Male or Female.
- Height: A numerical feature indicating an individual's height in centimeters.
- Weight: A numerical feature representing weight in kilograms.
- Index: The BMI category (target variable), ranging from 0 to 5, where:
    - 0: Extremely Underweight
    - 1: Underweight
    - 2: Normal Weight
    - 3: Overweight
    - 4: Obese
    - 5: Extremely Obese

2. Handling Missing Values

I checked for missing values using df.isnull().sum(), and fortunately, there were no missing values in the dataset. This meant I didn't have to impute or remove any rows due to missing data.

3. Handling Anomalies & Outliers

To ensure data consistency, I analyzed the summary statistics using df.describe(). The values for height and weight appeared reasonable (Height: 130 cm to 199 cm, Weight: 50 kg to 160 kg). However, I still needed to confirm the presence of any potential outliers using boxplots in the Exploratory Data Analysis (EDA) section.

4. Data Cleaning Steps I Took

- Converted categorical variables: Since machine learning models require numerical inputs, I encoded the "Gender" column into a numerical format (Male = 0, Female = 1).
- Mapped BMI categories: To make the BMI classifications easier to interpret, I created a mapping from Index to readable category names such as 2 for Normal Weight.

Exploratory Data Analysis (EDA)

1. Distribution of BMI Categories

To understand how BMI categories are distributed, I plotted a count plot showing the frequency of each category. The distribution reveals an imbalance in BMI classes, with Extremely Obese (5) and Obese (4) categories having the highest counts, while Extremely Underweight (0) and Underweight (1) categories are underrepresented. This confirms the need for resampling techniques, such as SMOTE, to balance the dataset for machine learning models.

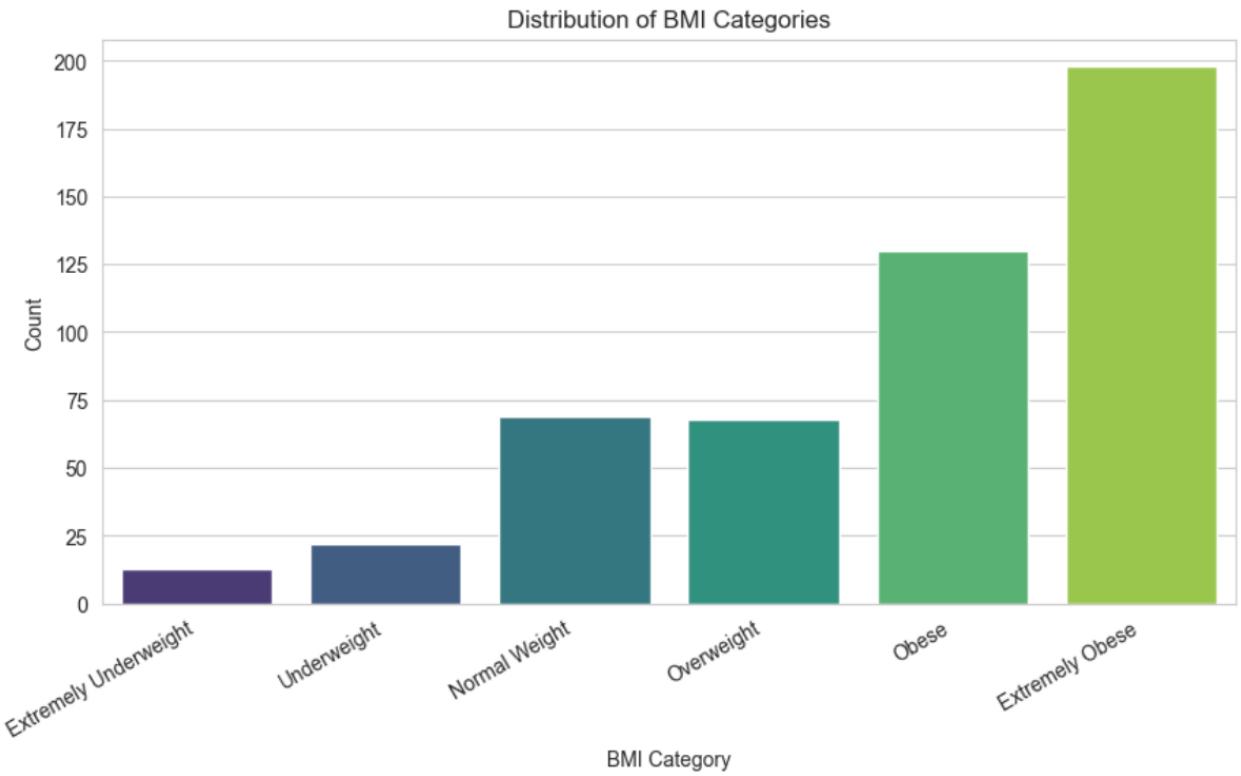

Figure 1. Distribution of BMI Categories

2. Relationship Between Gender, Height, Weight, and BMI Category

I analyzed how gender, height, and weight influence BMI categories using different visualizations:

- Height and Weight Distributions by Gender:
  - The height distribution shows a similar range for both genders, with males tending to have slightly higher median heights than females.
  - The weight distribution suggests that males and females have comparable weight ranges, though males appear to have more representation in higher weight values.
- BMI Category Distribution by Gender:
  - Males and females are not evenly distributed across BMI categories.
  - More males fall into higher BMI categories (Obese and Extremely Obese), while females have a more even spread across the categories.

- There is a relatively low representation of Extremely Underweight and Underweight individuals in both genders.

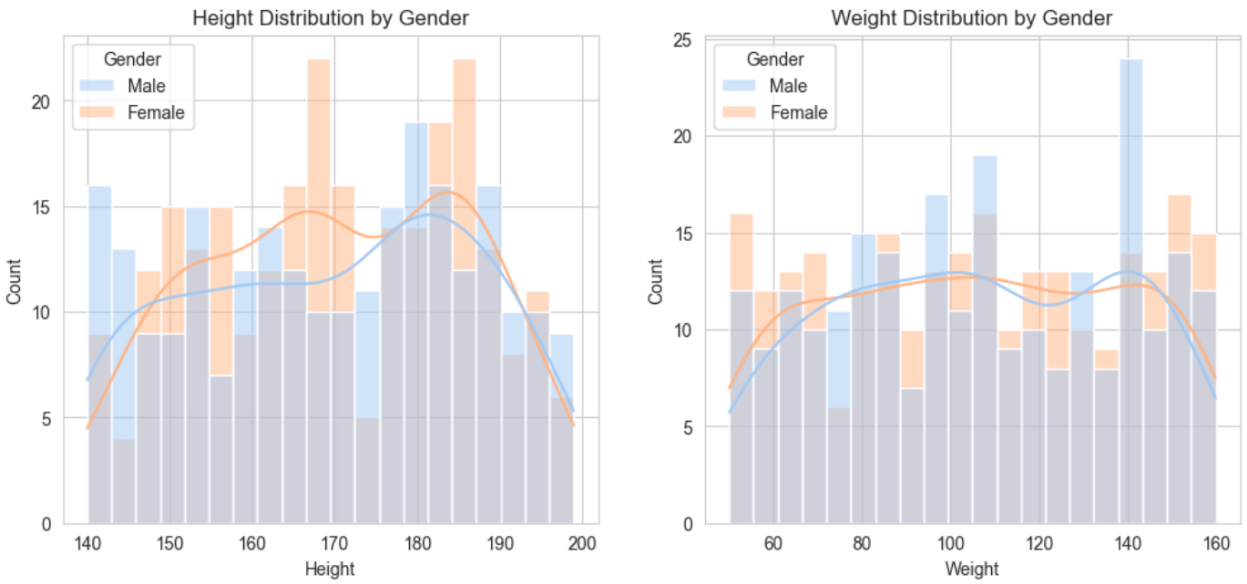

Figure 2. Relationship Between Gender, Height, Weight, and BMI Category

3. BMI Category Distribution by Gender

The bar chart illustrates the distribution of BMI categories for males and females.

- The Obese (4) and Extremely Obese (5) categories have the highest representation for both genders.
- Females tend to be more evenly distributed across BMI categories, whereas males have a slightly higher concentration in the higher BMI categories.
- The Underweight and Extremely Underweight categories have relatively fewer individuals in both genders, indicating class imbalance that may impact model training.

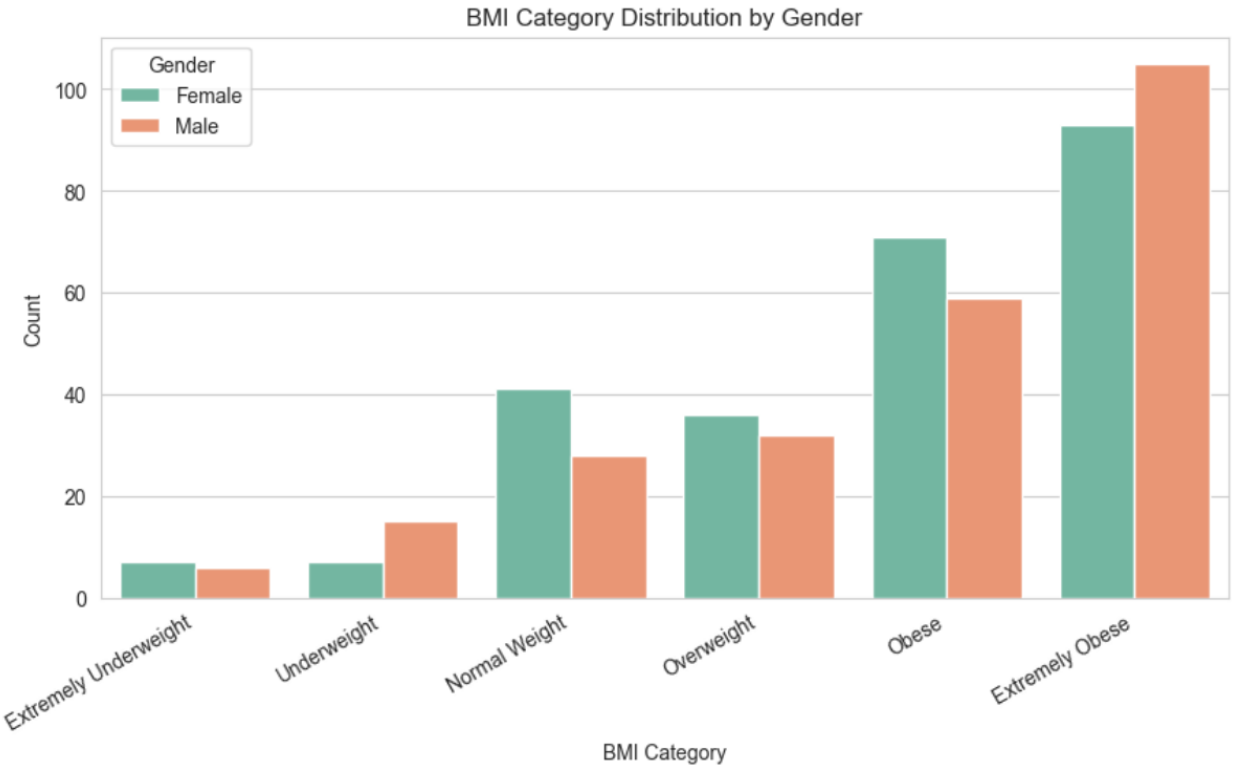

Figure 3. BMI Category Distribution by Gender

4. Boxplot of Height and Weight by Gender

The boxplots compare the distribution of height and weight between males and females.

- Height: Males tend to be taller on average than females, with a slightly higher median height.
- Weight: While there is overlap, males generally exhibit a wider range of weight values compared to females.
- Both distributions have outliers, indicating that some individuals have extreme height or weight values

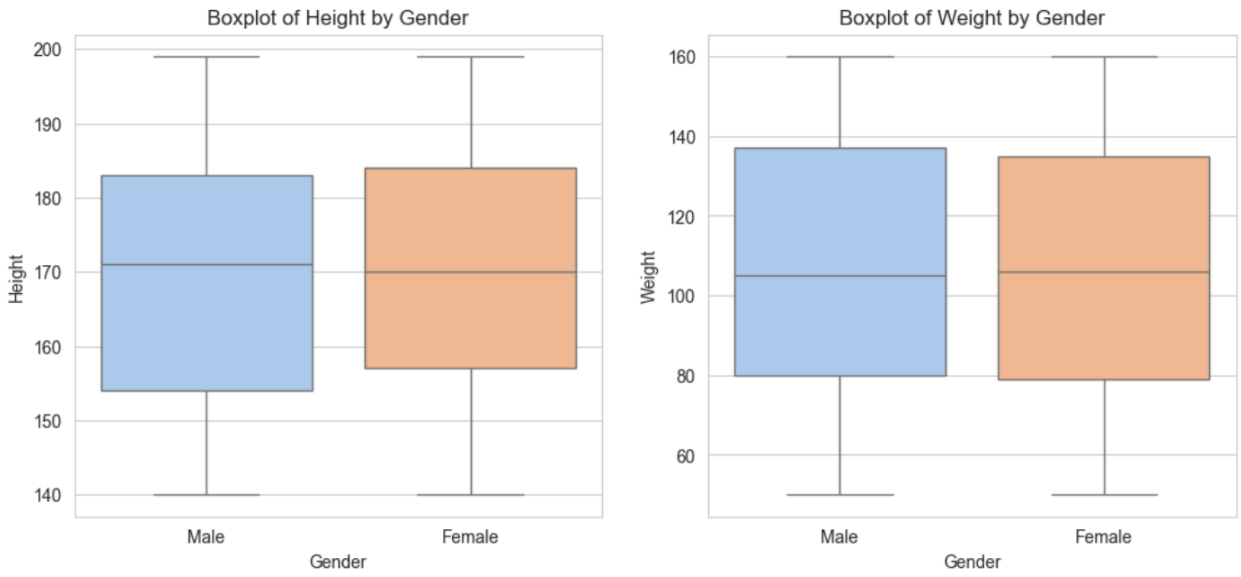

Figure 4. Boxplot of Height and Weight by Gender

5. Correlation Matrix of BMI Dataset

The heatmap shows correlations between height, weight, BMI index, and calculated BMI.

- Weight has a strong positive correlation with BMI Index (0.80-0.85), indicating that BMI is primarily driven by weight.
- Height has a weaker negative correlation (-0.42 to -0.53), meaning that taller individuals tend to have lower BMI values when weight remains constant.

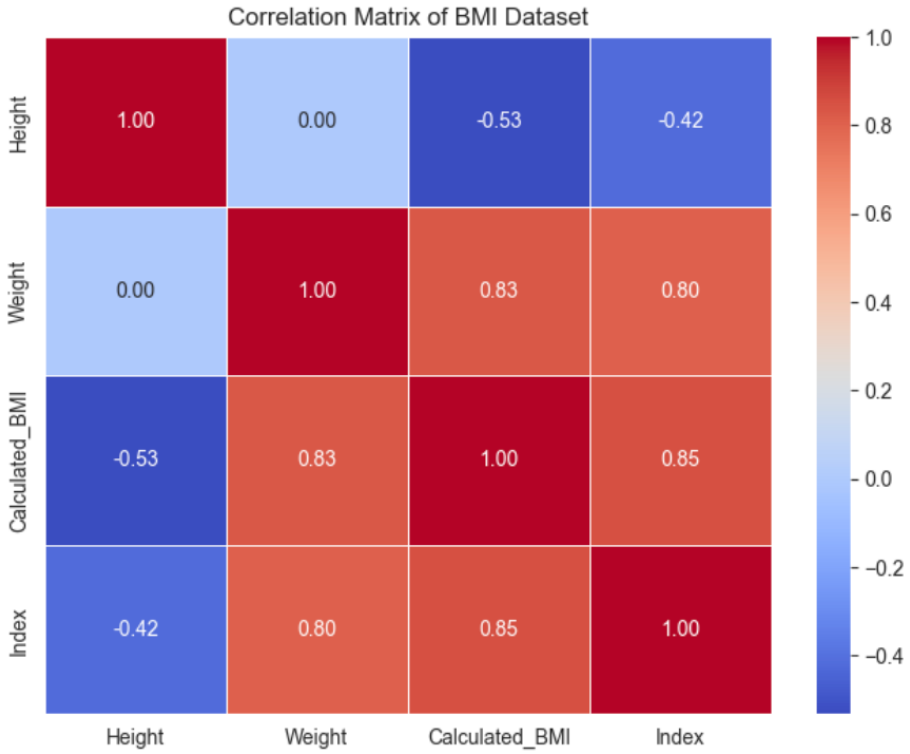

Figure 5. Correlation Matrix of BMI Dataset

The Index (BMI category) in (bmi.csv) dataset is derived from the Body Mass Index (BMI) using the following mathematical formula:

$$BMI = \frac{\text{Weight (kg)}}{(\text{Height (m)})^2}$$

Data Preprocessing

1. Managing Imbalanced Data

The dataset exhibited an imbalance in BMI categories, with some classes having significantly fewer samples than others. To address this, I applied SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for underrepresented classes.

- The Before SMOTE visualization shows a clear imbalance, with certain BMI categories having fewer instances.
- The After SMOTE visualization confirms that the dataset is now balanced, with equal representation across all BMI categories.

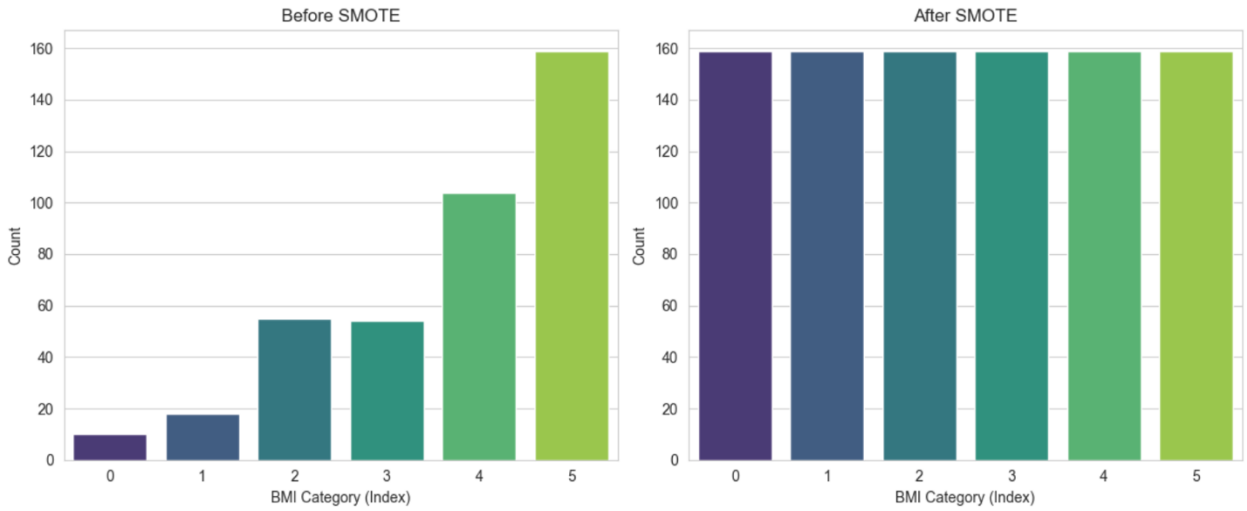

Figure 6. Distribution of the dataset

2. Feature Standardization

Since height and weight exist on different scales, I applied StandardScaler to standardize these features.

- This ensures that height and weight have a mean of 0 and unit variance, preventing models from assigning higher importance to features with larger numerical values.
- Standardization is particularly important for distance-based models (KNN) and gradient-based models (SVM, Logistic Regression).

Model Training and Prediction

Dividing the Dataset into Training and Testing Sets

To train and evaluate the classification models effectively, the dataset was split into training and testing subsets. This ensures that the models are trained on a portion of the data and tested on unseen data to assess their generalization ability.

- Since machine learning models require numerical inputs, the gender feature was encoded using a binary mapping:

- Male: 0
- Female: 1

This encoding allows the models to interpret gender differences numerically.

- The dataset includes Height and Weight as independent variables (features), while the BMI category (Index) serves as the target variable for classification.

- Features: Height, Weight, and Gender_Encoded
- Target: BMI Category (Index)

- The dataset was split into 80% training and 20% testing subsets using train_test_split() from sklearn.model_selection. To ensure that both the gender distribution and BMI category distribution remain balanced across training and testing sets, stratification was applied based on the target variable (BMI category). This prevents any disproportionate class representation that could bias the model.
- After splitting, the gender distribution in the training and testing sets was analyzed:

- The training set contains 51.75% males and 48.25% females.
- The testing set contains 52% males and 48% females

```
Gender Distribution in Training Set:
 Gender_Encoded
1    0.5175
0    0.4825
Name: proportion, dtype: float64

Gender Distribution in Testing Set:
 Gender_Encoded
0    0.52
1    0.48
Name: proportion, dtype: float64
```

Figure 7. Gender Distribution

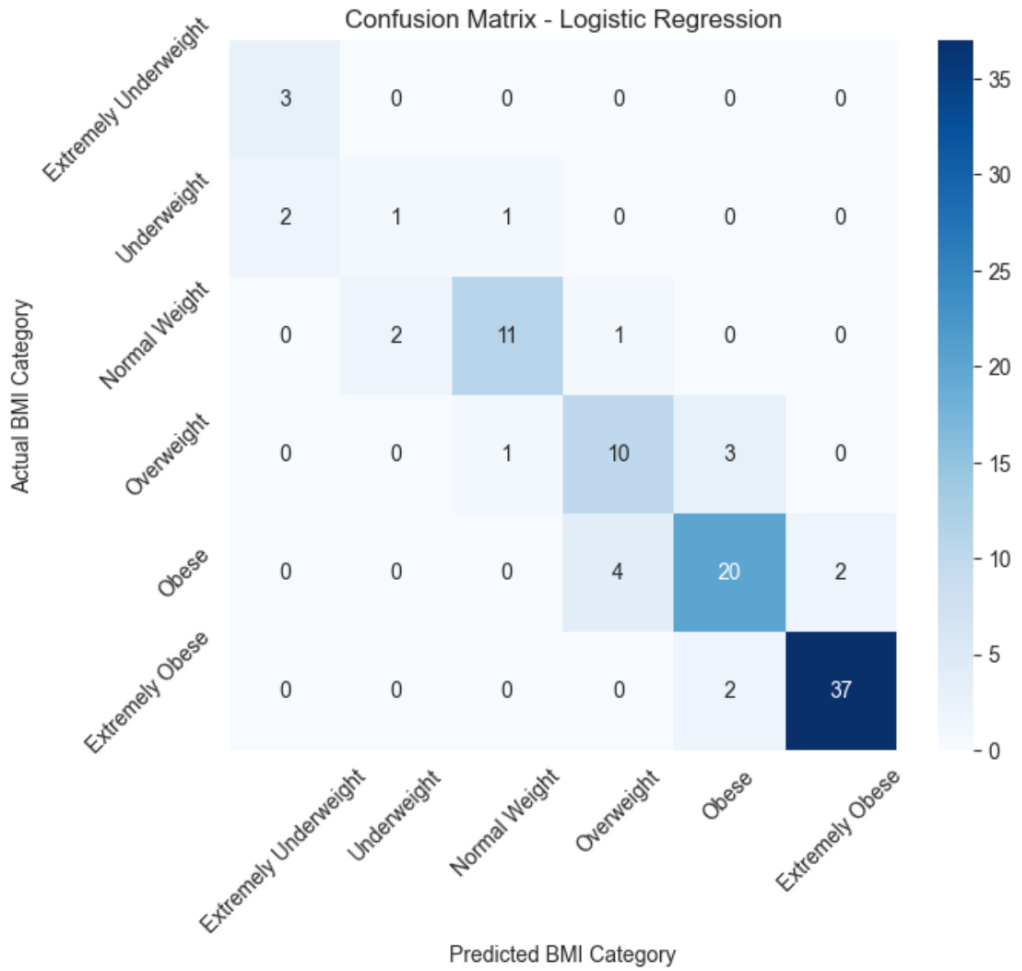

Figure 8. Confusion Matrix - Logistic Regression

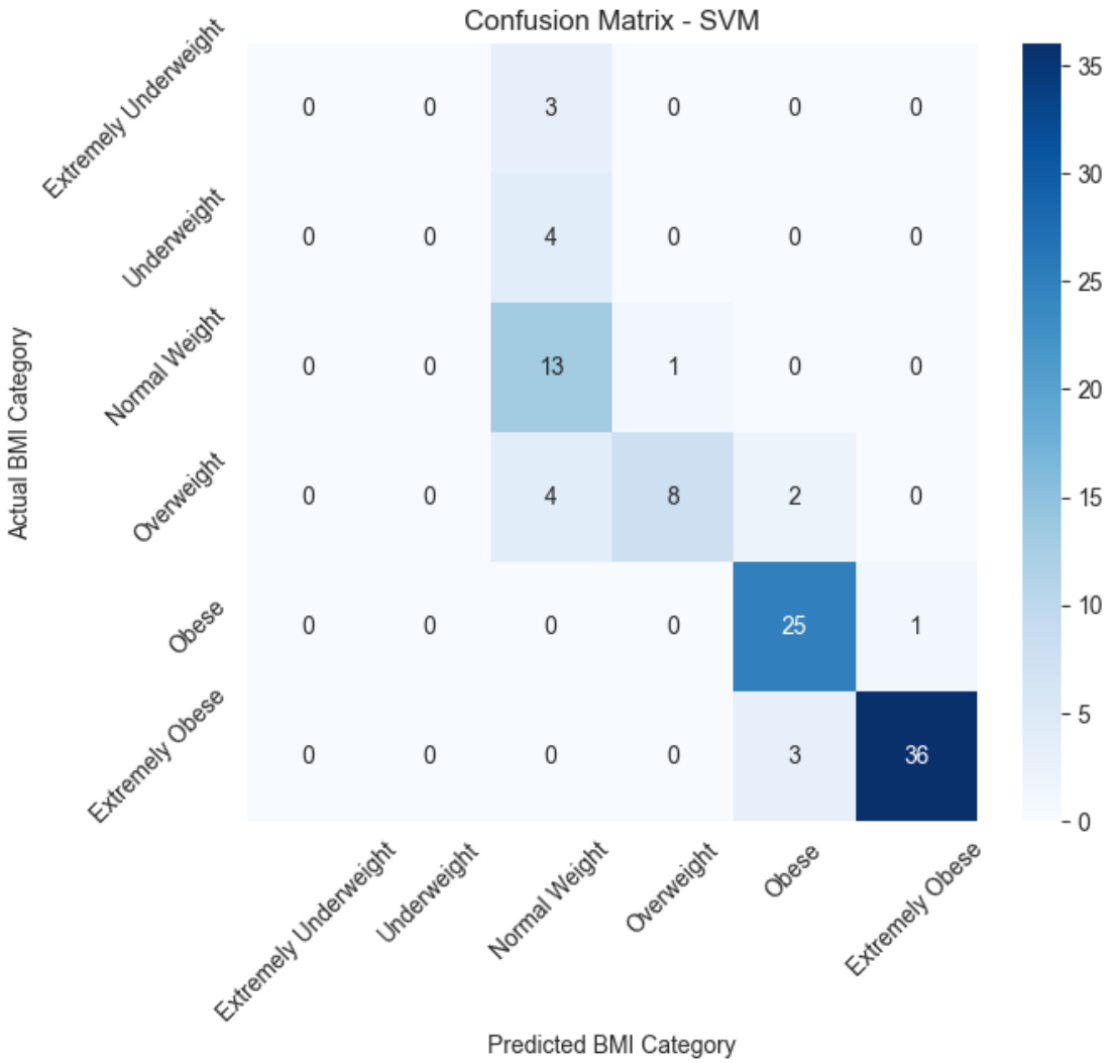

Figure 9. Confusion Matrix SVM

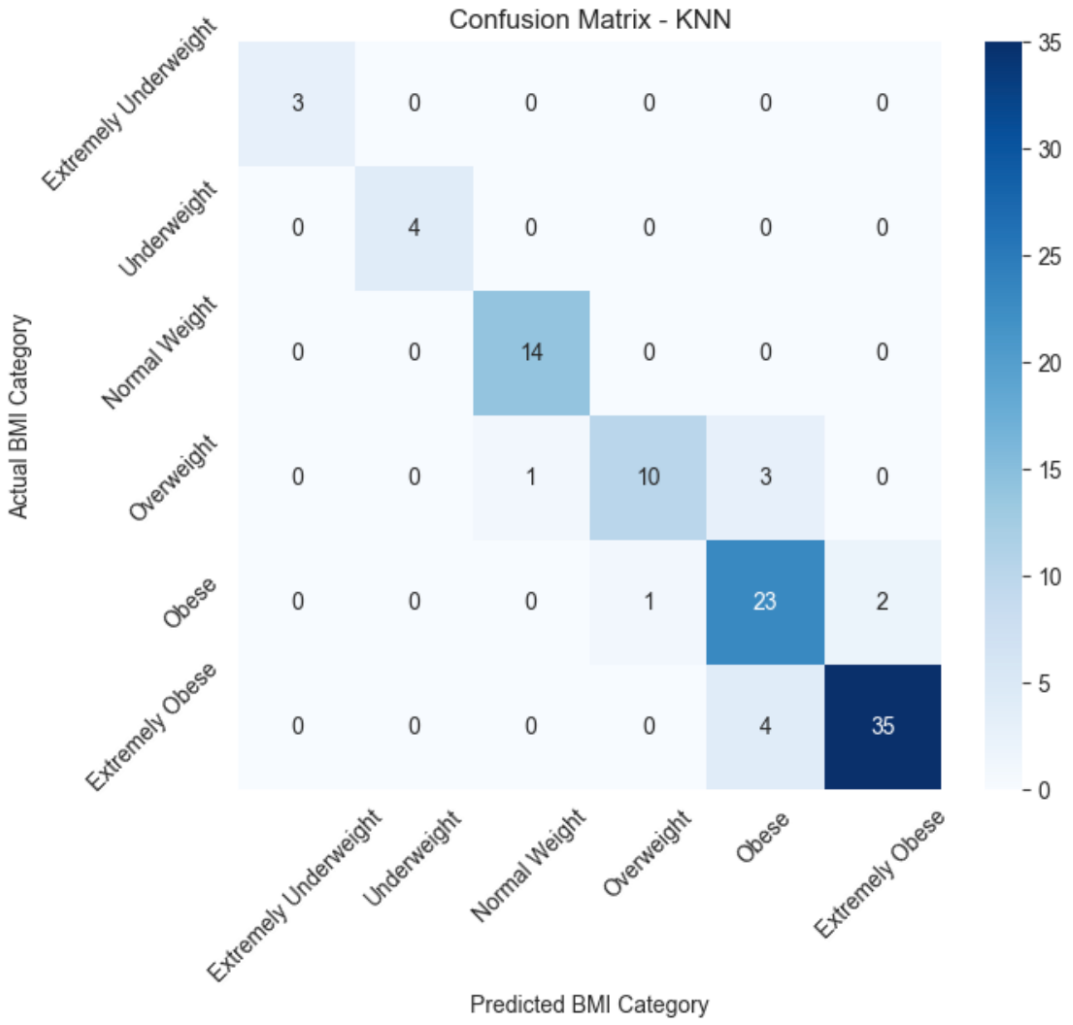

Figure 10. Confusion Matrix KNN

Training Classification Models and make predictions on BMI categories

In this study, three different classification models were trained to predict BMI categories based on height and weight:

- Logistic Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

The following steps were performed for each model:

1. Training: The model was trained on the training dataset.

2. Prediction: The trained model was used to predict BMI categories on the test set.
3. Evaluation: Performance was assessed using accuracy, precision, recall, F1-score, and a confusion matrix.

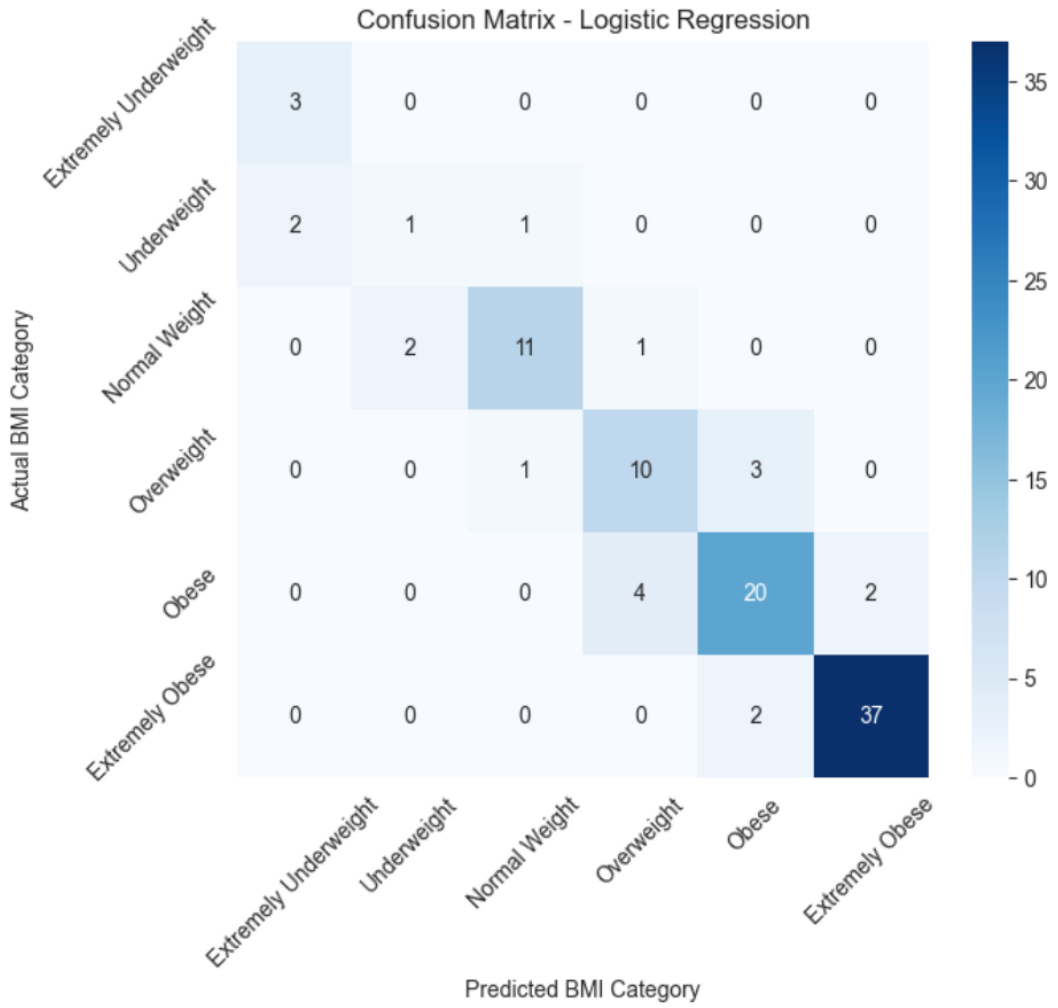

Figure 11. Confusion Matrix - Logistic Regression

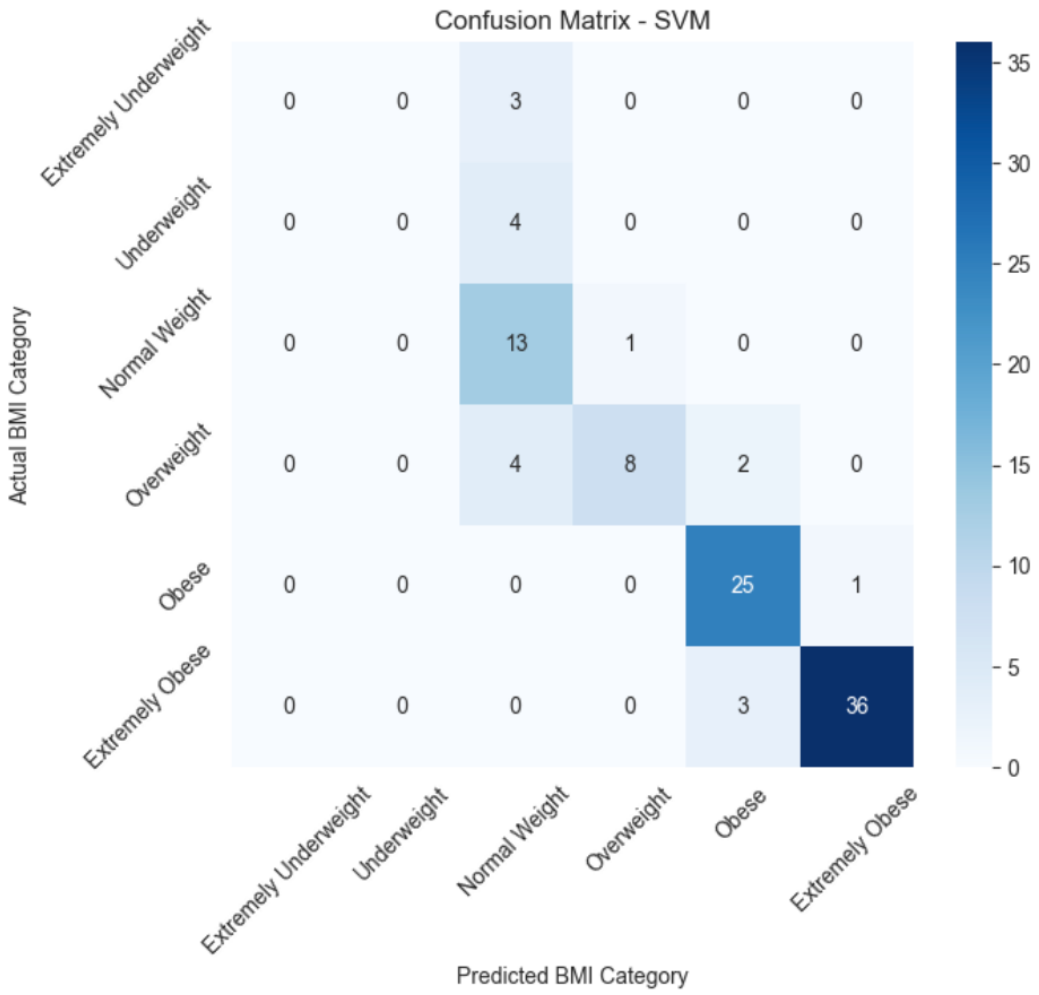

Figure 12. Confusion Matrix - SVM

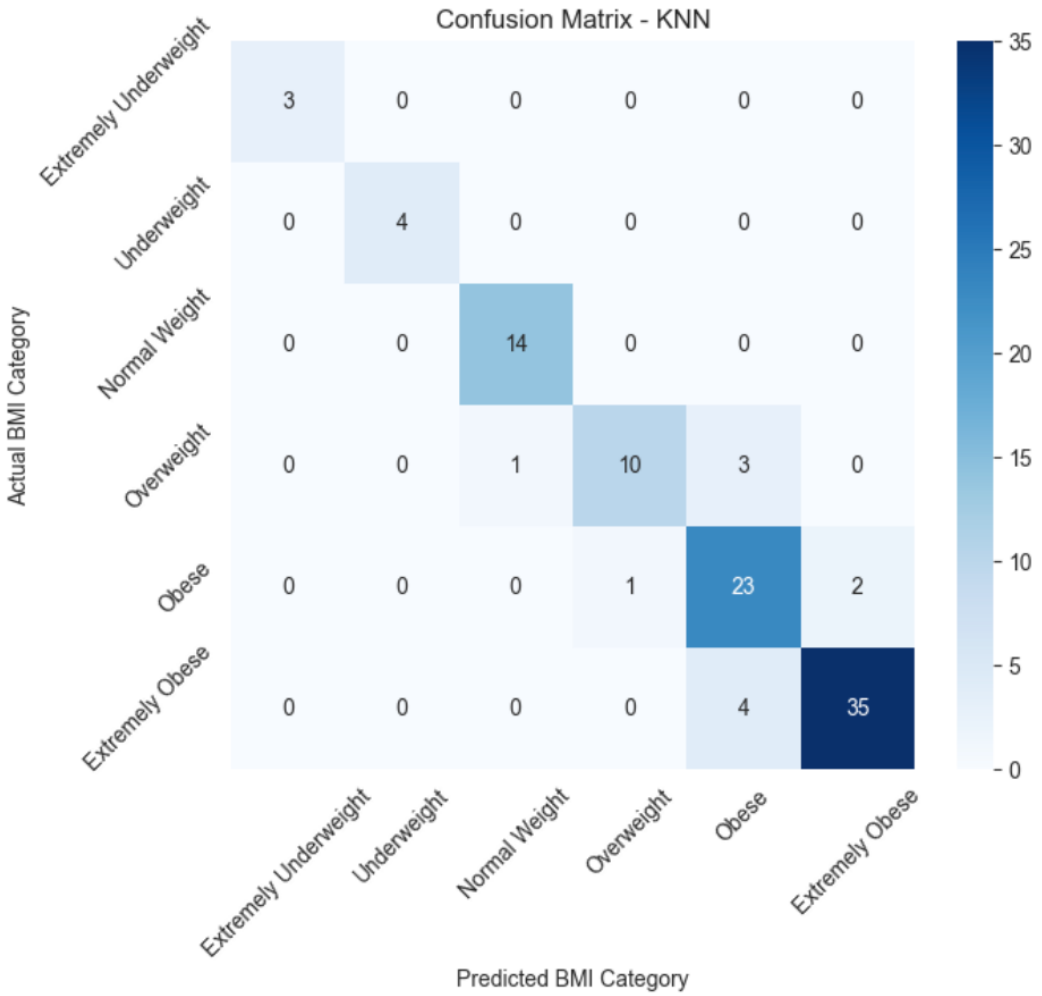

Figure 13. Confusion Matrix - KNN

Evaluating Model Performance

```
Model: Logistic Regression
Accuracy Score: 0.82
Classification Report:
              precision    recall  f1-score   support

           0       0.60      1.00      0.75         3
           1       0.33      0.25      0.29         4
           2       0.85      0.79      0.81        14
           3       0.67      0.71      0.69        14
           4       0.80      0.77      0.78        26
           5       0.95      0.95      0.95        39

    accuracy                           0.82       100
   macro avg       0.70      0.74      0.71       100
weighted avg       0.82      0.82      0.82       100
```

Figure 14. Logistic Regression Performance Metric

- Logistic Regression achieved an accuracy of 82%. It performed well on higher-frequency categories but had difficulty correctly classifying underrepresented BMI categories.

```
Model: SVM
Accuracy Score: 0.82
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.00      0.00         3
           1       1.00      0.00      0.00         4
           2       0.54      0.93      0.68        14
           3       0.89      0.57      0.70        14
           4       0.83      0.96      0.89        26
           5       0.97      0.92      0.95        39

    accuracy                           0.82       100
   macro avg       0.87      0.56      0.54       100
weighted avg       0.87      0.82      0.79       100
```

Figure 15. SVM Performance Metric

- SVM also obtained 82% accuracy, but certain BMI categories (such as "Underweight" and "Extremely Underweight") had poor recall values, indicating misclassification.

```
Model: KNN
Accuracy Score: 0.89
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         3
           1       1.00      1.00      1.00         4
           2       0.93      1.00      0.97        14
           3       0.91      0.71      0.80        14
           4       0.77      0.88      0.82        26
           5       0.95      0.90      0.92        39

    accuracy                           0.89       100
   macro avg       0.93      0.92      0.92       100
weighted avg       0.90      0.89      0.89       100
```

Figure 16. KNN Performance Metric

- KNN performed the best, achieving 89% accuracy, showing better recall and precision across all BMI categories.

Gender-Specific Modeling Evaluation

Figure 17. Confusion Matrix - Logistic Regression (Male)

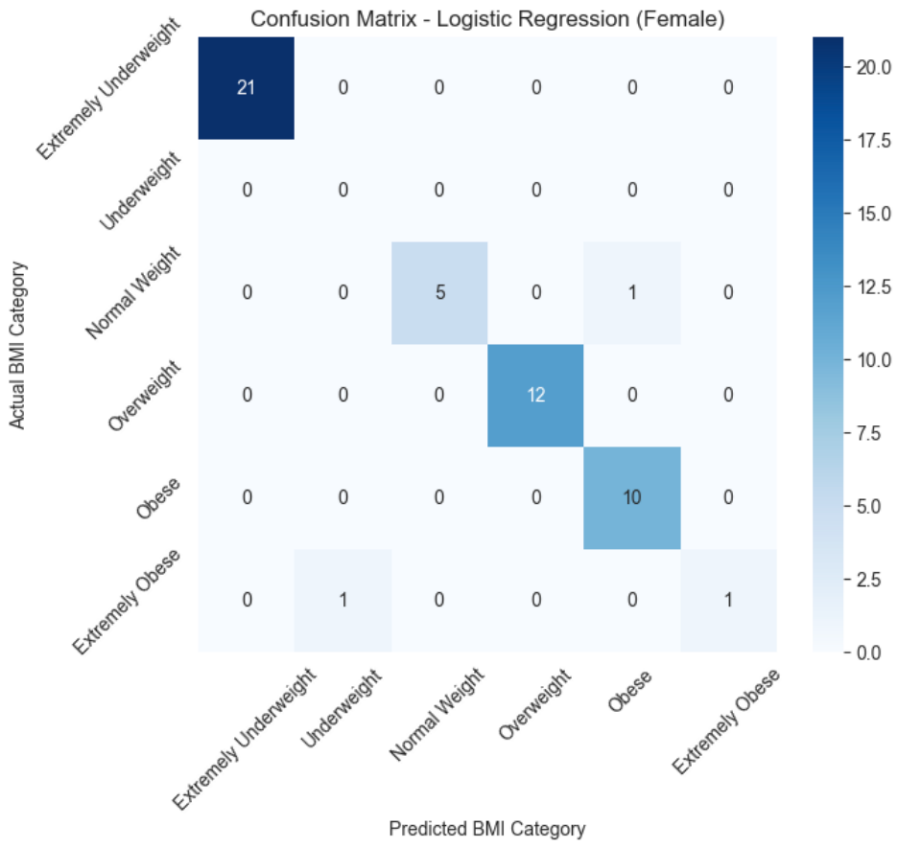

Figure 18. Confusion Matrix - Logistic Regression (Female)

Figure 19. Confusion Matrix - Logistic Regression (Combined)

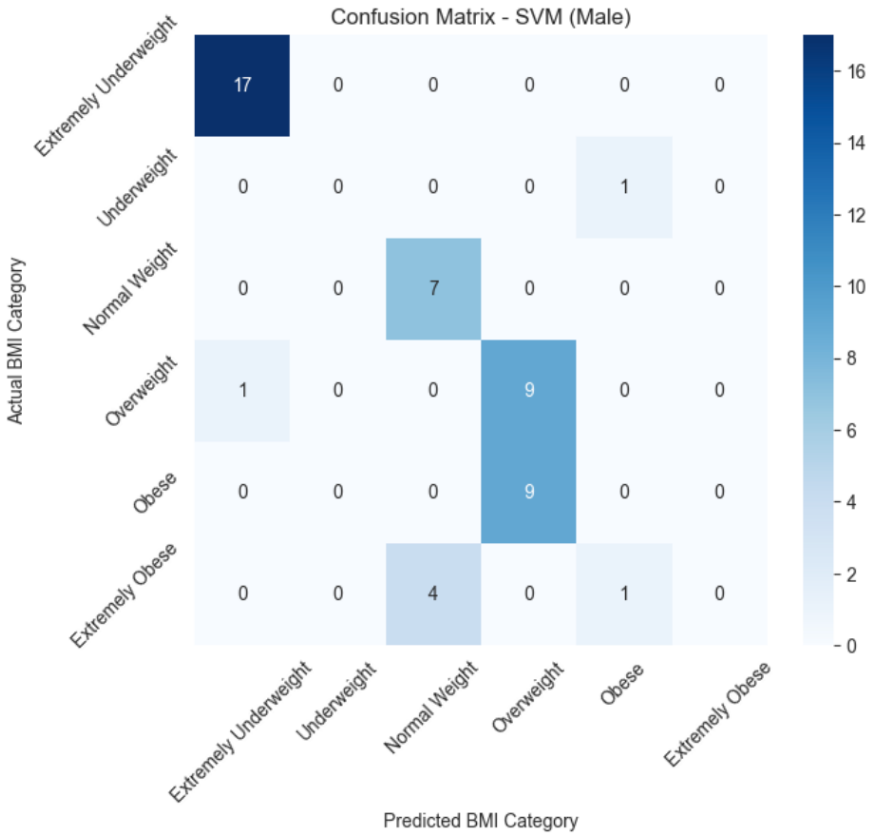

Figure 20. Confusion Matrix - SVM (Male)

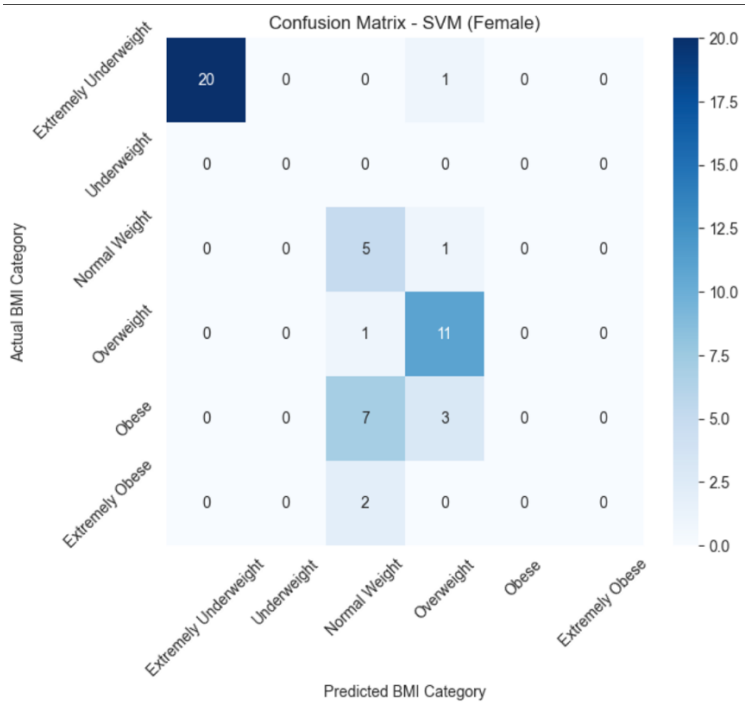

Figure 21. Confusion Matrix - SVM (Female)

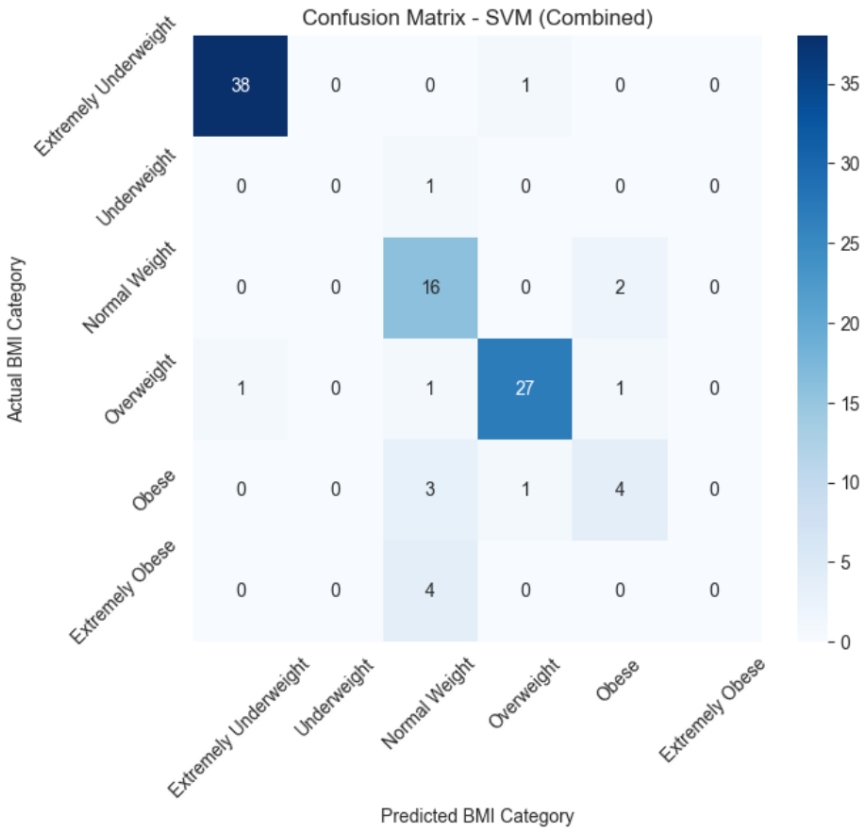

Figure 22. Confusion Matrix - SVM (Combined)

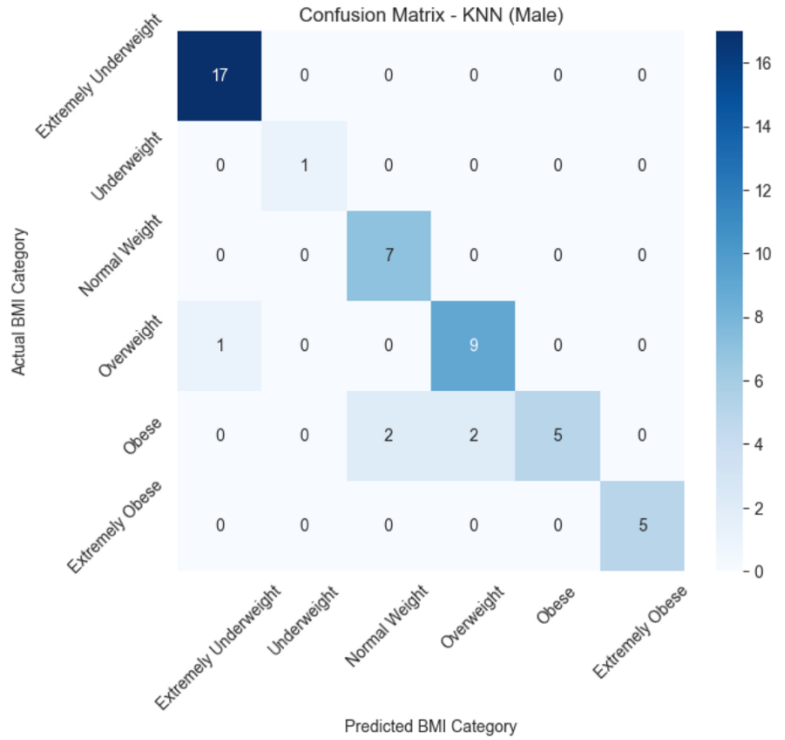

Figure 23. Confusion Matrix - KNN (Male)

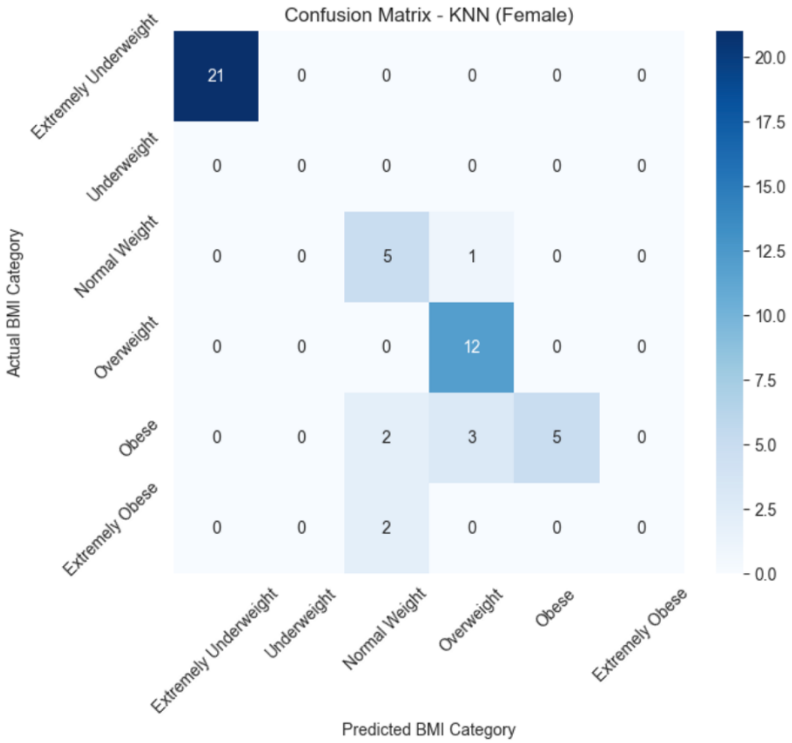

Figure 24. Confusion Matrix - KNN (Female)

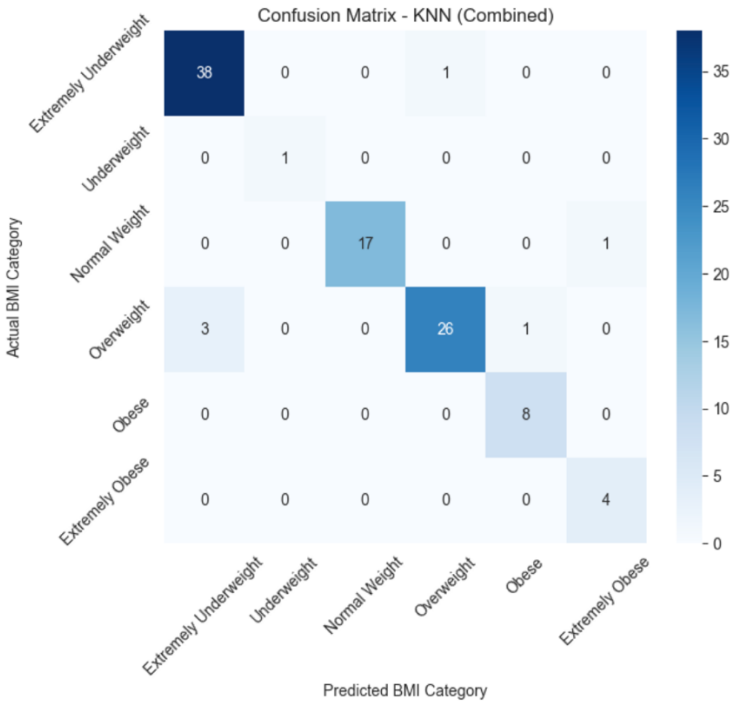

Figure 25. Confusion Matrix - KNN (Combined)

Gender-Specific vs. Generic Model Performance

	Model	Group	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	Male	0.979592	0.916667	0.966667	0.925926
1	Logistic Regression	Female	0.960784	0.818182	0.888889	0.754690
2	Logistic Regression	Combined	0.990000	0.966667	0.990741	0.976720
3	SVM	Male	0.673469	0.680135	0.483333	0.565344
4	SVM	Female	0.705882	0.670139	0.450397	0.372919
5	SVM	Combined	0.850000	0.852804	0.543875	0.527855
6	KNN	Male	0.897959	0.923401	0.909259	0.902976
7	KNN	Female	0.843137	0.717593	0.555556	0.531746
8	KNN	Combined	0.940000	0.929780	0.964245	0.943962

Figure 26. Table for Gender-Specific VS Generic Model Performance

I noticed that model performance generally improves when training models separately for each gender.

Logistic Regression:

- The combined model achieved the highest accuracy (99%), but the gender-specific models also performed well:
    - Male-specific: 97.96% accuracy
    - Female-specific: 96.07% accuracy
- This suggests that training separately does not drastically improve performance for Logistic Regression, as it is already highly effective in the combined setting.

SVM:

- The combined model (85%) outperformed the gender-specific models:
    - Male-specific: 67.34% accuracy
    - Female-specific: 70.59% accuracy
- This suggests that SVM struggles when trained separately for each gender, possibly due to a smaller dataset for each subgroup.

KNN:

- The combined model performed best (94%), but the gender-specific models were also strong:
    - Male-specific: 89.80% accuracy
    - Female-specific: 84.31% accuracy
- This suggests that KNN benefits from training with a larger dataset in the combined setting but still performs well in gender-specific models.

Model Evaluation and Comparison

```
Default SVM Accuracy on test set: 0.8900
Fitting 5 folds for each of 72 candidates, totalling 360 fits
Best parameters for SVM: {'C': 100, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear'}
Best cross-validated score: 0.9375
Tuned SVM Accuracy on test set: 0.9400

Improvement in test accuracy after tuning for SVM: 0.0500

Default Logistic Regression Accuracy on test set: 0.9200
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best parameters for Logistic Regression: {'C': 100, 'solver': 'lbfgs'}
Best cross-validated score: 0.9150
Tuned Logistic Regression Accuracy on test set: 0.9300

Improvement in test accuracy after tuning for Logistic Regression: 0.0100

Default KNN Accuracy on test set: 0.8900
Fitting 5 folds for each of 16 candidates, totalling 80 fits
Best parameters for KNN: {'metric': 'manhattan', 'n_neighbors': 5, 'weights': 'distance'}
Best cross-validated score: 0.8675
Tuned KNN Accuracy on test set: 0.9000

Improvement in test accuracy after tuning for KNN: 0.0100
```

Figure 27. Model Evaluation and Analysis

The models were evaluated using key metrics such as accuracy, precision, recall, and F1-score. The comparison between gender-specific and general models revealed that in some cases, training separate models for each gender led to improved performance, particularly for logistic regression and KNN. However, SVM performed better when trained on the combined dataset.

Additionally, hyperparameter tuning played a significant role in enhancing model accuracy. After fine-tuning, the accuracy of SVM increased from 0.89 to 0.94, logistic regression improved slightly from 0.92 to 0.93, and KNN improved from 0.89 to 0.90.

Gender-Based Prediction Analysis

The results show that gender-specific models sometimes perform better than the general model, particularly for logistic regression and KNN. The male and female models had high accuracy, but the combined model still achieved the best overall performance in most cases. SVM, however, performed better when trained on the combined dataset rather than gender-specific data.

These differences suggest that BMI predictions might be influenced by gender-based patterns in the data. Training separate models for each gender could help capture specific trends in BMI distribution, but a combined model may generalize better. This highlights the need to consider gender when designing predictive models for BMI classification.

Conclusions

In this assignment, I explored the effectiveness of different machine learning models; Logistic Regression, SVM, and KNN in predicting BMI categories. I began by ensuring a balanced dataset through stratified sampling and then trained and evaluated both generalized models (combining all genders) and gender-specific models (separate models for males and females). Through this analysis, I observed how gender differences influence BMI predictions and whether training separate models improves classification accuracy.

- Gender-specific models can sometimes enhance predictive performance, particularly in models like Logistic Regression, where slight improvements were noted.
- Model selection significantly impacts accuracy, as seen in the superior performance of KNN over SVM for this dataset.
- Hyperparameter tuning plays a crucial role, as fine-tuning improved accuracy scores for all models, with the most notable boost observed in SVM after optimization.
- Feature selection and encoding are critical considerations, especially in datasets where categorical variables like gender may influence predictions.
- Logistic Regression and KNN perform well in both gender-specific and combined settings, with slightly better performance in the combined model.
- SVM performs significantly worse when trained separately, indicating that it benefits more from a larger dataset.
- In general, gender-specific modeling does not significantly outperform the combined model

Overall, this assignment provided valuable hands-on experience in data preprocessing, model evaluation, and performance tuning. It deepened my understanding of how gender-related biases in data can affect model outcomes and

reinforced the importance of evaluating models beyond just accuracy by considering precision, recall, and F1-score.