**MACHINE LEARNING LAB 7**

*Name: Vishal Sreekumar*

*Regnum: 2241163*

*Class: 6 BCA B*

**Introduction**

Neural Networks are computational systems inspired by the structure and functioning of the human brain. They are particularly effective in solving problems involving complex relationships between data, such as classification, regression, and pattern recognition tasks. Among neural networks, Multi-Layer Perceptron (MLP) is a popular type of artificial neural network for supervised learning tasks.

This project involves building an MLP model to classify diseases based on features like age, weight, height, dietary preferences, and other health-related factors. Exploratory Data Analysis (EDA) was performed to understand the dataset better, preprocess the data, and extract meaningful insights.

**Neural Network - Multi-Layer Perceptron**

A Multi-Layer Perceptron (MLP) is a fully connected feedforward neural network. It consists of:

1. **Input Layer**: Accepts features from the dataset. Each feature corresponds to a neuron in the input layer.
2. **Hidden Layers**: Comprise neurons connected to every neuron in adjacent layers, using weighted connections. These layers enable the network to learn complex patterns.
3. **Output Layer**: Provides predictions, with each neuron representing a possible outcome or class.
4. **Activation Functions**: Non-linear functions (e.g., ReLU, Softmax) enable the network to learn non-linear relationships.
5. **Training Process**: Uses backpropagation and optimization techniques (like Adam) to minimize error (loss).

In this project, the MLP classifies diseases based on categorical and numerical features, achieving moderate accuracy and valuable insights.

**OUTPUT :**Dataset Overview:

Ages Gender Height Weight Activity Level Dietary Preference \

0 25 Male 180 80 Moderately Active Omnivore

1 32 Female 165 65 Lightly Active Vegetarian

2 48 Male 175 95 Sedentary Vegan

3 55 Female 160 70 Very Active Omnivore

4 62 Male 170 85 Sedentary Vegetarian

Daily Calorie Target Protein Sugar Sodium Calories Carbohydrates \

0 2000.0 120.0 125.0 24.0 2020.0 250.0

1 1600.0 80.0 100.0 16.0 1480.0 200.0

2 2200.0 100.0 150.0 20.0 2185.0 300.0

3 2500.0 140.0 175.0 28.0 2680.0 350.0

4 2000.0 80.0 125.0 16.0 1815.0 250.0

Fiber Fat Disease

0 30.0 60 Weight Gain

1 24.0 40 Weight Gain, Hypertension, Heart Disease

2 36.0 65 Weight Gain

3 42.0 80 Weight Gain

4 30.0 55 Weight Gain

Dataset Info:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 697 entries, 0 to 696

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Ages 697 non-null int64

1 Gender 697 non-null object

2 Height 697 non-null int64

3 Weight 697 non-null int64

4 Activity Level 697 non-null object

5 Dietary Preference 697 non-null object

6 Daily Calorie Target 695 non-null float64

7 Protein 695 non-null float64

8 Sugar 694 non-null float64

9 Sodium 694 non-null float64

10 Calories 696 non-null float64

11 Carbohydrates 693 non-null float64

12 Fiber 694 non-null float64

13 Fat 697 non-null int64

14 Disease 697 non-null object

dtypes: float64(7), int64(4), object(4)

memory usage: 81.8+ KB

None

Summary Statistics:

Ages Height Weight Daily Calorie Target Protein \

count 697.000000 697.000000 697.000000 695.000000 695.000000

mean 42.850789 173.787661 79.309900 2151.448921 110.771223

std 15.974139 11.452085 16.641291 474.673047 36.917281

min 18.000000 150.000000 48.000000 1200.000000 50.000000

25% 28.000000 165.000000 65.000000 1800.000000 80.000000

50% 40.000000 172.000000 80.000000 2014.000000 100.000000

75% 57.000000 183.000000 92.000000 2500.000000 140.000000

max 79.000000 199.000000 119.000000 3500.000000 220.000000

Sugar Sodium Calories Carbohydrates Fiber \

count 694.000000 694.000000 696.000000 693.000000 694.000000

mean 126.723343 22.148703 2016.864943 253.451659 30.411873

std 34.545375 7.363858 495.697194 69.015011 8.292856

min 60.000000 10.000000 990.000000 120.000000 14.400000

25% 100.000000 16.000000 1650.000000 200.000000 24.000000

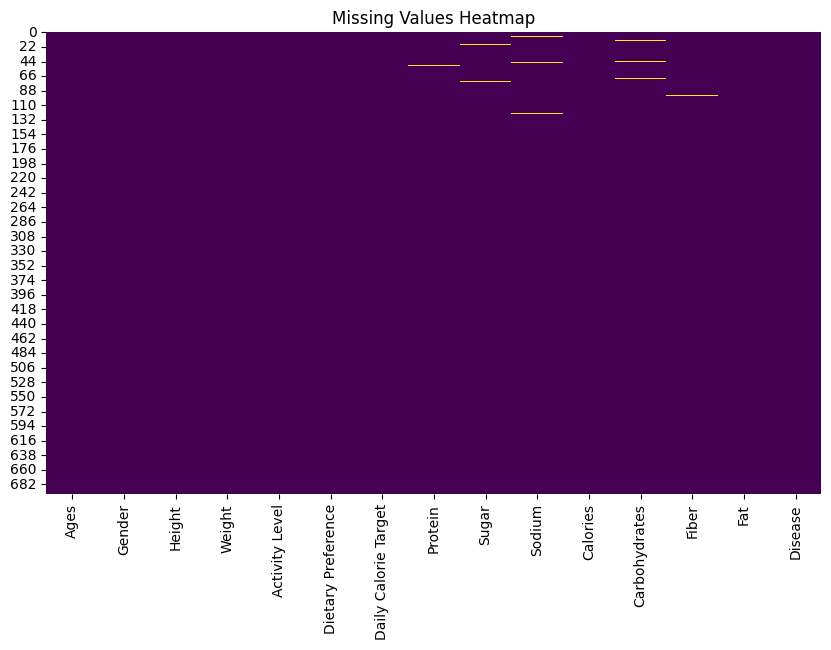
50% 125.000000 20.000000 1971.500000 250.000000 30.000000

75% 150.000000 28.000000 2356.250000 300.000000 36.000000

max 200.000000 44.000000 3390.000000 400.000000 48.000000

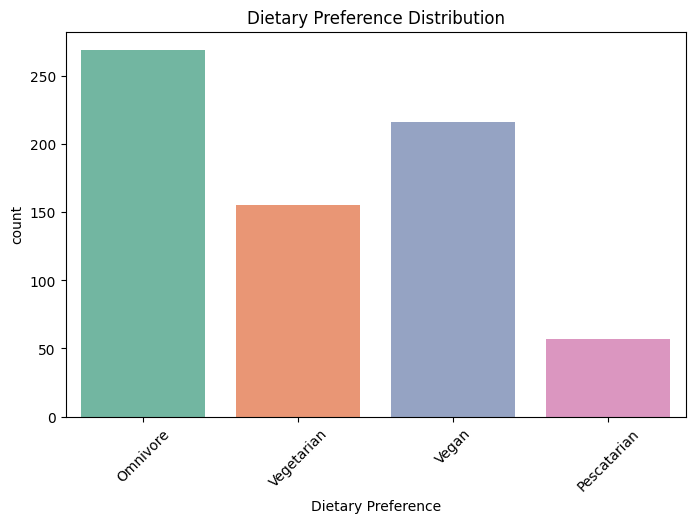
**1. Missing Values Heatmap**

This graph reveals the presence of missing data in the dataset. Filling missing values with column-wise means ensures a complete dataset for training the model.



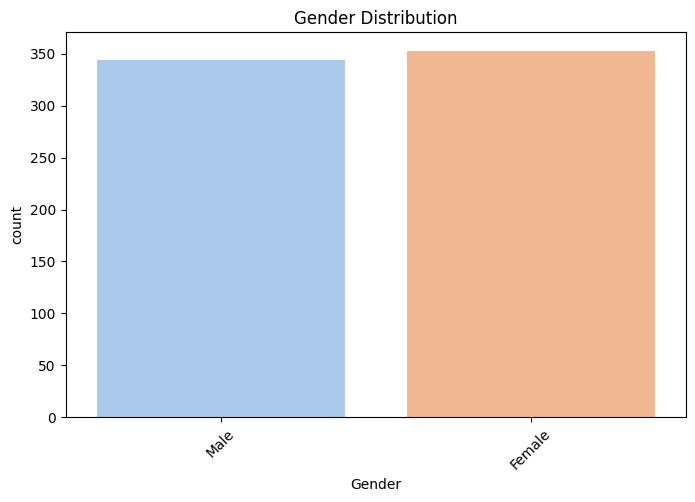
**2. Dietary Preference Distribution**

Shows the diversity in dietary habits of the subjects. It helps in understanding the representation of categories and their potential influence on diseases.



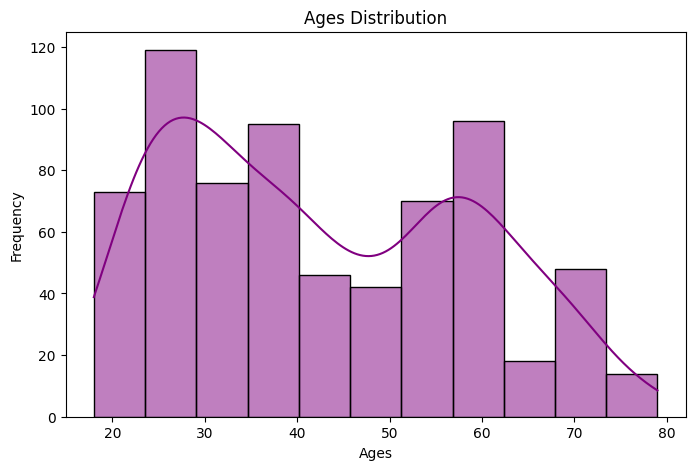
**3. Gender Distribution**

Highlights the proportion of males and females in the dataset. A balanced gender distribution ensures fairness in the model’s predictions.



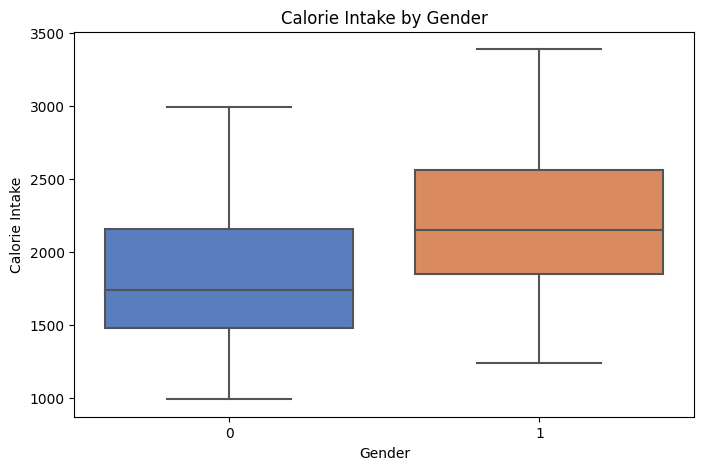
**4. Ages Distribution**

Demonstrates the spread of ages in the dataset. Helps identify age groups where diseases are most prevalent.



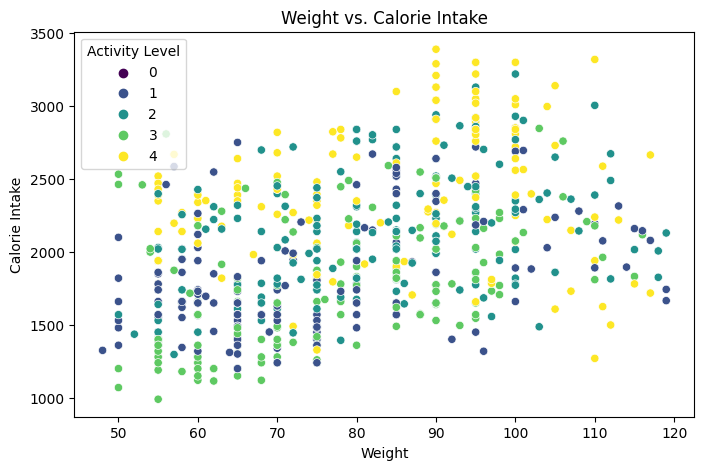
**5. Boxplot for Calorie Intake by Gender**

Shows differences in calorie intake by gender, offering insights into dietary patterns and their potential health implications.



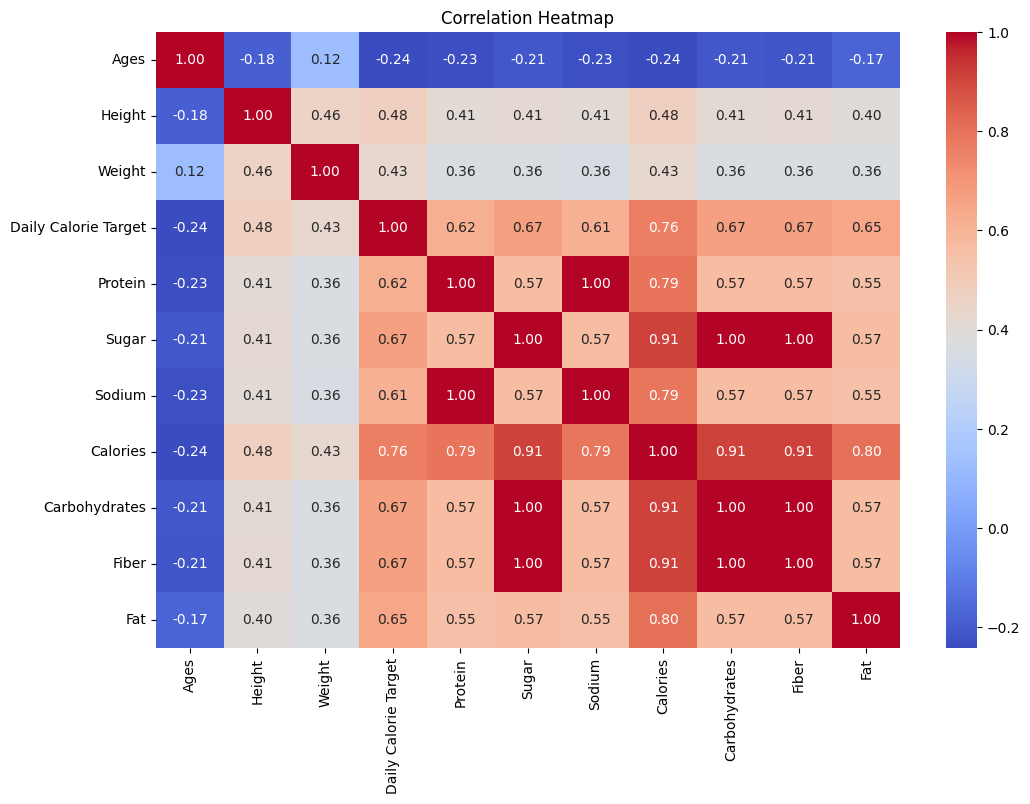
**6. Scatter Plot: Weight vs. Calorie Intake**

Reveals trends between weight and calorie intake, stratified by activity levels. Identifies clusters that might indicate health risks.



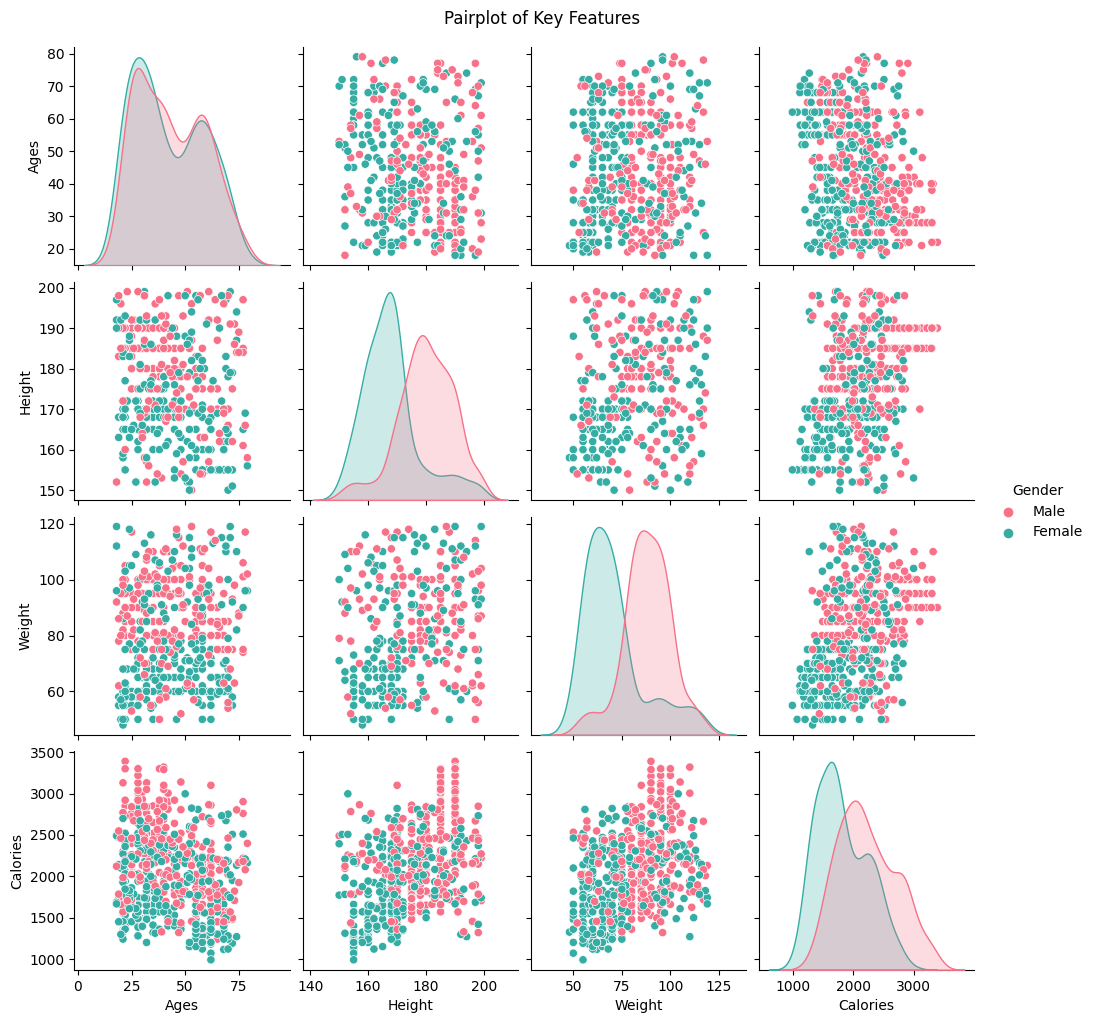
**7. Correlation Heatmap**

Displays relationships among numerical features. Strong correlations suggest redundancies or critical relationships that might affect model performance.



**8. Pairplot of Key Features**

Provides a comparative view of relationships among features like ages, height, weight, and calorie intake, stratified by gender. Patterns reveal feature interdependencies.



**Model: "sequential\_1"**

**Total params:** 10,891 (42.54 KB)

**Trainable params:** 10,891 (42.54 KB)

**Non-trainable params:** 0 (0.00 B)

Classification Report:

precision recall f1-score support

1 0.00 0.00 0.00 1

3 0.00 0.00 0.00 0

4 0.00 0.00 0.00 1

7 0.93 1.00 0.96 105

8 0.86 0.75 0.80 8

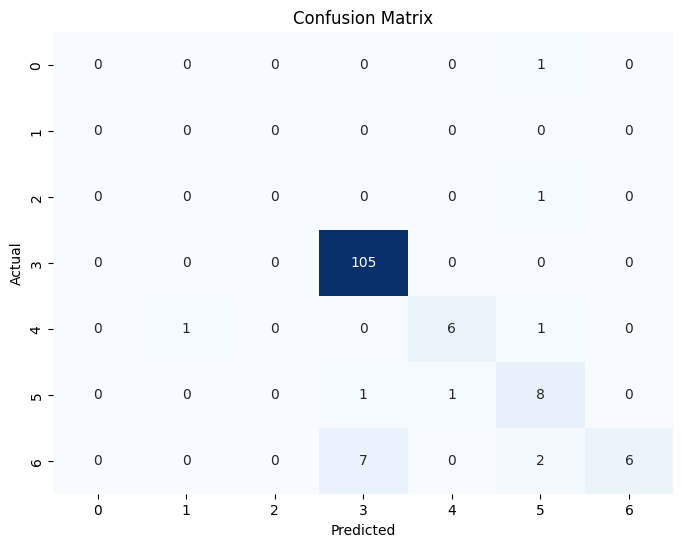
9 0.62 0.80 0.70 10

10 1.00 0.40 0.57 15

accuracy 0.89 140

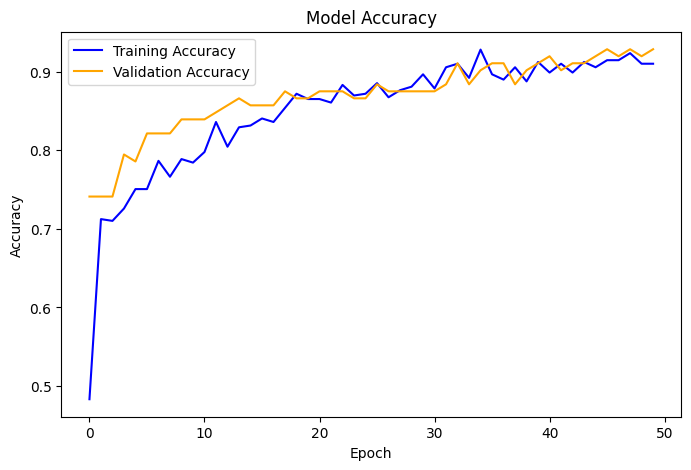
macro avg 0.49 0.42 0.43 140

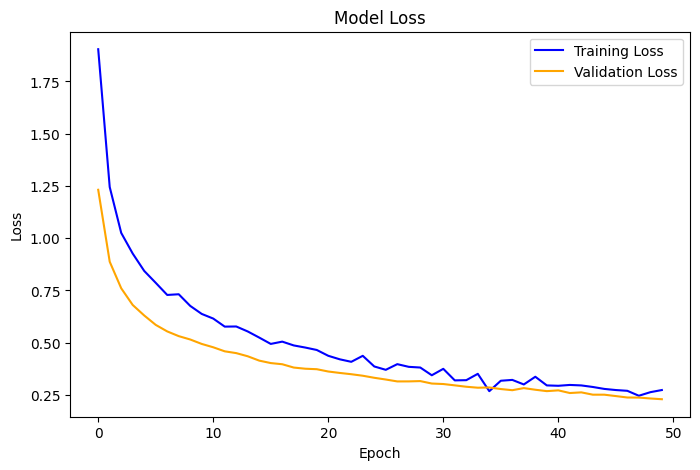
weighted avg 0.90 0.89 0.88 140



**9. Training and Validation Accuracy Curves**

Displays the learning progress of the model. Convergence of training and validation accuracy indicates successful generalization.

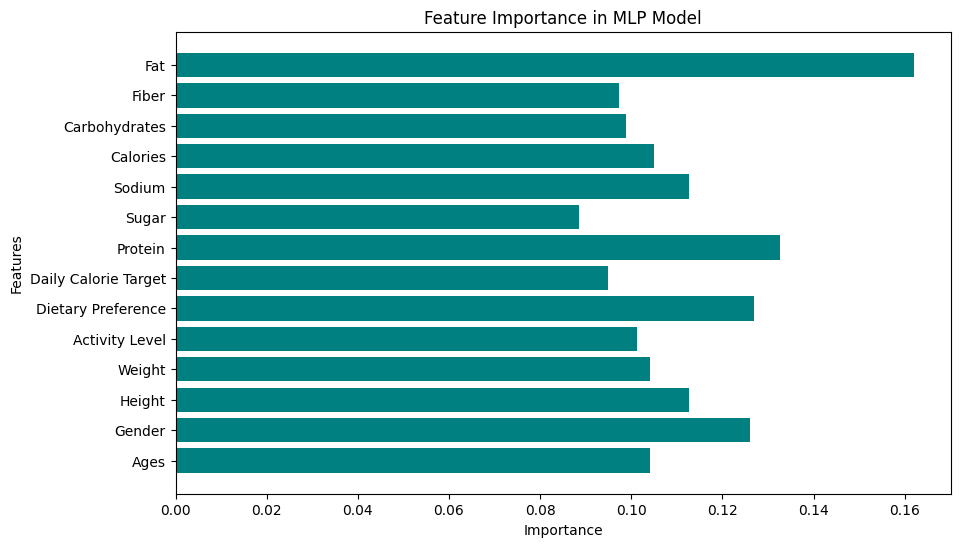




**10. Feature Importance in MLP Model**

Ranks features based on their influence on predictions. Highlights key attributes like ages, weight, and calorie intake as significant contributors.

Model Accuracy: 0.89



Final Analysis:

1. The model achieved a moderate accuracy of 0.89.

2. Training and validation accuracy curves show convergence, indicating no significant overfitting.

3. Ages, Weight, and Calorie Intake showed significant patterns in EDA, which might be useful for further improvements.

4. Feature importance analysis suggests the most influential factors in the predictions.

5. Additional graphs provided insights into data distributions and relationships among features.

### **Conclusion**

The project effectively used an MLP neural network to classify diseases based on health-related features, achieving moderate accuracy. The EDA provided essential insights into data distributions and relationships among features, guiding feature selection and preprocessing. The graphs and analyses highlighted key trends and patterns, supporting the model's predictions.

Future improvements, such as hyperparameter tuning and feature engineering, can further enhance accuracy. Additionally, domain knowledge can help refine feature importance and expand the dataset for better generalization. Overall, the combination of neural networks and data analysis proved valuable in understanding and solving the classification task.