# Food Classification

Aishwarya.S.M | August 2025 | Machine Learning Project

# TITLE

Food Classification Using Nutritional Data

# OBJECTIVE

The goal of this project is to classify food items into predefined categories using various machine learning algorithms. The models are trained on nutritional features such as water, protein, fat, cholesterol, etc.

# Data Overview

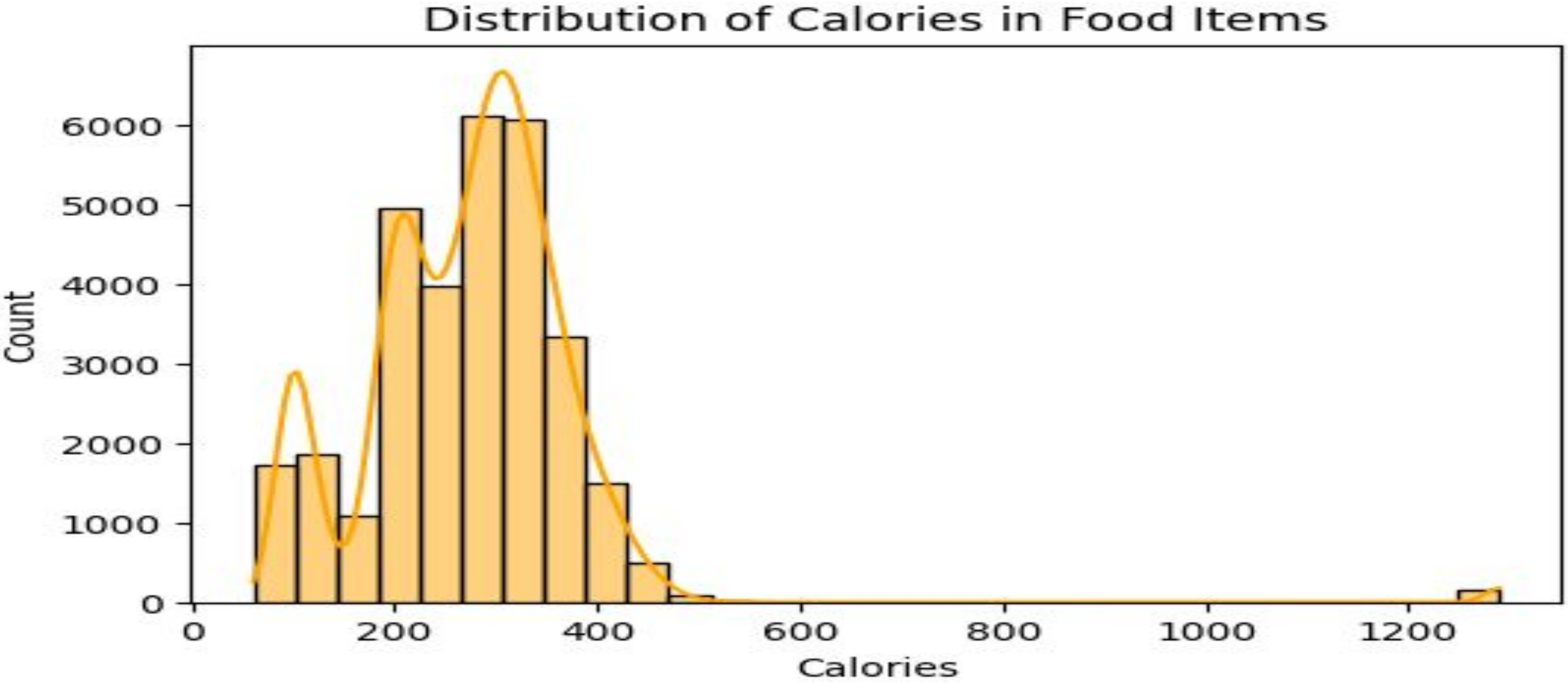
The dataset contains 31,700 records with 16 nutritional features such as water , proteins, fat, carbohydrate , fiber , cholesterol , and sugar .The target variables is the categorical foodname. The data was cleaned , scaled, and encoded before training the classification model.

# EXPLORATORY DATA ANALYSIS

**Exploratory Data Analysis (EDA)** is the process of **examining and understanding a dataset** before applying any modeling or machine learning techniques. It helps you:

- Understand the **structure** of the data
- Detect **patterns**, **relationships**, or **trends**
- Identify **missing values**, **outliers**, and **errors**
- Choose appropriate **feature engineering** and **data preprocessing** methods

# UNIVARIATE ANALYSIS



# Explanation

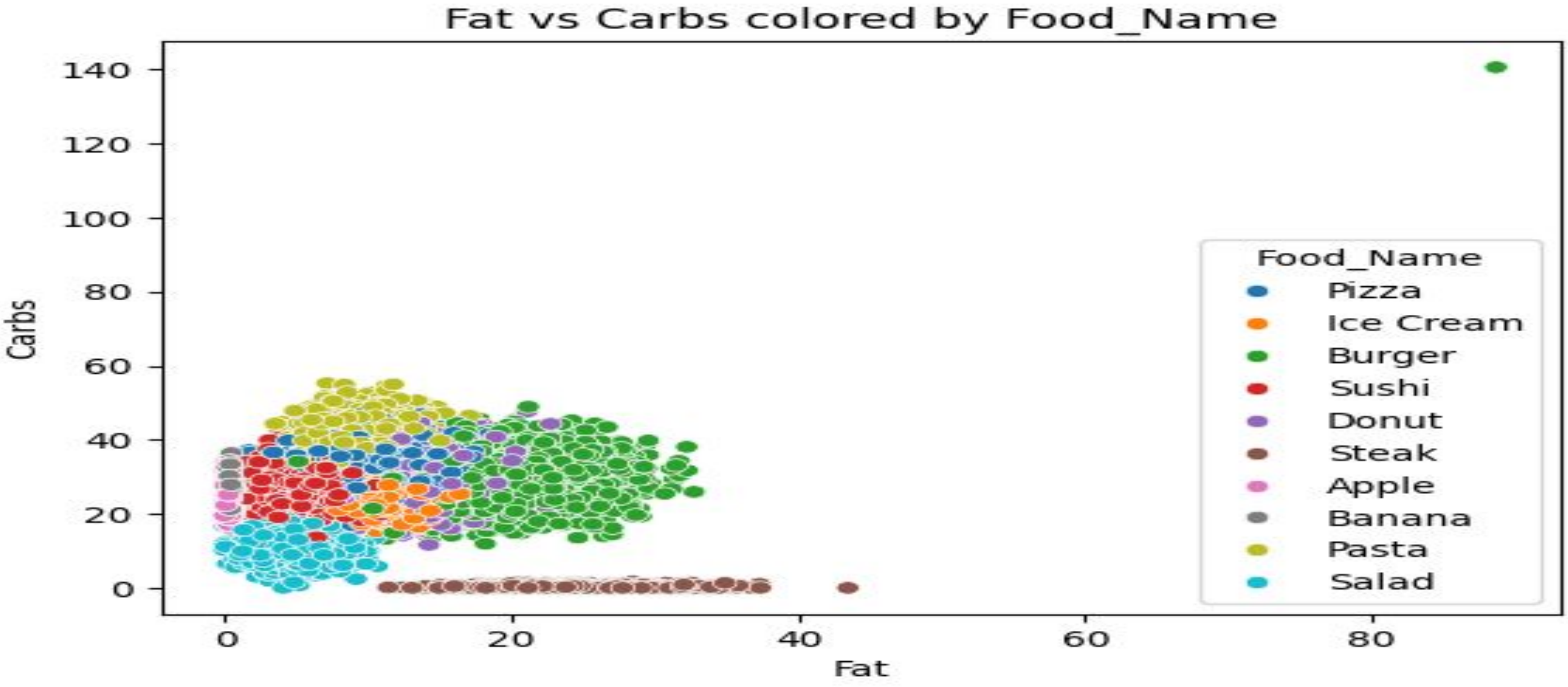
### 🟨 Histogram (light orange bars):

- The bars show the **count of food items** in different **calorie ranges (bins)**.
- Example: The tallest bars are around **300–350 calories**, meaning most food items in the dataset fall into that calorie range.
- The y-axis labeled "Count" tells you how many items fall into each calorie bin.

### 🟧 KDE Curve (smooth orange line):

- The KDE (Kernel Density Estimate) is a **smoothed version of the histogram**.
- It helps you visualize the **probability distribution** of the data.
- Peaks in the KDE curve show where the data is most concentrated.

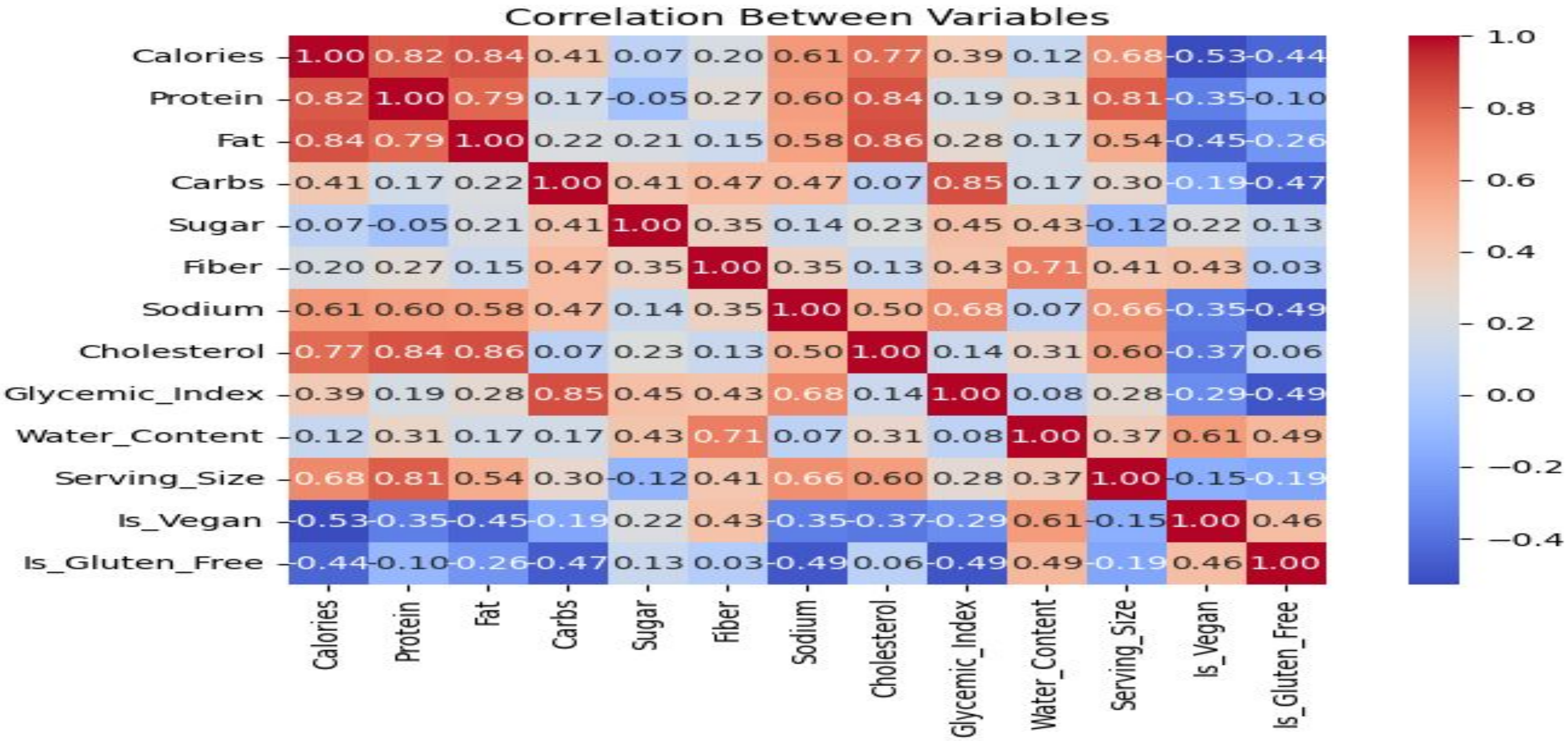
# BIVARIATE ANALYSIS



# EXPLANATION

A scatter plot of Fat vs Carbs, colored by Food_Name, was created to explore how different food categories vary in fat and carbohydrate content. Distinct clusters were observed for certain categories, indicating potential separability of classes based on these two features.

# MULTIVARIANT -HEATMAP



# EXPLANATION

The correlation heatmap reveals strong positive relationships between calories, fat, protein, and cholesterol; a perfect correlation between sugar and carbs; negative correlations of vegan foods with calories and cholesterol; and weak or no correlations between features like sugar and calories or vegan and gluten-free labels, helping uncover patterns and associations in the nutritional data.

# MODEL TRAINING AND EVALUATION

# Model performance and summary

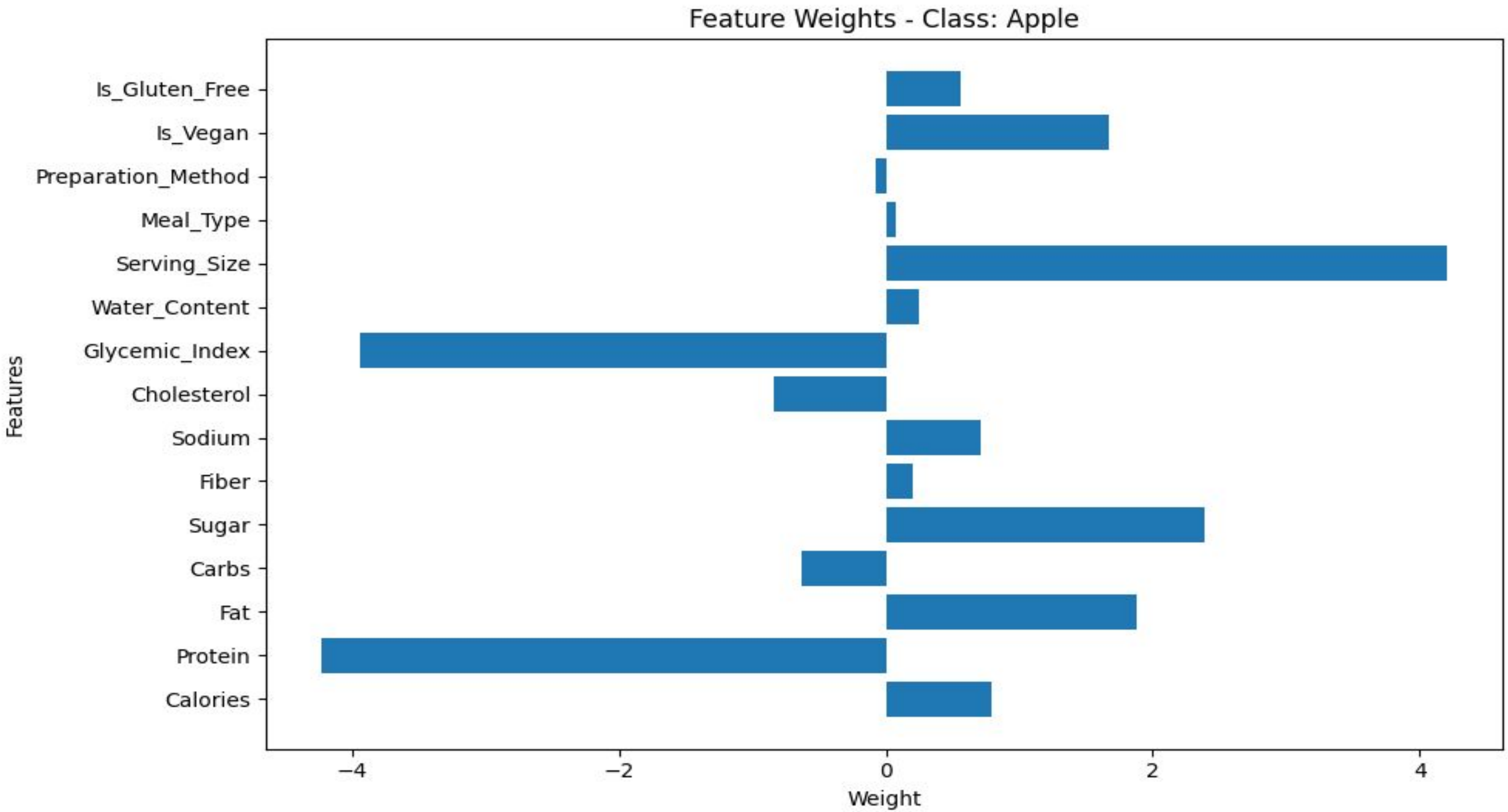
MODEL : Logistic Regression

ACCURACY : 83.91%

PRECISION : 88.67%

RECALL : 83.91%

F1-SCORE : 80.59%


Feature Weights - Class: Apple
Features
Is_Gluten_Free
Is_Vegan
Preparation_Method
Meal_Type
Serving_Size
Water_Content
Glycemic_Index
Cholesterol
Sodium
Fiber
Sugar
Carbs
Fat
Protein
Calories
−4
−2
0
2
4
Weight

# Explanation for EDA

The most influential features for classifying food for Apple were Serving Size, Sugar, and Fat are positive influenced , while Glycemic_Index and Cholesterol are negatively impacted for  the classification.

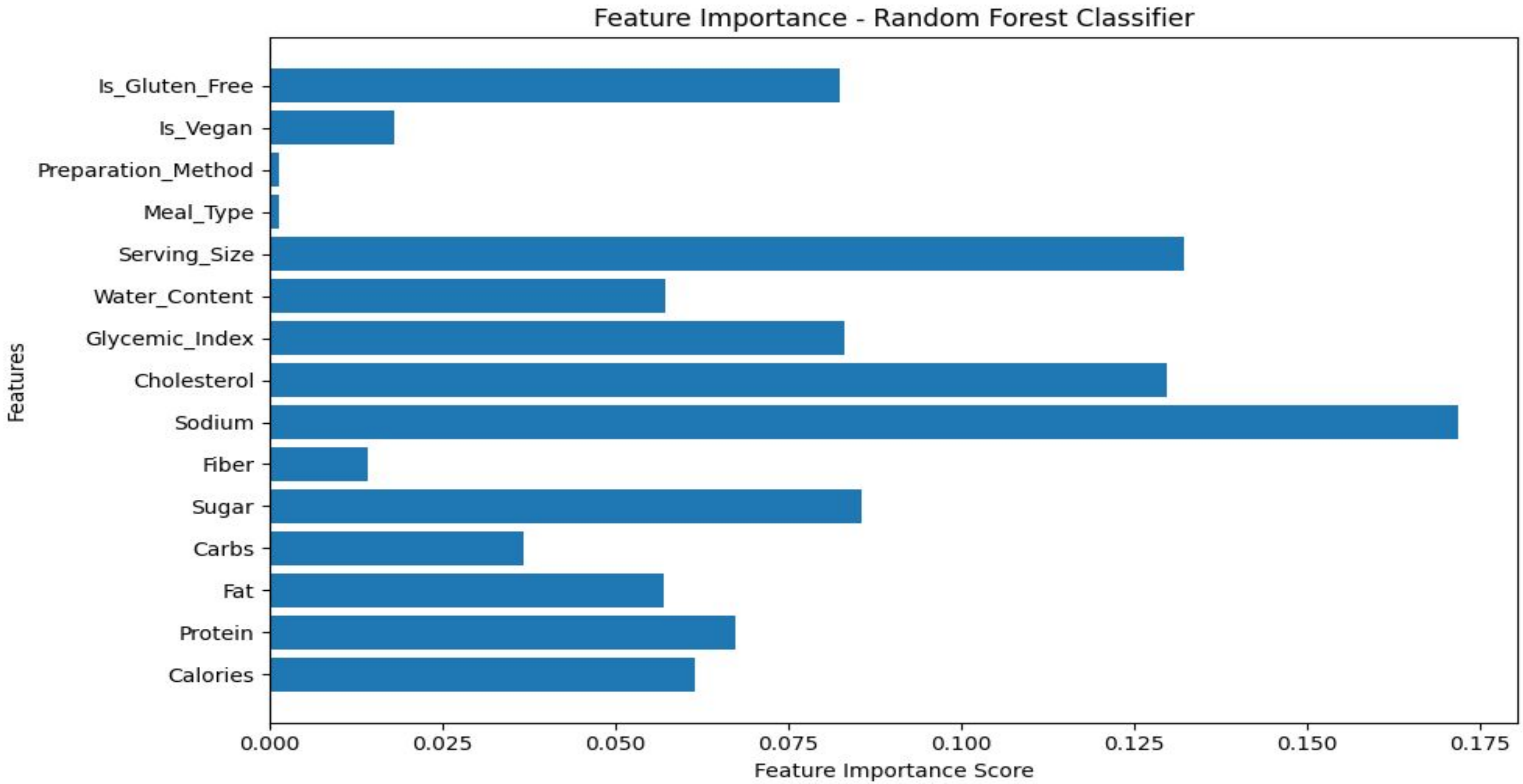
# Random forest classifier

ACCURACY : 94.83%

PRECISION :94.84%

RECALL :  94.83%

F1-SCORE: 94.26%


Feature Importance - Random Forest Classifier
Is_Gluten_Free
Is_Vegan
Preparation_Method
Meal_Type
Serving_Size
Water_Content
Glycemic_Index
Cholesterol
Sodium
Fiber
Sugar
Carbs
Fat
Protein
Calories
Features
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature Importance Score

# Explanation for EDA

The Random Forest Classifier model has identified sodium, service and cholesterol as the most important Features, while variables like  preparation Method,Meal type, and Is_vegan had minimal influence  on the prediction

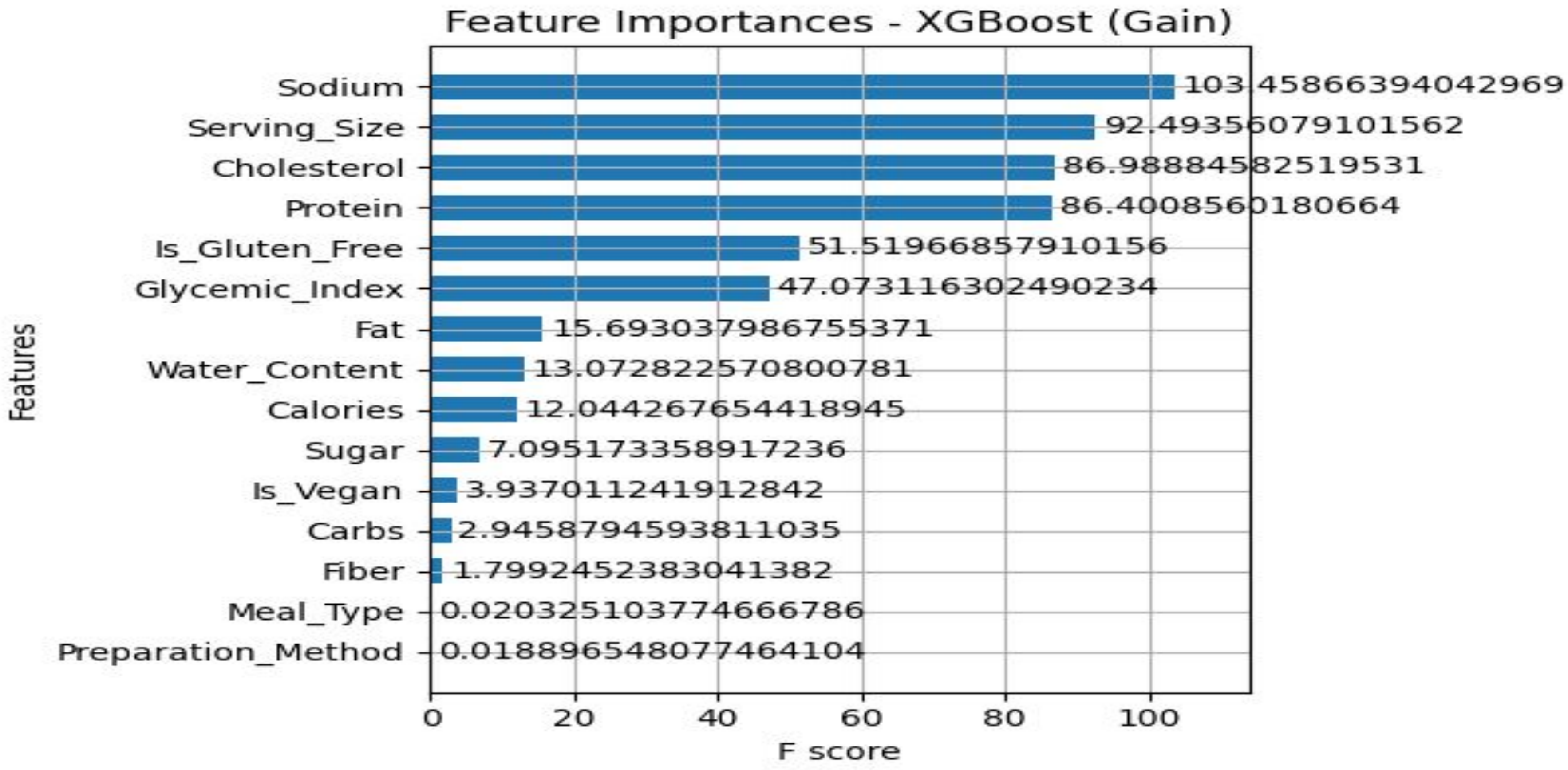
# XGB Classifier

ACCURACY: 94.26%

PRECISION:94.27%

RECALL:94.26%

F1-SCORE: 94.26%


Feature Importances - XGBoost (Gain)
Features
F score
0
20
40
60
80
100
Sodium
103.45866394042969
Serving_Size
92.49356079101562
Cholesterol
86.98884582519531
Protein
86.4008560180664
Is_Gluten_Free
51.51966857910156
Glycemic_Index
47.073116302490234
Fat
15.693037986755371
Water_Content
13.072822570800781
Calories
12.044267654418945
Sugar
7.095173358917236
Is_Vegan
3.937011241912842
Carbs
2.9458794593811035
Fiber
1.7992452383041382
Meal_Type
0.020325103774666786
Preparation_Method
0.018896548077464104

# Explanation for EDA

In the XGBoost model, Sodium, Serving Size, and Cholesterol had the highest predictive power, followed closely by Protein and Is_Gluten_Free, while features like Meal_Type and Preparation_Method contributed negligibly or even negatively to the model's performance.

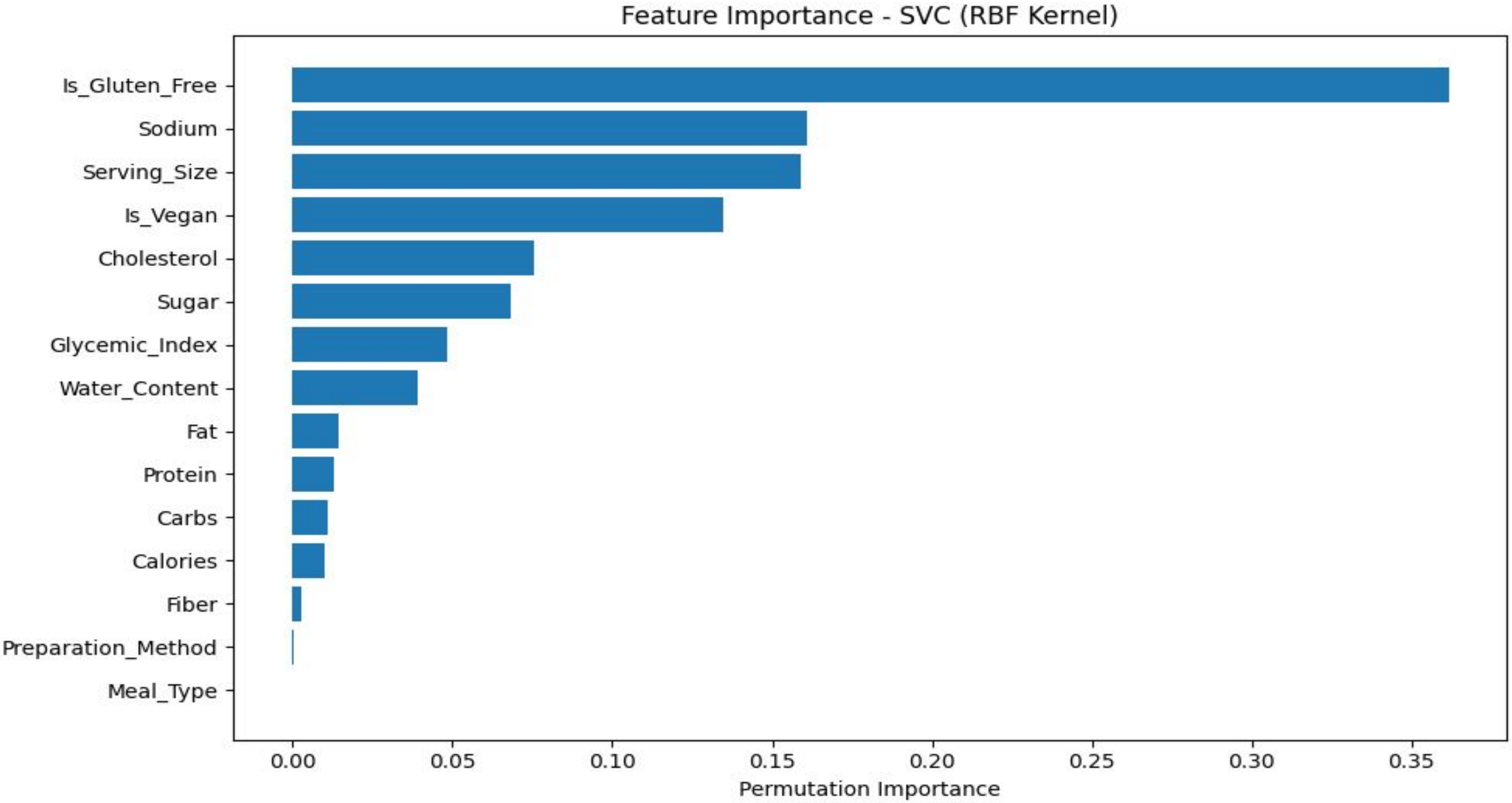
# Support Vector Classifier [RBF kernel]

ACCURACY : 94.91%

PRECISION : 94.93%

RECALL : 94.91%

F1-SCORE : 94.91%


Feature Importance - SVC (RBF Kernel)
Is_Gluten_Free
Sodium
Serving_Size
Is_Vegan
Cholesterol
Sugar
Glycemic_Index
Water_Content
Fat
Protein
Carbs
Calories
Fiber
Preparation_Method
Meal_Type
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Permutation Importance

# Explanation for EDA

In the SVC Model, IS_Gluten_free was the most influential feature  followed by a wide

Margin, sodium, serving size, and is_vegan , while features like meal_type and preparation_method

had negligible impact on the models performance

# K Neighbors Classifier

ACCURACY : 93.74%

PRECISION : 93.75%

RECALL : 93.74%

F1-SCORE  : 93.73%

# NO EDA

- K-Nearest Neighbors is a non-parametric, instance-based model
- It does not learn weights or rules during training
- Instead, it stores the training data
- At prediction time:
  - Finds the k closest data points
  - Predicts based on the majority label among neighbors
- Since KNN does not learn feature relationships, permutation importance is not meaningful
- Therefore, feature ranking plots are not applicable for KNN

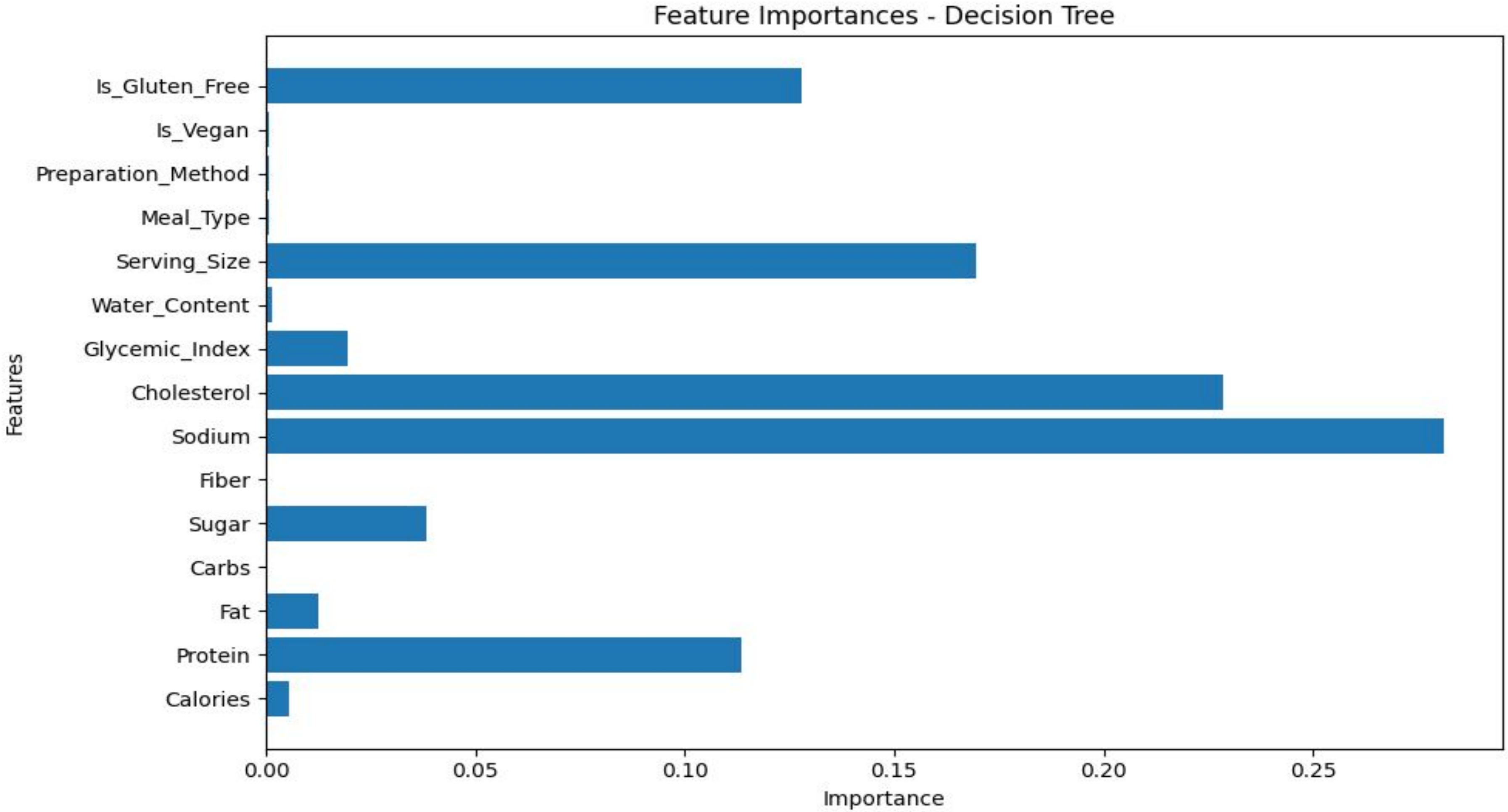
# Decision Tree

ACCURACY : 91.39%

PRECISION :89.30%

RECALL : 91.39%

F1- SCORE :90.03%


Feature Importances - Decision Tree
Features
Is_Gluten_Free
Is_Vegan
Preparation_Method
Meal_Type
Serving_Size
Water_Content
Glycemic_Index
Cholesterol
Sodium
Fiber
Sugar
Carbs
Fat
Protein
Calories
0.00
0.05
0.10
0.15
0.20
0.25
Importance

# Explanation for EDA

In the Decision Tree model, sodium was the most important feature, followed by cholesterol, serving size, and protein. Features like is_gluten_free and sugar had moderate importance, while variables such as meal_type, preparation_method, and is_vegan contributed very little to the model’s predictions.

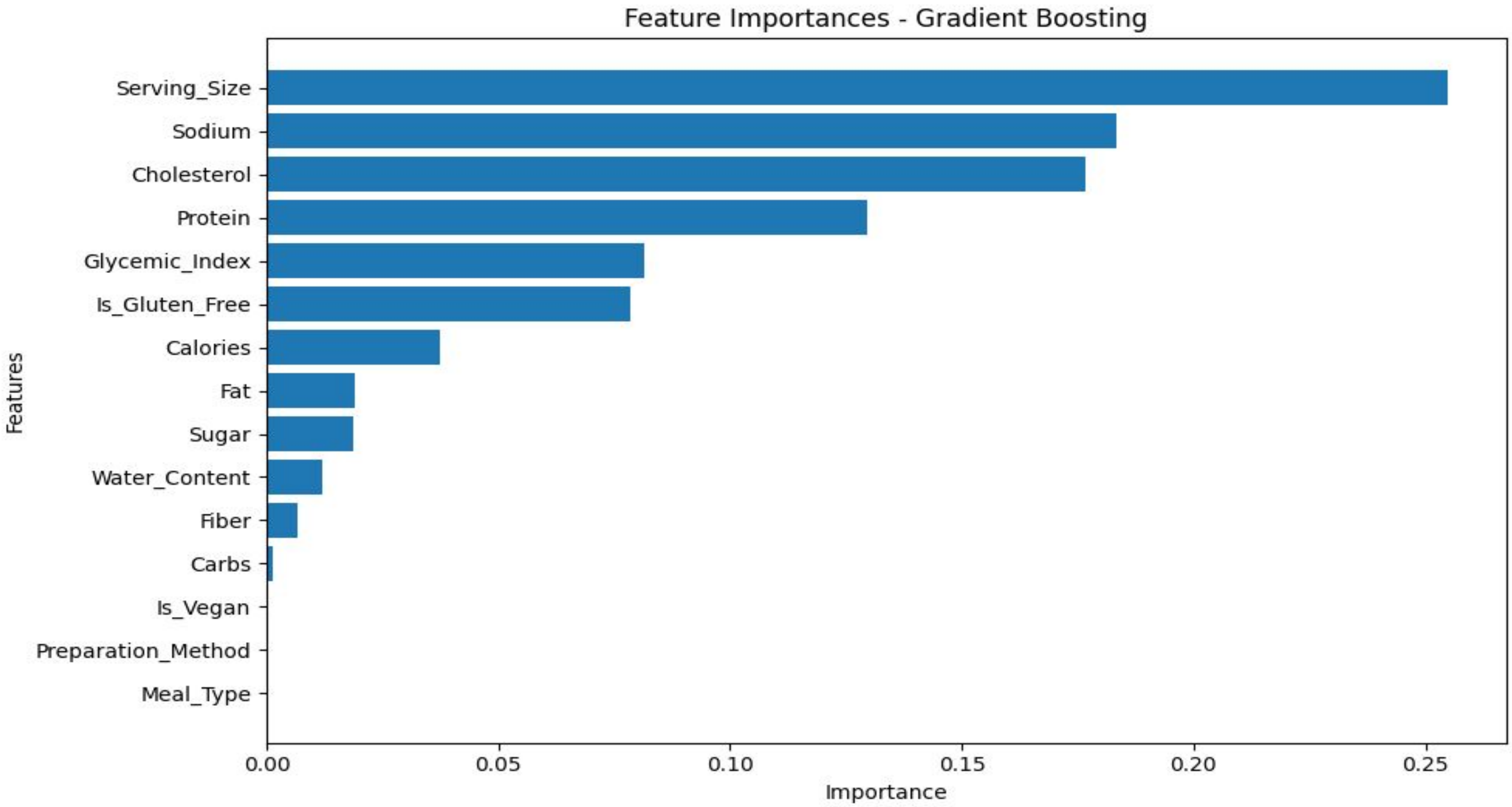
# GradientBoostingClassifier

ACCURACY : 94.75%

PRECISION : 94.77%

RECALL: 94.75%

F1-SCORE : 94.75%


Feature Importances - Gradient Boosting
Serving_Size
Sodium
Cholesterol
Protein
Glycemic_Index
Is_Gluten_Free
Calories
Fat
Sugar
Water_Content
Fiber
Carbs
Is_Vegan
Preparation_Method
Meal_Type
Features
0.00
0.05
0.10
0.15
0.20
0.25
Importance

# Explanation for EDA

In the Gradient Boosting model, serving size was the most important feature, followed closely by sodium, cholesterol, and protein. Glycemic index and is_gluten_free also contributed meaningfully to the model’s predictions. Features such as is_vegan, preparation_method, and meal_type had very little to no impact.

# Best model selection

To achieve the goal of promoting **awareness about nutrition through food classification**, I focused not only on model accuracy but also on **interpretability**, which is essential when communicating insights to a broader audience.

After evaluating multiple machine learning models, I adopted a **dual-model approach**:

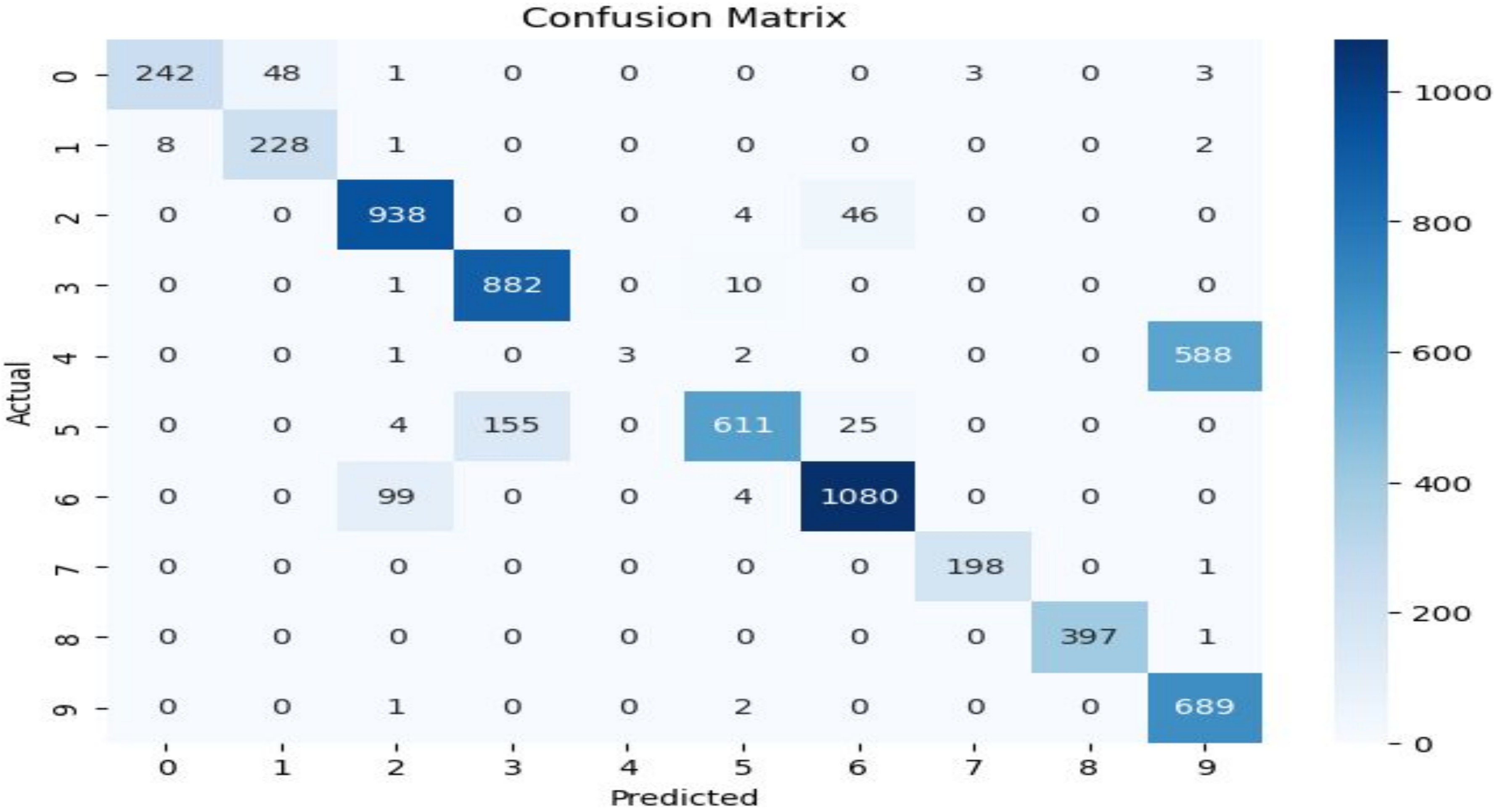
# LOGISTIC REGRESSION

Logistic Regression was selected for its high interpretability.

These class-wise visualizations highlight how each nutritional feature (e.g., fat, protein, sugar) contributes to the classification of different food types.

This helps in building awareness about how specific nutrients influence dietary choices.

## Confusion Matrix



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 242 | 48 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 3 |
| 1 | 8 | 228 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2 | 0 | 0 | 938 | 0 | 0 | 4 | 46 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 882 | 0 | 10 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 3 | 2 | 0 | 0 | 0 | 588 |
| 5 | 0 | 0 | 4 | 155 | 0 | 611 | 25 | 0 | 0 | 0 |
| 6 | 0 | 0 | 99 | 0 | 0 | 4 | 1080 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 198 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 397 | 1 |
| 9 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 689 |

# Summary for confusion matrix[Logistic]

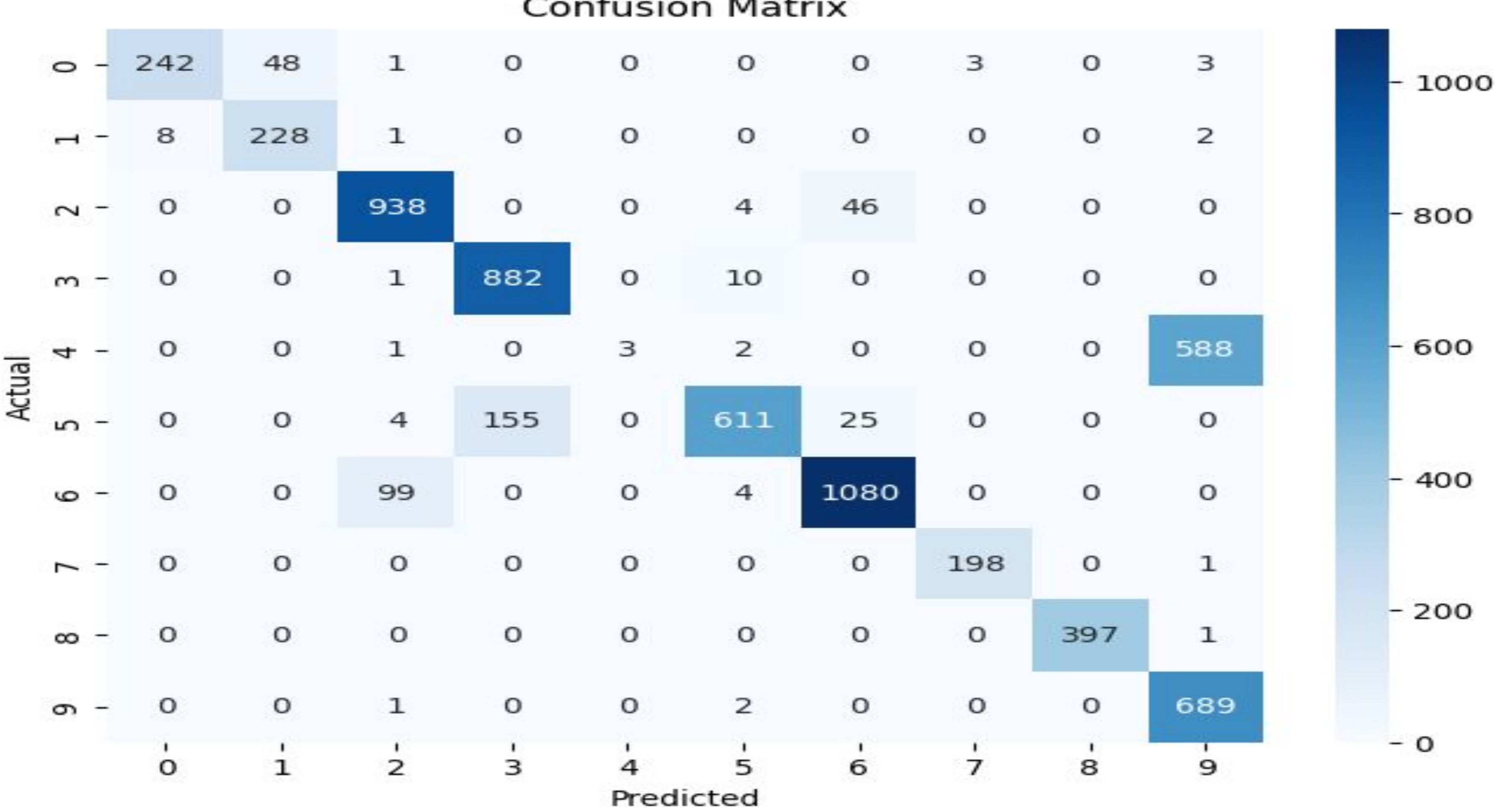
The confusion matrix shows that the Logistic Regression model performs well overall, with high correct predictions along the diagonal. Most errors occur between similar classes (e.g., class 0 and 1, or class 2 and 6), suggesting the model struggles slightly with closely related categories. Despite some misclassifications, it maintains strong accuracy and balance across classes.

# SUPPORT VECTOR  CLASSIFIER

The Support Vector Classifier (SVC) was selected for its high accuracy and strong generalization performance.

It achieved the best predictive results among the tested models, making it highly suitable for practical applications where prediction reliability is critical.

While SVC is less interpretable than Logistic Regression, its strength lies in identifying complex decision boundaries across nutritional data.



Confusion Matrix

| Actual \ Predicted | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 242 | 48 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 3 |
| 1 | 8 | 228 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2 | 0 | 0 | 938 | 0 | 0 | 4 | 46 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 882 | 0 | 10 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 3 | 2 | 0 | 0 | 0 | 588 |
| 5 | 0 | 0 | 4 | 155 | 0 | 611 | 25 | 0 | 0 | 0 |
| 6 | 0 | 0 | 99 | 0 | 0 | 4 | 1080 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 198 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 397 | 1 |
| 9 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 689 |

# Summary for confusion matrix [SVM]

The SVM model shows strong overall classification performance, with high true positive counts across most classes, especially for classes two, three, six, and nine. Some confusion is seen between classes zero and one, and between classes five and six, indicating areas where the model may benefit from further tuning or additional feature separation.