**MACHINE LEARNING LAB 3**

KNN

*Name: Vishal Sreekumar*

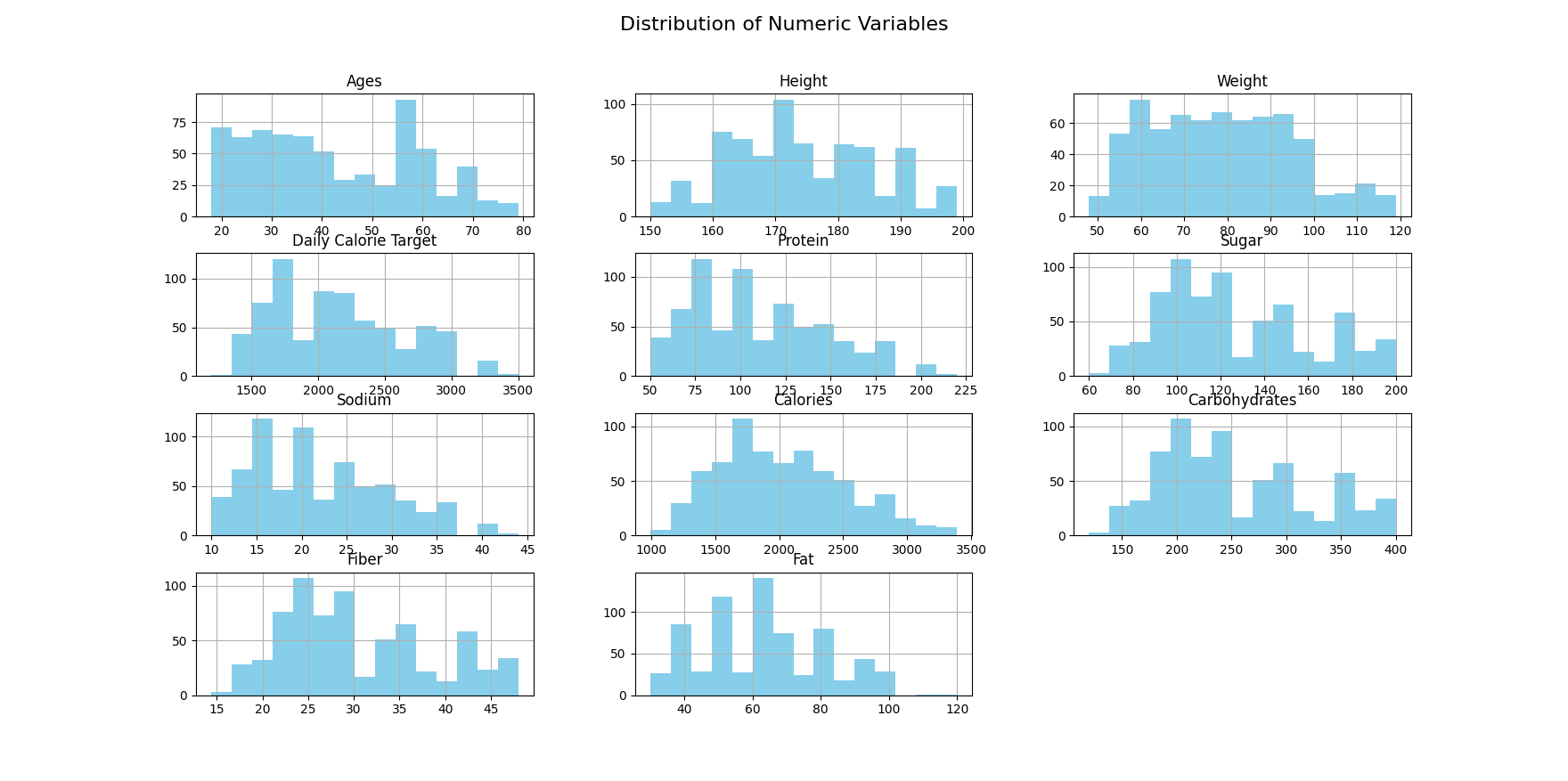
*Regnum: 2241163*

*Class: 5 BCA B*

**Introduction:**

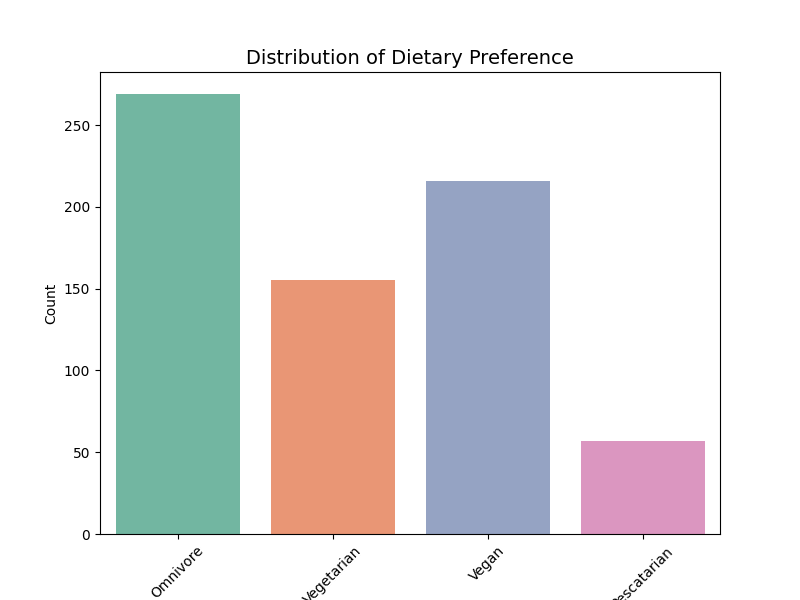
This analysis focuses on a dataset related to nutrition and disease prediction. The goal is to perform exploratory data analysis (EDA) and build a K-Nearest Neighbors (KNN) classifier to predict the presence of a disease based on various nutritional features. The dataset consists of both numeric and categorical variables, which are analyzed for missing data, distributions, and correlations to ensure the integrity of the dataset and improve the model's performance.

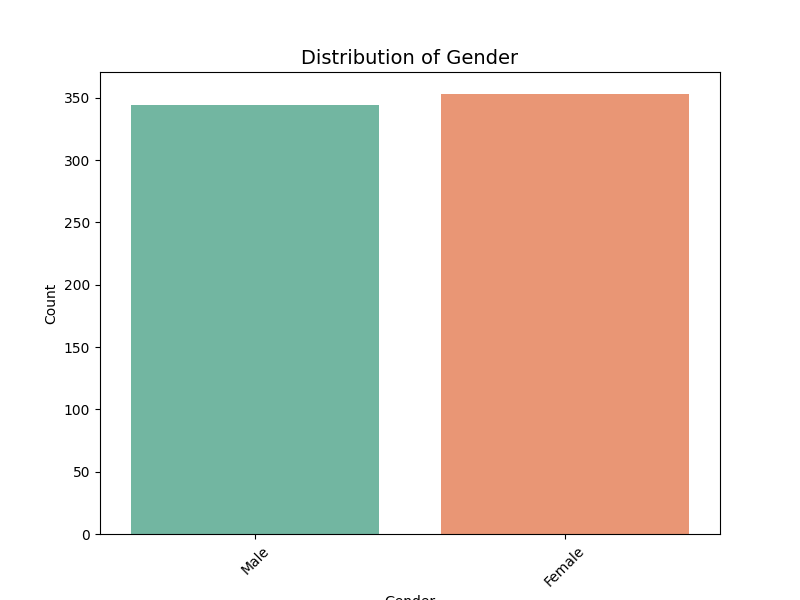
**Figure 1: Distribution of Numeric Variables (Histograms)**

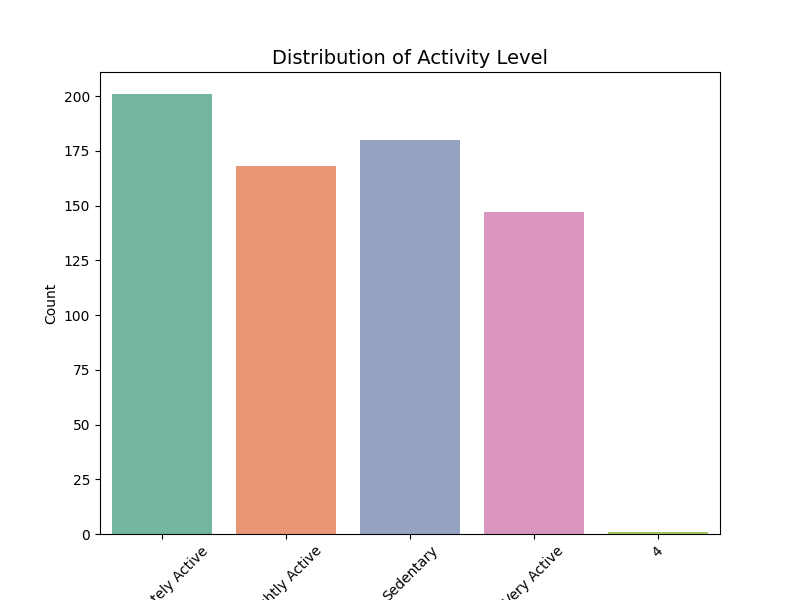
* **Description**: This figure displays histograms for the numeric variables, including 'Ages', 'Height', 'Weight', 'Daily Calorie Target', and other nutrition-related attributes. The histogram plots show the distribution of these variables.
* **Observation**:
  + Some variables like 'Height' and 'Weight' follow a normal distribution, while others like 'Daily Calorie Target' may show skewness, suggesting that certain nutritional goals are more commonly set than others.
  + Some variables may exhibit outliers, particularly with values at the extreme ends of the spectrum.
* **Implication**:
  + Skewed distributions might require data transformations or binning to normalize them for better model performance.
  + Outliers may need further investigation or could be removed if they distort the model's results.  
      
    

**Figure 2: Distribution of Nominal Variables (Count Plots)**

* **Description**: This set of count plots shows the distribution of the categorical variables, including 'Gender', 'Activity Level', 'Dietary Preference', and 'Disease'. The count plots illustrate how many instances there are for each category in these variables.
* **Observation**:
  + The 'Gender' variable shows a near-equal distribution of male and female participants.
  + Other variables like 'Activity Level' and 'Dietary Preference' may exhibit imbalances, with certain categories being more common than others.
  + The 'Disease' variable shows an uneven distribution, indicating that some diseases are underrepresented in the dataset.
* **Implication**:
  + Imbalanced categories in 'Disease' might lead to biased model predictions, so techniques like over-sampling or under-sampling may be considered to handle such imbalances.
  + For variables like 'Activity Level', where some categories are more frequent, it might be important to investigate if these biases reflect the real-world distribution or if they are dataset artifacts.

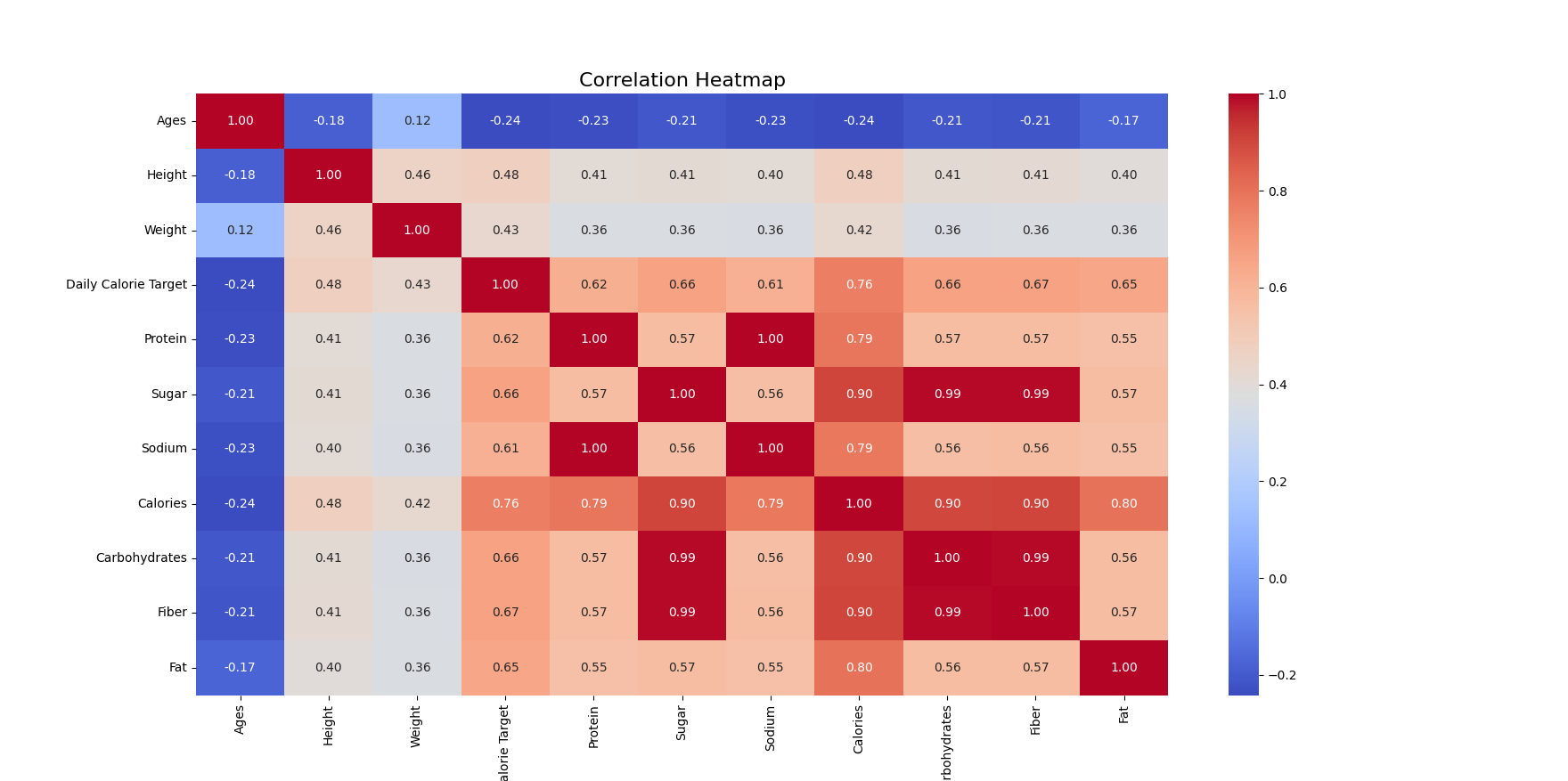






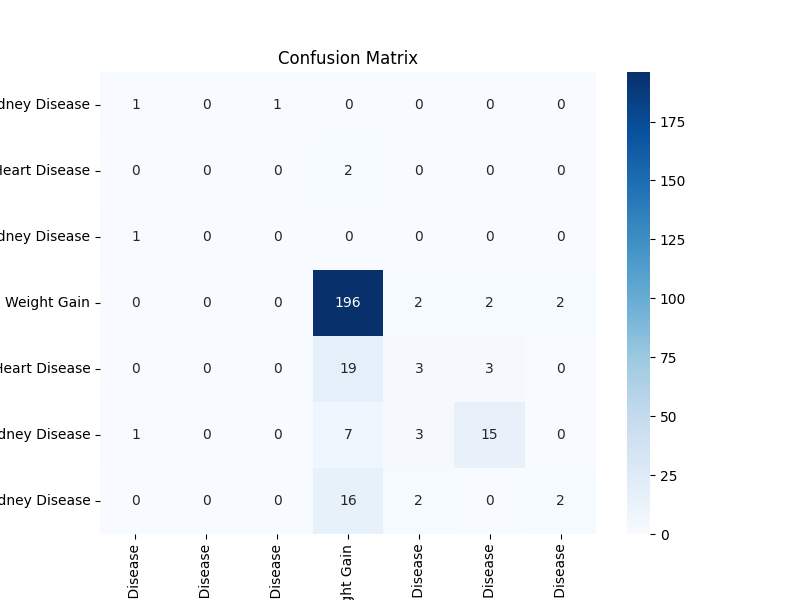
**Figure 3: Heatmap for Correlation**

* **Description**: This heatmap represents the correlation matrix for the numeric features in the dataset. It shows the pairwise correlations between variables, with color intensity indicating the strength of the relationship.
* **Observation**:
  + Some nutritional variables such as 'Calories', 'Carbohydrates', and 'Sugar' show strong correlations, suggesting that they may be related to one another in terms of dietary habits.
  + The correlation between 'Protein' and 'Fat' is moderate, indicating that they are often balanced together in diets.
  + 'Ages' and 'Height' show a weaker correlation, suggesting that height may not significantly change as individuals age within the given dataset range.
* **Implication**:
  + Highly correlated features can lead to multicollinearity, which can affect the performance of certain machine learning models. Dimensionality reduction or feature selection methods could be used to mitigate this.

Understanding these correlations is valuable for feature engineering, ensuring the model focuses on relevant variables.  
****

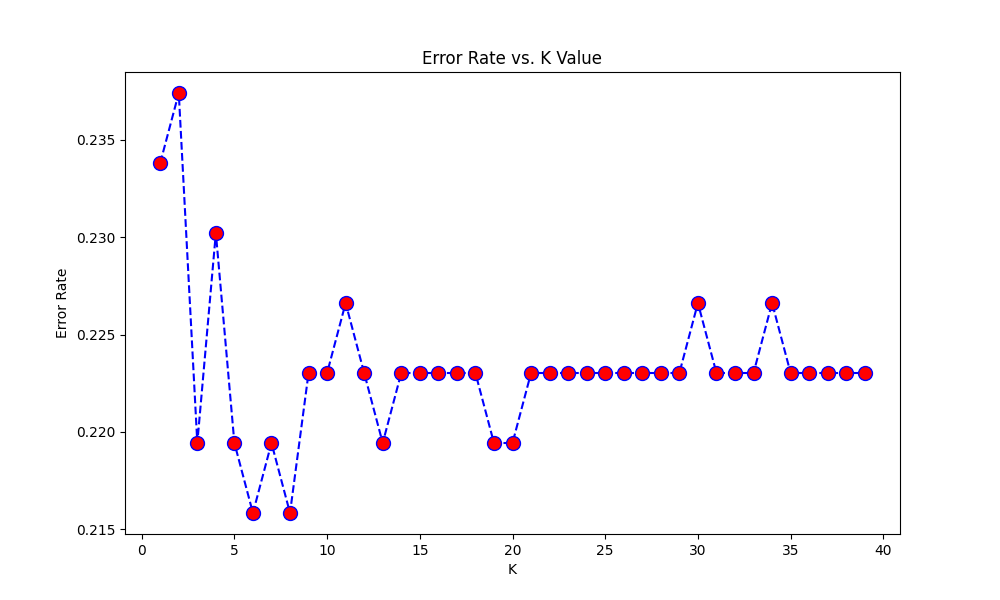
**Figure 4: Confusion Matrix**

* **Description**: This confusion matrix visualizes the model’s predictions versus the actual outcomes, showing the true positives, true negatives, false positives, and false negatives for each class.
* **Observation**:
  + The diagonal elements indicate correctly predicted disease categories.
  + Off-diagonal elements represent misclassifications, and the overall pattern gives insight into how well the model is distinguishing between different diseases.
* **Implication**:
  + A high number of off-diagonal elements may suggest that the model is struggling to differentiate between certain diseases. Further hyperparameter tuning, data augmentation, or model adjustments may be needed.



**Figure 5: Error Rate vs. K Value (Elbow Method)**

* **Description**: This plot shows how the error rate of the KNN model changes with different values of K (the number of neighbors). The 'elbow' point is typically used to identify the optimal K value.
* **Observation**:
  + The plot shows that as the value of K increases, the error rate initially decreases and then starts to stabilize, which indicates that an optimal K value has been reached.
  + The plot highlights the best K value that minimizes the error rate.
* **Implication**:
  + The elbow method helps in selecting the best K value to prevent overfitting (for small values of K) or underfitting (for large values of K).
  + Choosing the right K is critical for optimizing the model’s performance and ensuring its generalization on unseen data.



**OUTPUT :**

**Class Distribution in Target Variable 'Disease'**

| **Class** | **Count** |
| --- | --- |
| Weight Gain | 503 |
| Weight Gain, Hypertension, Heart Disease, Kidney Disease | 66 |
| Weight Gain, Hypertension, Heart Disease | 62 |
| Weight Gain, Kidney Disease | 50 |
| Diabetes, Acne, Weight Gain, Hypertension, Heart Disease, Kidney Disease | 5 |
| Hypertension, Heart Disease | 4 |
| Hypertension, Heart Disease, Kidney Disease | 3 |
| Diabetes, Acne, Hypertension, Kidney Disease | 1 |
| Hypertension, Kidney Disease | 1 |
| Diabetes, Acne, Weight Loss, Hypertension, Heart Disease, Kidney Disease | 1 |
| Kidney Disease | 1 |

**Underrepresented Classes**

|  |
| --- |
| Diabetes, Acne, Hypertension, Kidney Disease |

**Updated Class Distribution**

| **Class** | **Count** |
| --- | --- |
| Weight Gain | 503 |
| Weight Gain, Hypertension, Heart Disease, Kidney Disease | 66 |
| Weight Gain, Hypertension, Heart Disease | 62 |
| Weight Gain, Kidney Disease | 50 |
| Diabetes, Acne, Weight Gain, Hypertension, Heart Disease, Kidney Disease | 5 |
| Hypertension, Heart Disease | 4 |
| Hypertension, Heart Disease, Kidney Disease | 3 |

**Classification Report (Initial KNN with K = 5)**

| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| Diabetes, Acne, Weight Gain, Hypertension, Heart Disease, Kidney Disease | 0.33 | 0.50 | 0.40 | 2 |
| Hypertension, Heart Disease | 1.00 | 0.00 | 0.00 | 2 |
| Hypertension, Heart Disease, Kidney Disease | 0.00 | 0.00 | 0.00 | 1 |
| Weight Gain | 0.82 | 0.97 | 0.89 | 202 |
| Weight Gain, Hypertension, Heart Disease | 0.30 | 0.12 | 0.17 | 25 |
| Weight Gain, Hypertension, Heart Disease, Kidney Disease | 0.75 | 0.58 | 0.65 | 26 |
| Weight Gain, Kidney Disease | 0.50 | 0.10 | 0.17 | 20 |

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 0.78 |
| Macro Avg (Precision, Recall, F1) | 0.53 / 0.32 / 0.33 |
| Weighted Avg (Precision, Recall, F1) | 0.74 / 0.78 / 0.74 |

**Best K Value**

| **Metric** | **Value** |
| --- | --- |
| Best K Value | 6 |

**Classification Report with Best K (K = 6)**

| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| Diabetes, Acne, Weight Gain, Hypertension, Heart Disease, Kidney Disease | 0.33 | 0.50 | 0.40 | 2 |
| Hypertension, Heart Disease | 1.00 | 0.00 | 0.00 | 2 |
| Hypertension, Heart Disease, Kidney Disease | 1.00 | 0.00 | 0.00 | 1 |
| Weight Gain | 0.81 | 0.98 | 0.89 | 202 |
| Weight Gain, Hypertension, Heart Disease | 0.40 | 0.16 | 0.23 | 25 |
| Weight Gain, Hypertension, Heart Disease, Kidney Disease | 0.71 | 0.58 | 0.64 | 26 |
| Weight Gain, Kidney Disease | 0.50 | 0.05 | 0.09 | 20 |

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 0.784 |
| Macro Avg (Precision, Recall, F1) | 0.68 / 0.32 / 0.32 |
| Weighted Avg (Precision, Recall, F1) | 0.74 / 0.78 / 0.73 |

**Accuracy with Best K**

| **Metric** | **Value** |
| --- | --- |
| Accuracy with Best K | 0.784 |

**Conclusion:**

The analysis of the dataset has provided several insights into the nutritional and disease-related features. Missing values have been effectively handled, and visualizations have uncovered the distribution and relationships between various features. The KNN model, after optimizing the K value using the elbow method, offers a promising approach for disease prediction. The confusion matrix and classification report further highlight areas where the model performs well, though the class imbalance issue remains a challenge that may need to be addressed for improved accuracy. Further exploration and refinement of the dataset and model will be crucial for building a more robust and reliable prediction system.