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.
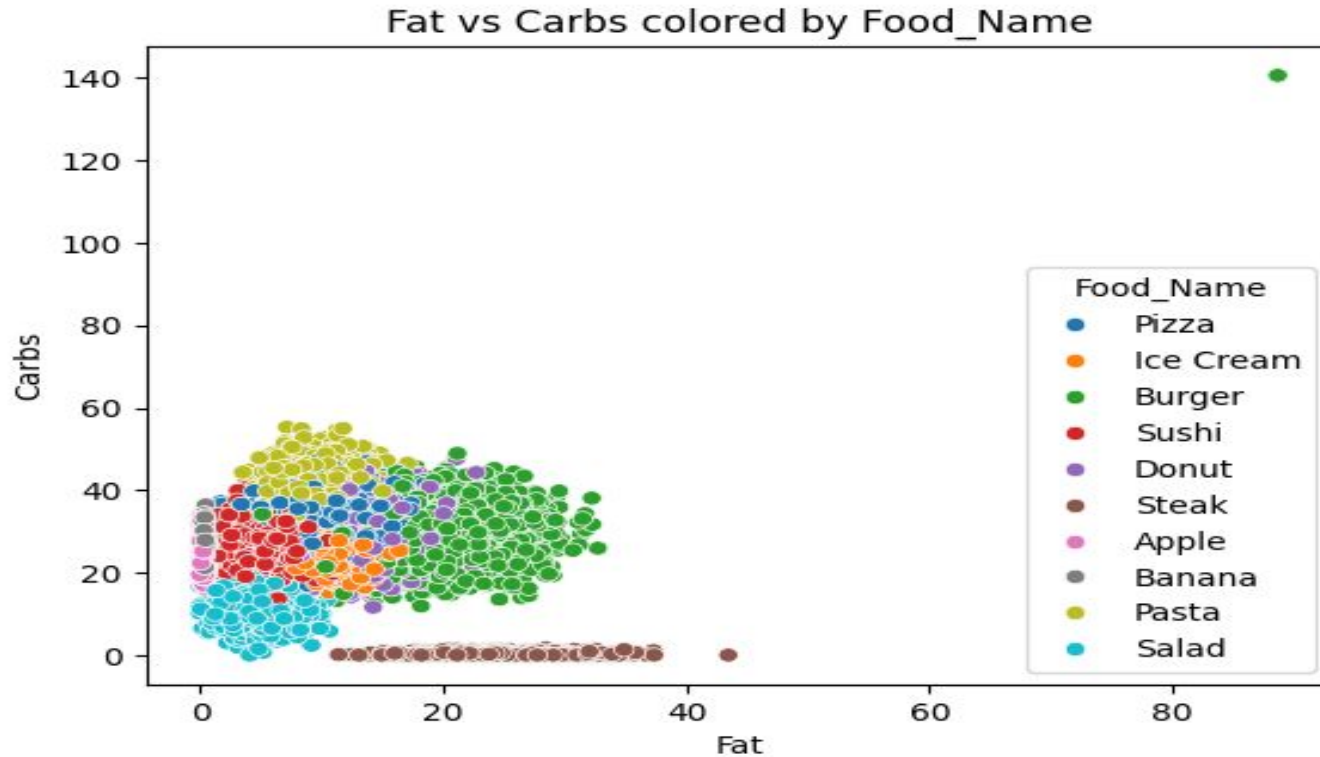
Data Overview

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 (EDA)

Fat vs Carbs colored by Food_Name

MODEL TRAINING AND EVALUATION

Model performance and summary

MODEL : Logistic Regression
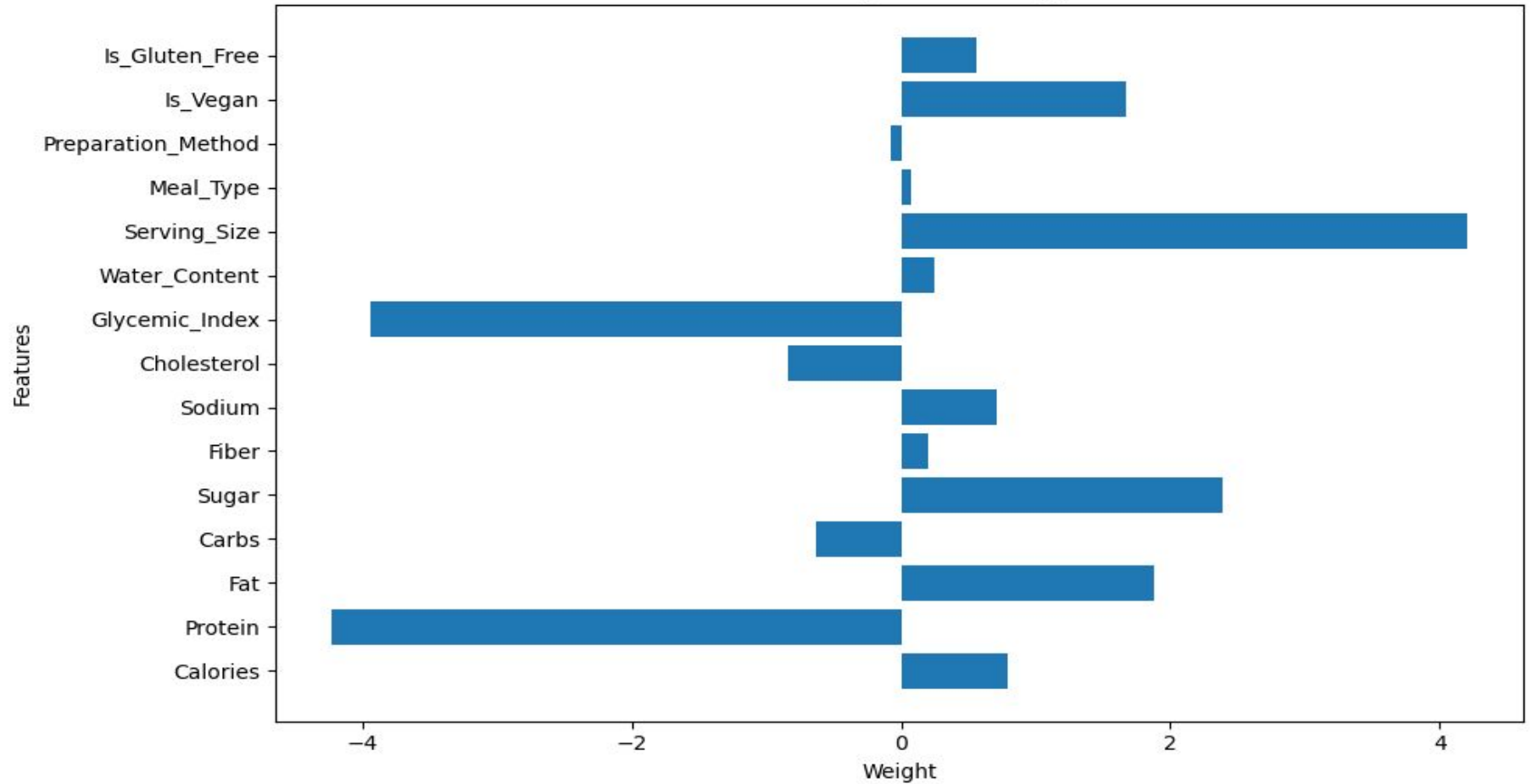
ACCURACY : 83.91%

PRECISION : 88.67%

RECALL : 83.91%

F1-SCORE : 80.59%

Feature Weights - Class: Apple

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
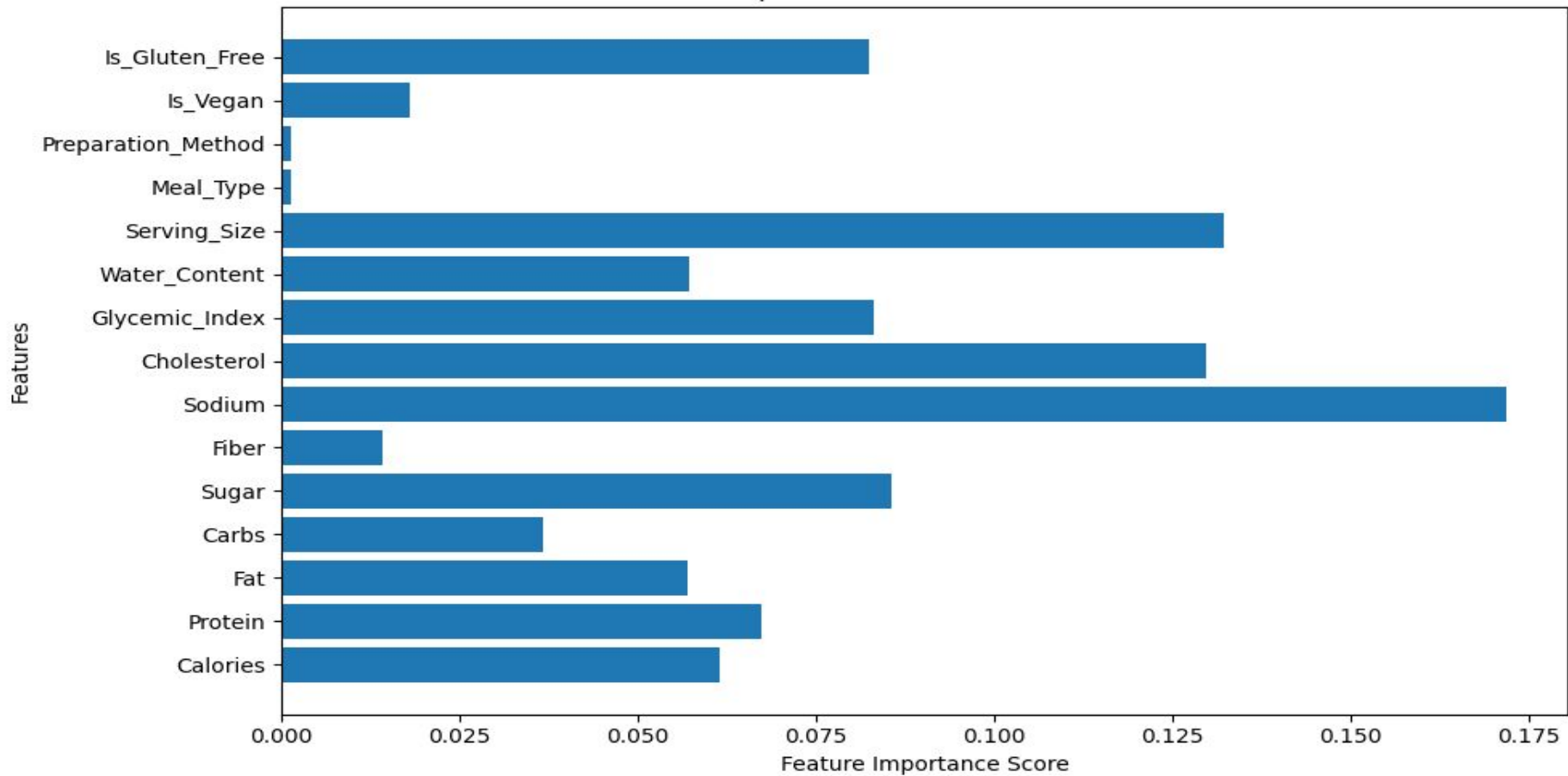
ACCURACY : 94.83%

PRECISION :94.84%

RECALL : 94.83%

F1-SCORE: 94.26%

Feature Importance - Random Forest Classifier

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

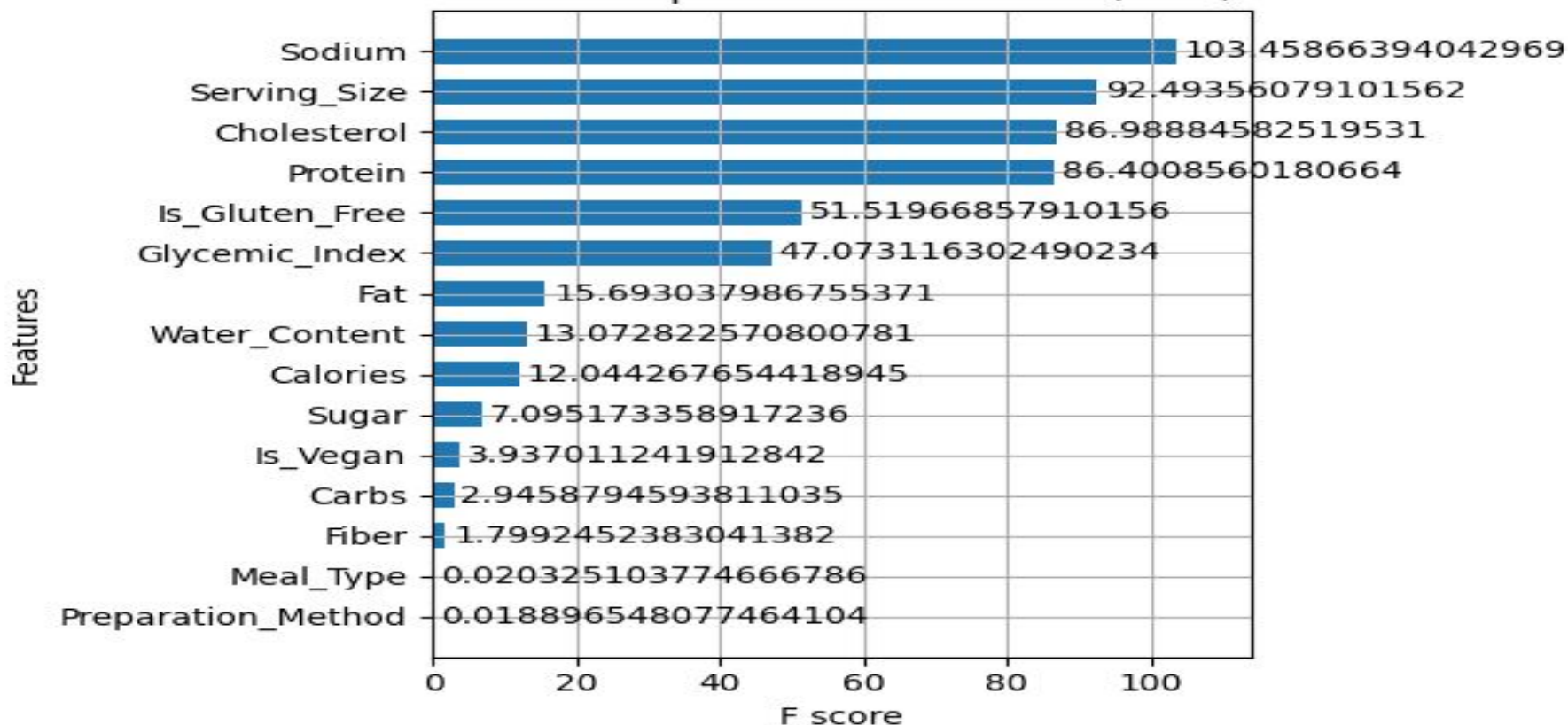prediction

XGB Classifier

ACCURACY: 94.26%

PRECISION:94.27%

RECALL:94.26%

F1-SCORE: 94.26%

Feature Importances - XGBoost (Gain)

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]
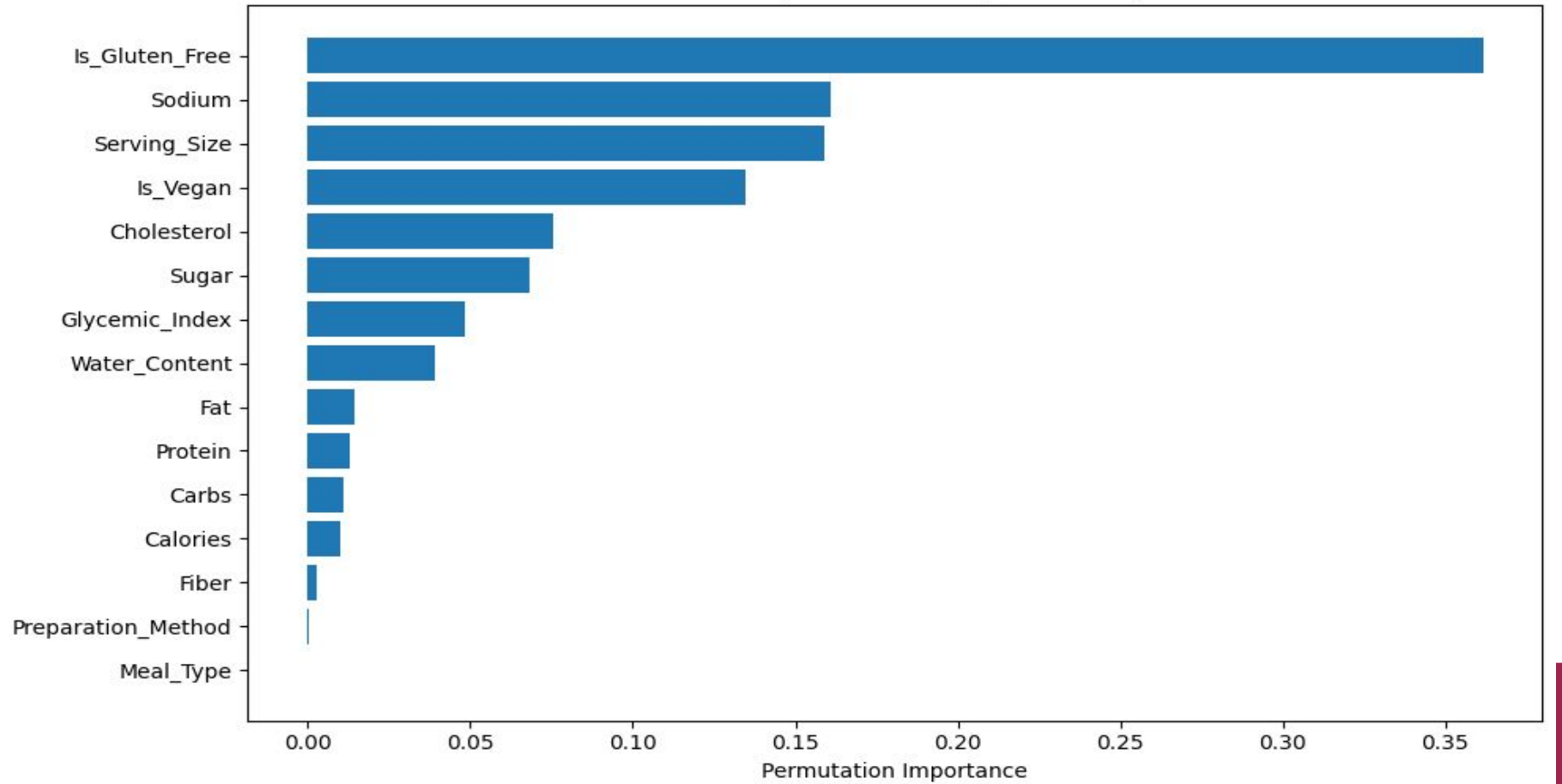
ACCURACY : 94.91%

PRECISION : 94.93%

RECALL : 94.91%

F1-SCORE : 94.91%

Feature Importance - SVC (RBF Kernel)

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
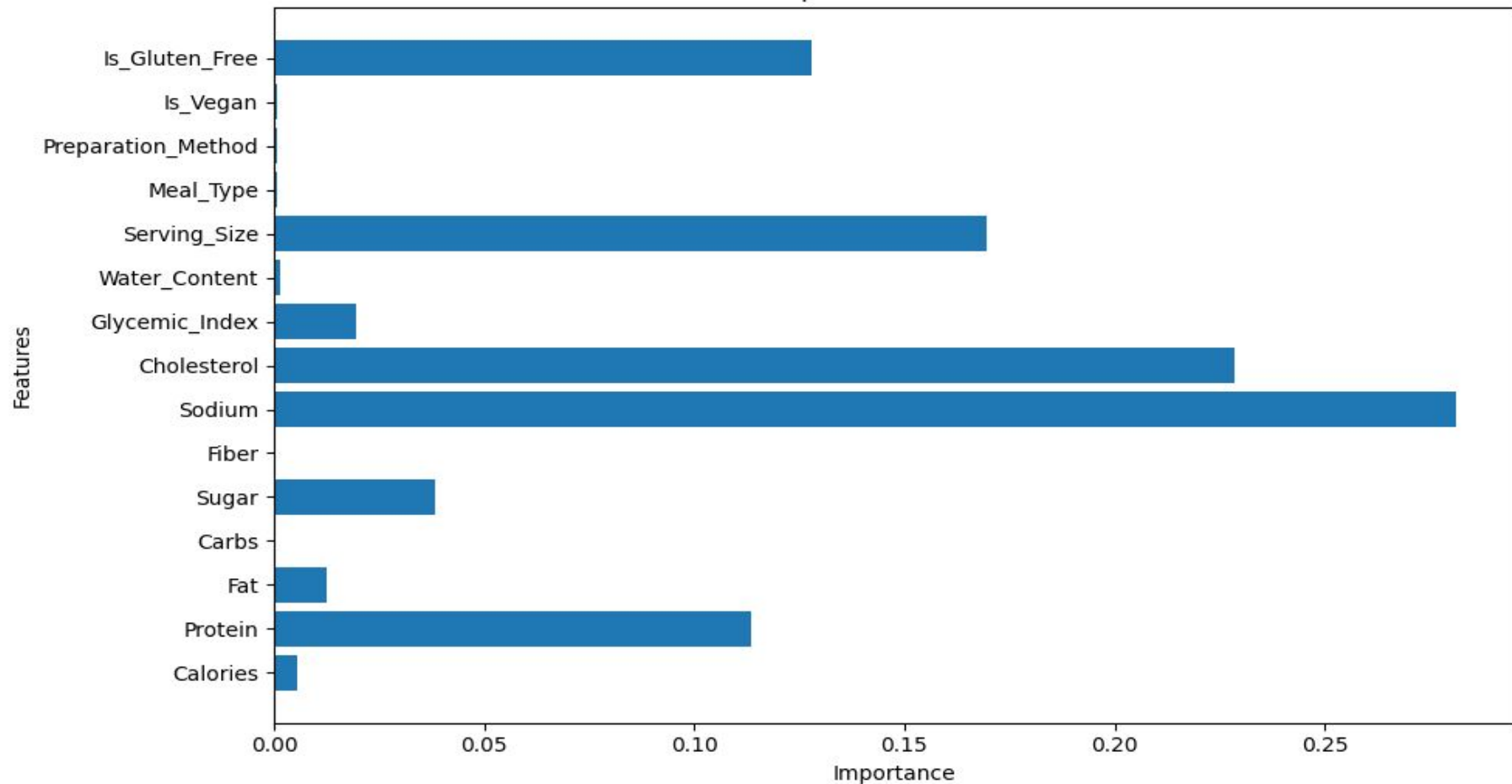
Decision Tree

ACCURACY : 91.39%

PRECISION :89.30%

RECALL : 91.39%

F1- SCORE :90.03%

Feature Importances - Decision Tree

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.

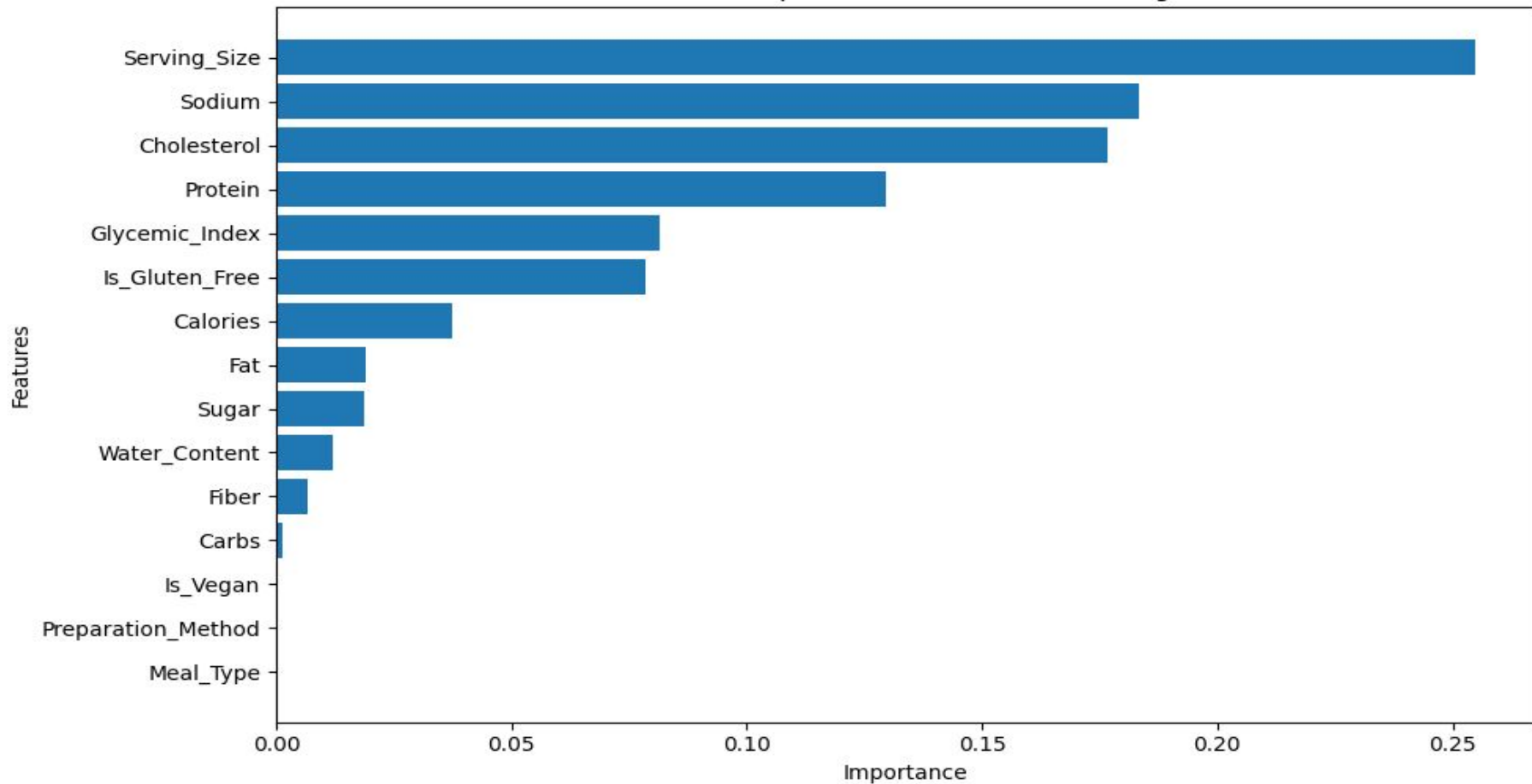# GradientBoostingClassifier

ACCURACY : 94.75%

PRECISION : 94.77%

RECALL: 94.75%

F1-SCORE : 94.75%

Feature Importances - Gradient Boosting

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**:
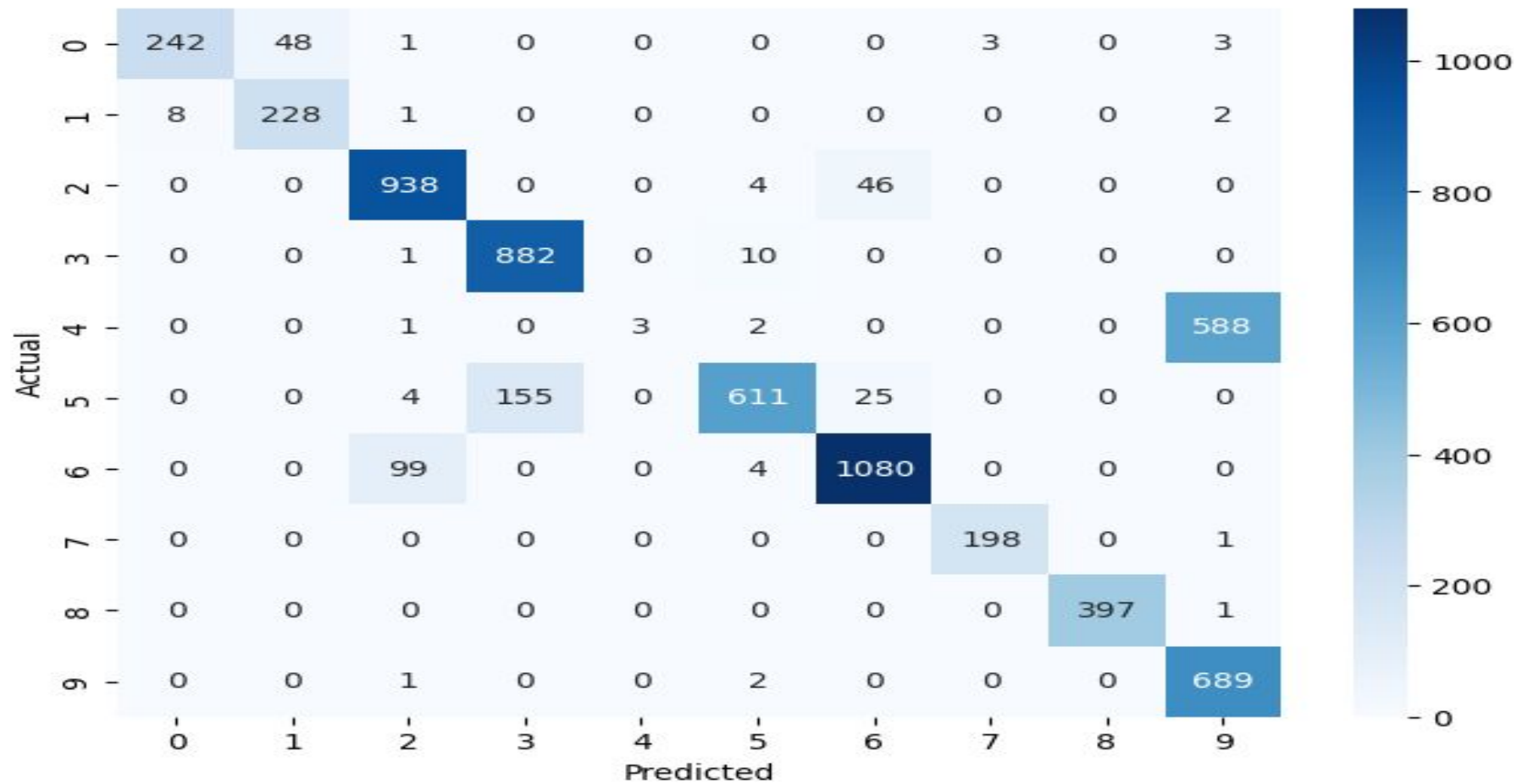
LOGISTIC REGRESSION

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

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
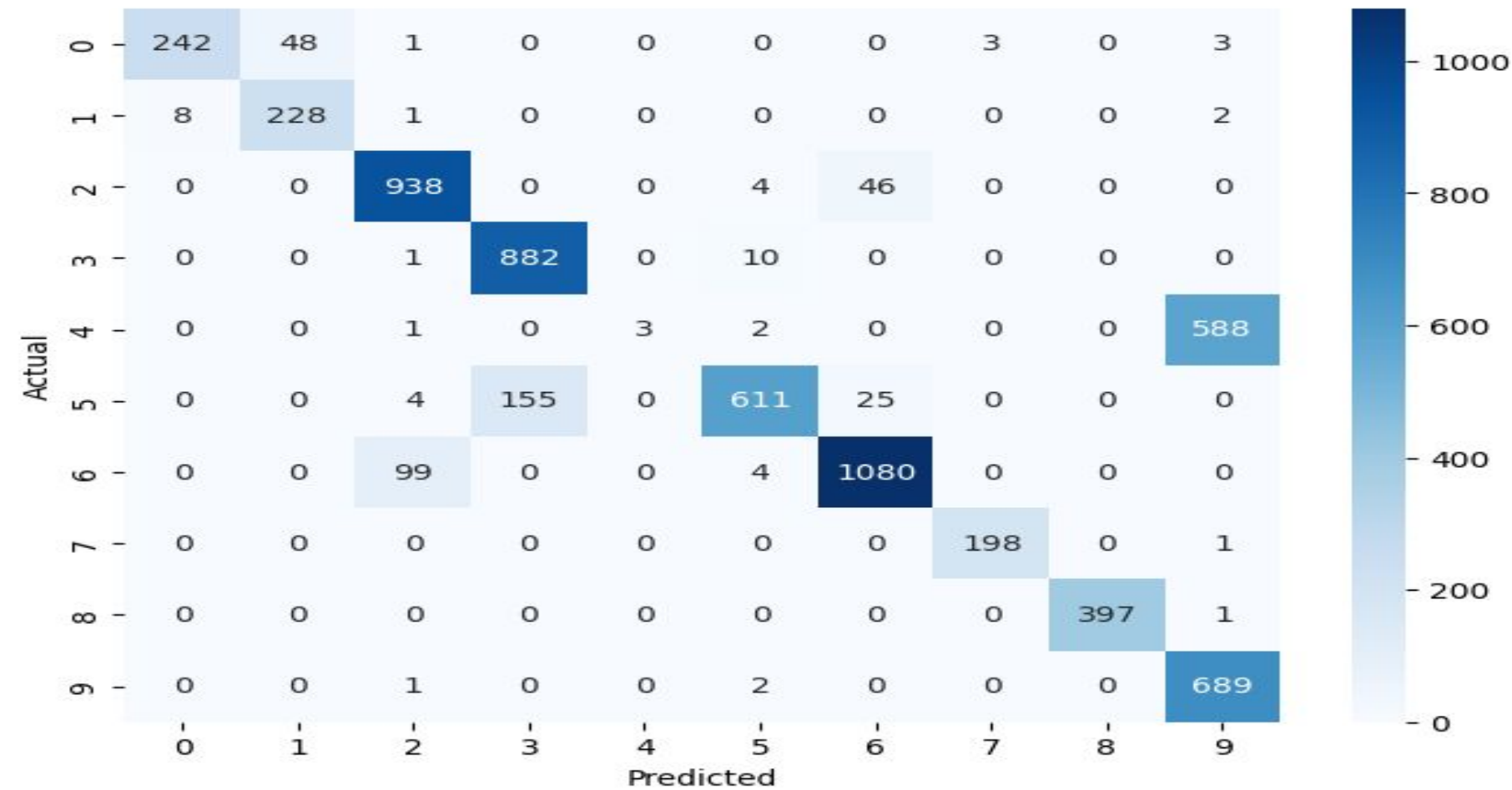
Summary for confusion matrix [SVM]

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.