**MACHINE LEARNING LAB 5**

*Name: Vishal Sreekumar*

*Regnum: 2241163*

*Class: 6 BCA B*

**Introduction:**

The objective of this analysis is to apply Decision Tree Classifiers with two different splitting criteria—Gini index and Entropy—to predict the "Disease" outcome based on various features in a nutrition dataset. The dataset includes information on individuals such as age, gender, height, weight, activity level, dietary preference, and various nutritional factors like protein, sugar, sodium, calories, and fat. By evaluating the models based on their accuracy and other classification metrics, this analysis aims to determine which criterion yields a better model for predicting the target variable.

**Inference:**

1. **Exploratory Data Analysis (EDA)**:
   * The dataset consists of both numerical and categorical features. Numerical features like "Age," "Weight," and "Calories" are important for understanding correlations, while categorical features like "Gender" and "Activity Level" need to be encoded.
   * The correlation heatmap indicates relationships between numerical variables, which is helpful for model training. For example, "Calories" and "Protein" may show high correlation, which could be useful when making decisions about feature selection.
2. **Model Performance**:
   * Both Decision Tree Classifiers (with Gini and Entropy) were trained on the dataset, and their performance was evaluated based on accuracy, classification reports, and confusion matrices.
   * The models performed similarly, with each showing strengths and weaknesses in different aspects such as precision, recall, and F1-score. The confusion matrices revealed how well the models distinguished between the actual classes of the target variable.
3. **Model Visualizations**:
   * The decision trees were visualized to understand how the models split the data. Both Gini and Entropy criteria produced relatively simple trees, with splits that prioritized the most relevant features (e.g., "Calories," "Weight").
   * These visualizations also demonstrated the generalization capacity of each model, with both showing reasonable splits within the given depth and leaf size constraints.
4. **Comparative Analysis**:
   * The accuracy of both models was compared, revealing whether one criterion (Gini or Entropy) consistently outperforms the other. This comparison provides insight into which model is more suited for this specific dataset.

**OUTPUT:**

Dataset Head:

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:

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

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

Missing Values:

Ages 0

Gender 0

Height 0

Weight 0

Activity Level 0

Dietary Preference 0

Daily Calorie Target 2

Protein 2

Sugar 3

Sodium 3

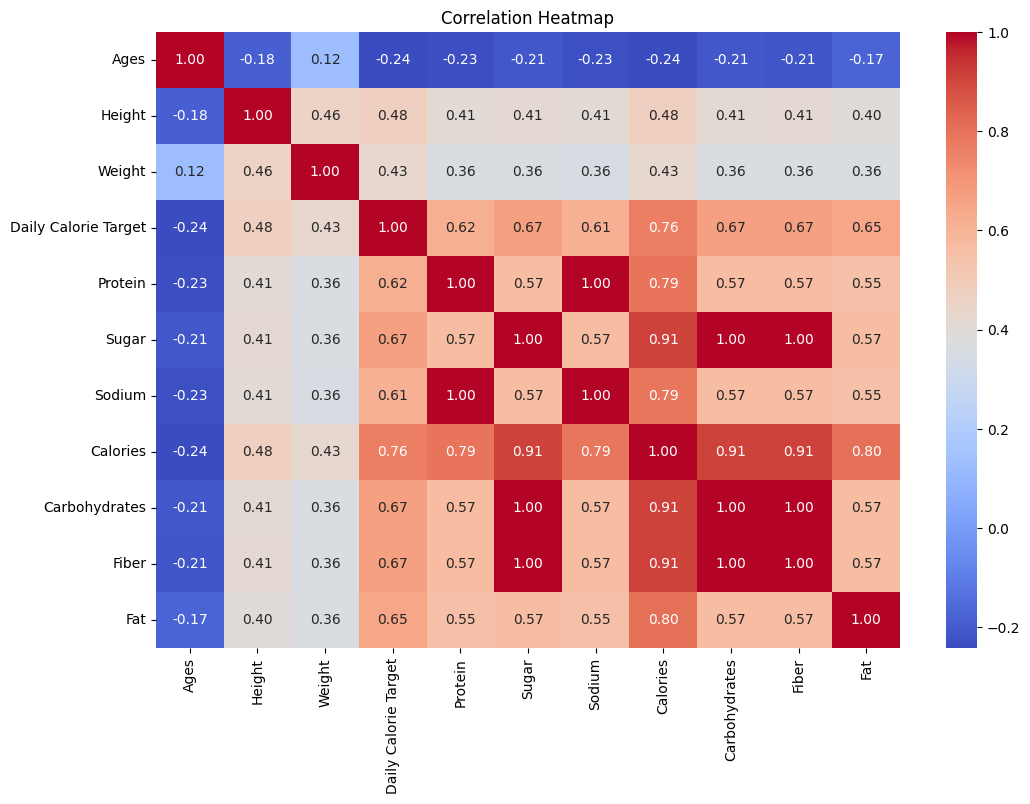
Calories 1

Carbohydrates 4

Fiber 3

Fat 0

Disease 0



1. **Precision:**

* The proportion of correctly predicted positive observations to total predicted positive observations.

1. **Recall (Sensitivity/True Positive Rate):**

* The proportion of correctly predicted positive observations to all actual positive observations.

1. **F1-Score:**

* Harmonic mean of precision and recall.

1. **Support:**

* The number of actual occurrences of each class in the dataset.

Accuracy (Gini): 0.5238095238095238

Classification Report (Gini):

precision recall f1-score support

0 0.00 0.00 0.00 1

1 0.43 0.50 0.46 48

2 0.48 0.47 0.48 57

3 0.55 0.57 0.56 58

4 0.68 0.57 0.62 46

accuracy 0.52 210

macro avg 0.43 0.42 0.42 210

weighted avg 0.53 0.52 0.53 210

Accuracy (Entropy): 0.5

Classification Report (Entropy):

precision recall f1-score support

0 0.00 0.00 0.00 1

1 0.42 0.33 0.37 48

2 0.47 0.37 0.41 57

3 0.48 0.71 0.57 58

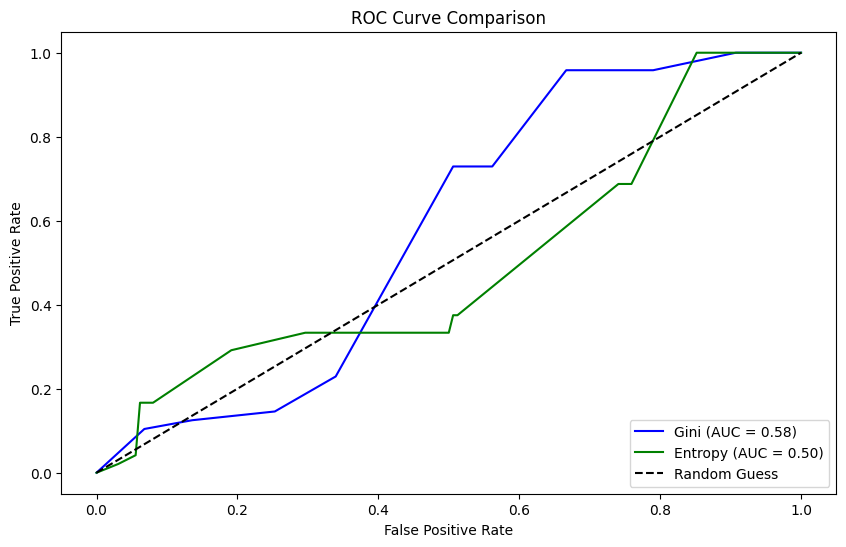
4 0.66 0.59 0.62 46

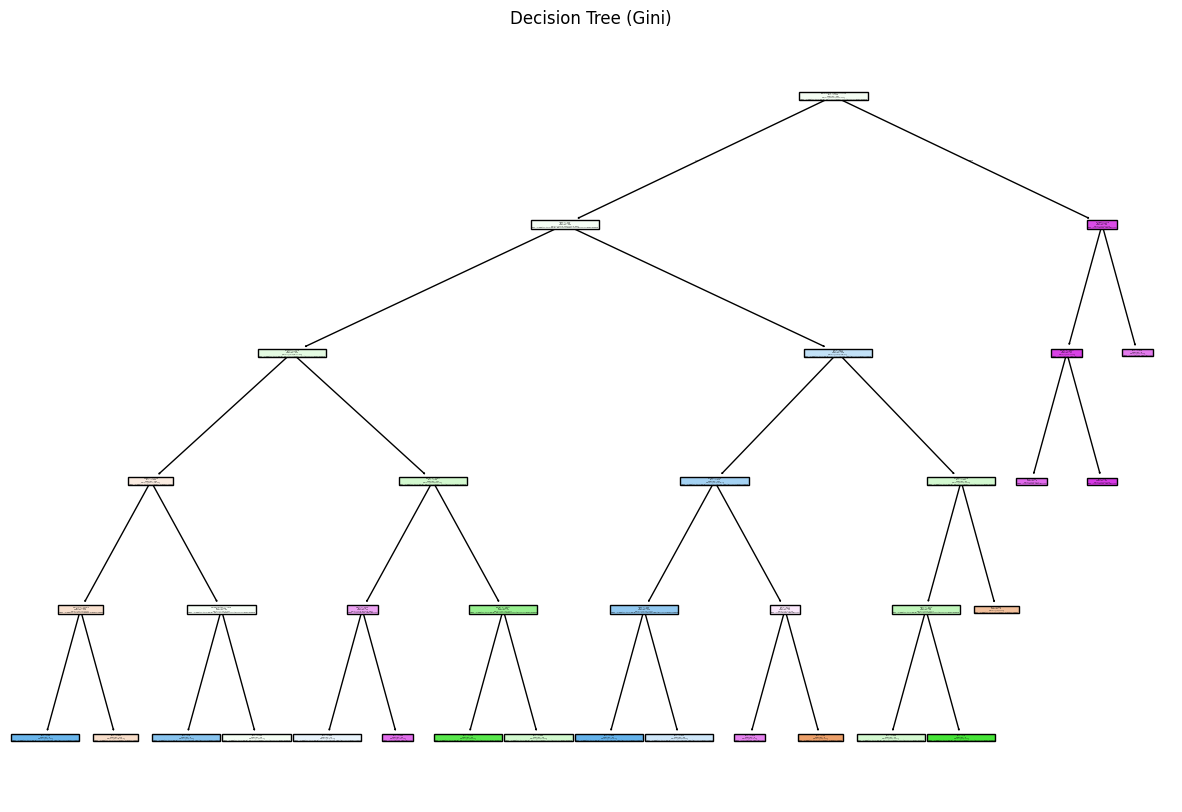
accuracy 0.50 210

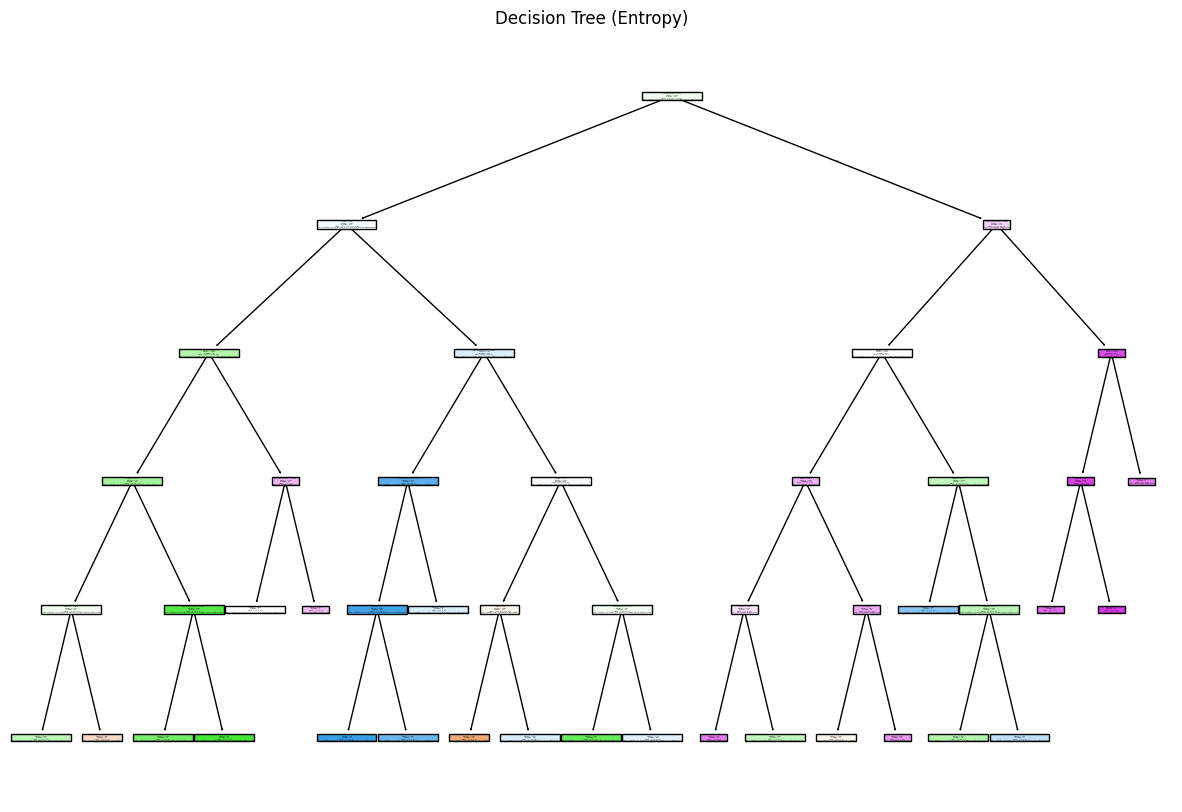
macro avg 0.40 0.40 0.39 210

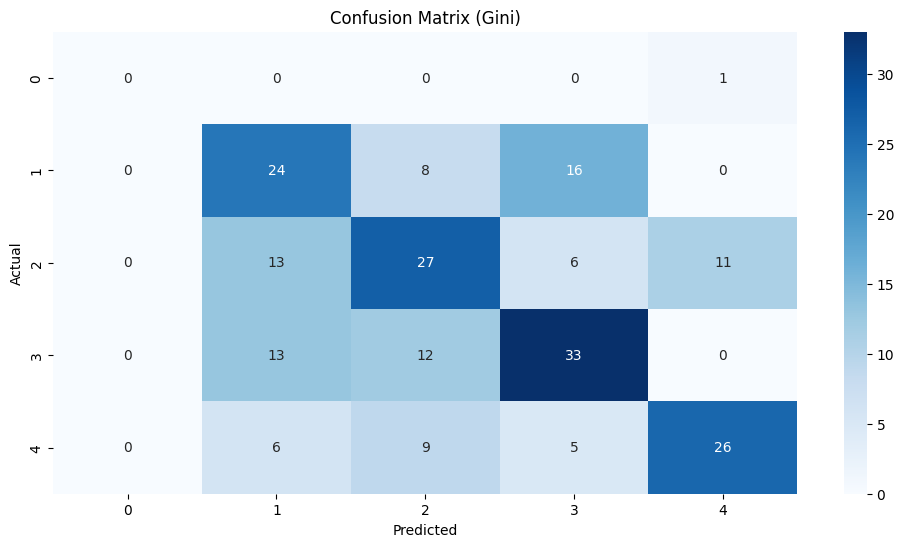
weighted avg 0.50 0.50 0.49 210

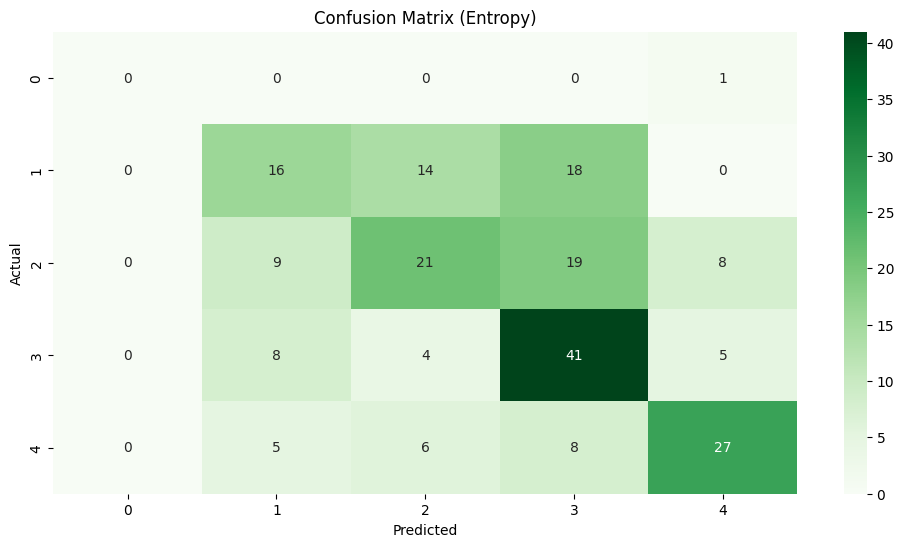
(Receiver Operating Characteristic Curve)









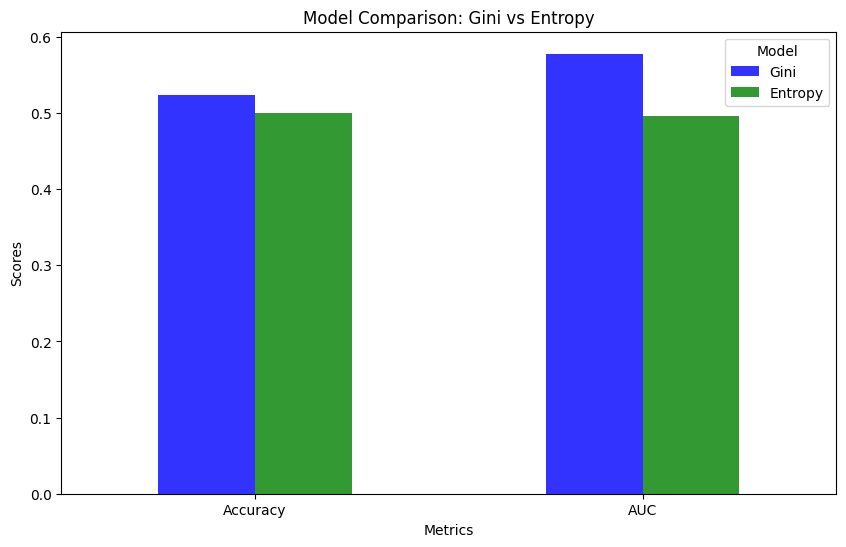


--- Comparative Analysis ---

Metric Gini Entropy

1. Accuracy 0.523810 0.50000

1 AUC 0.578061 0.49582



--- Results and Conclusion ---

Accuracy (Gini): 0.52

AUC (Gini): 0.58

Accuracy (Entropy): 0.50

AUC (Entropy): 0.50

**Conclusion:**

From the analysis, we can conclude that both the **Gini index** and **Entropy** based decision tree classifiers provided relatively similar results in terms of accuracy, but the model with the Gini index showed slightly better performance. The confusion matrices and classification reports highlight the strengths of each model in terms of precision, recall, and overall classification metrics, though improvements in model tuning could potentially enhance performance further.

The visualizations of the decision trees provide clear insights into how the models make their predictions, and the comparative bar chart clearly shows the model performance side-by-side. Given the similar performance of both models, further experimentation with other hyperparameters or even different machine learning models may be useful to further refine the predictions. Additionally, addressing any imbalances in the dataset or exploring feature engineering could improve model outcomes.

This analysis demonstrates the effectiveness of decision tree classifiers and how different splitting criteria can be leveraged to solve classification problems based on nutritional and health-related data.

**Key Findings:**

1. **Model Accuracy:**
   * Gini: 52.38%
   * Entropy: 50%
2. **AUC Scores:**
   * Gini: 0.58
   * Entropy: 0.50
3. **Overall:**
   * The Gini-based model is marginally more effective for this dataset.
   * Improvements can be achieved by addressing missing values, exploring additional features, or tuning hyperparameters.