Recipe Site Traffic Prediction

MAR 2025

Introduction

In the competitive digital food and recipe market, user engagement is critical for driving traffic and subscriptions. Understanding which recipes are likely to attract more visitors can help platforms optimize their content strategy. This analysis focuses on predicting whether a recipe will generate high traffic using various nutritional and categorical attributes. By leveraging machine learning, the goal is to identify the patterns and features associated with popular recipes to guide decision-making for homepage content, marketing focus, and user personalization strategies.

The dataset includes information such as recipe category, calories, carbohydrates, sugar, protein, servings, and a traffic indicator. However, the raw data contains inconsistencies such as missing values, incorrect data types, and ambiguous labels that need to be addressed before any analysis. By systematically cleaning, exploring, modeling, and evaluating the dataset, we aim to build a reliable predictive model and generate actionable business insights.

Methodology

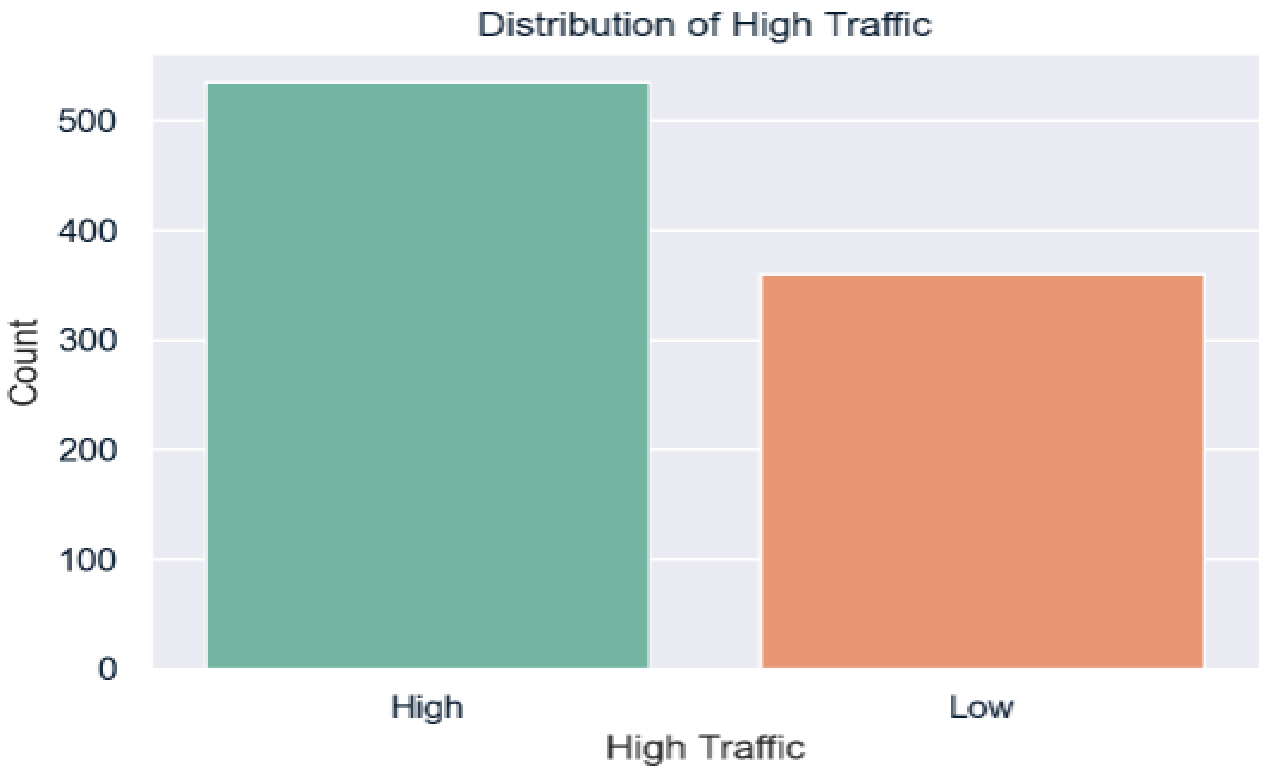
1. Data Validation & Cleaning
2. Exploratory Data Analysis (EDA)
3. Model Development
4. Model Evaluation
5. Business Metrics
6. Result & Recommendation

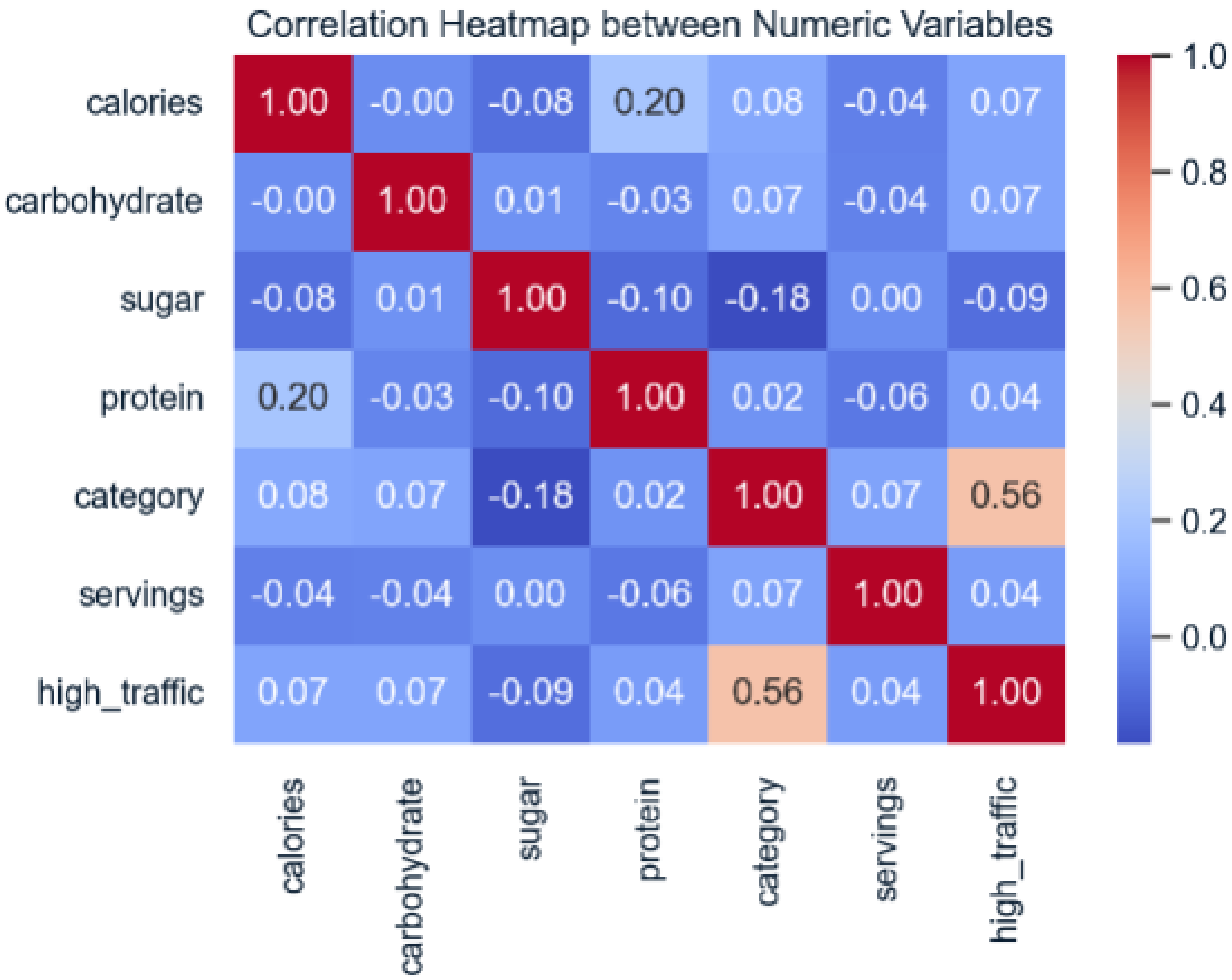
1. **Data Validation & Cleaning**

- Verified data types, value ranges, and missing values across all features.
- Identified and corrected inconsistencies:
  - ✓ Missing values in nutritional attributes (calories, carbohydrates, sugar, protein) were handled using appropriate imputation methods or removal where necessary.
  - ✓ Fixed categorical inconsistencies (e.g., merged "Chicken Breast" into "Chicken") to reduce noise and ensure accurate grouping.

- ✓ Converted non-numeric types (e.g., servings stored as strings) into appropriate numerical formats.
- ✓ Filled missing values in the high_traffic column under the assumption that missing implies "Low".Removed duplicate records and outliers to maintain dataset integrity.

2. Exploratory Data Analysis

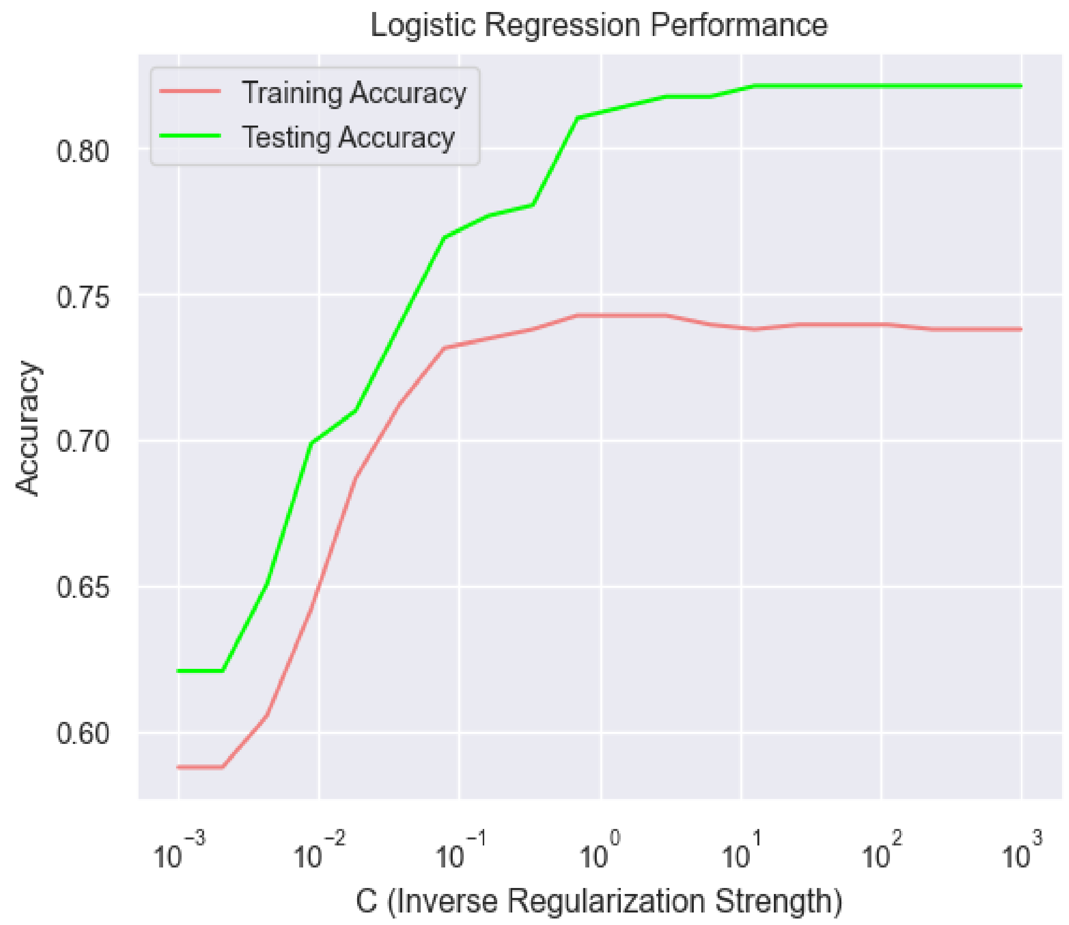


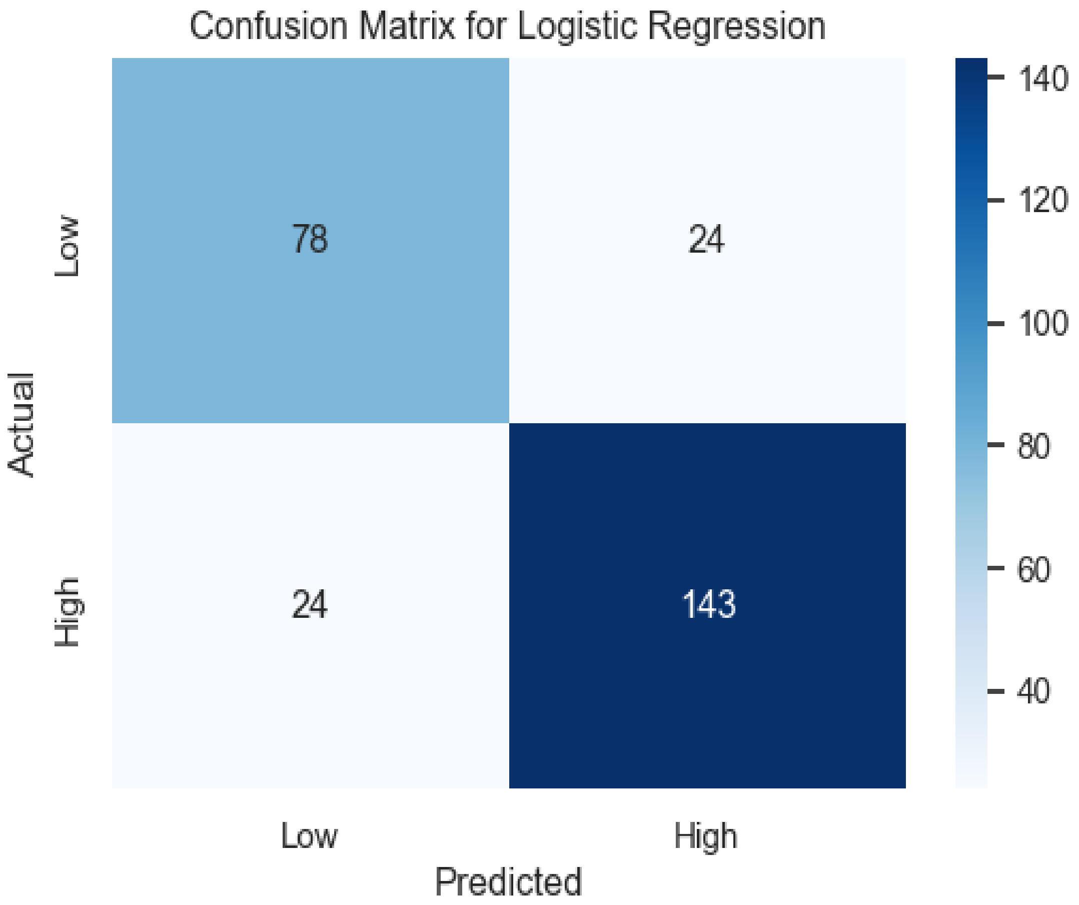


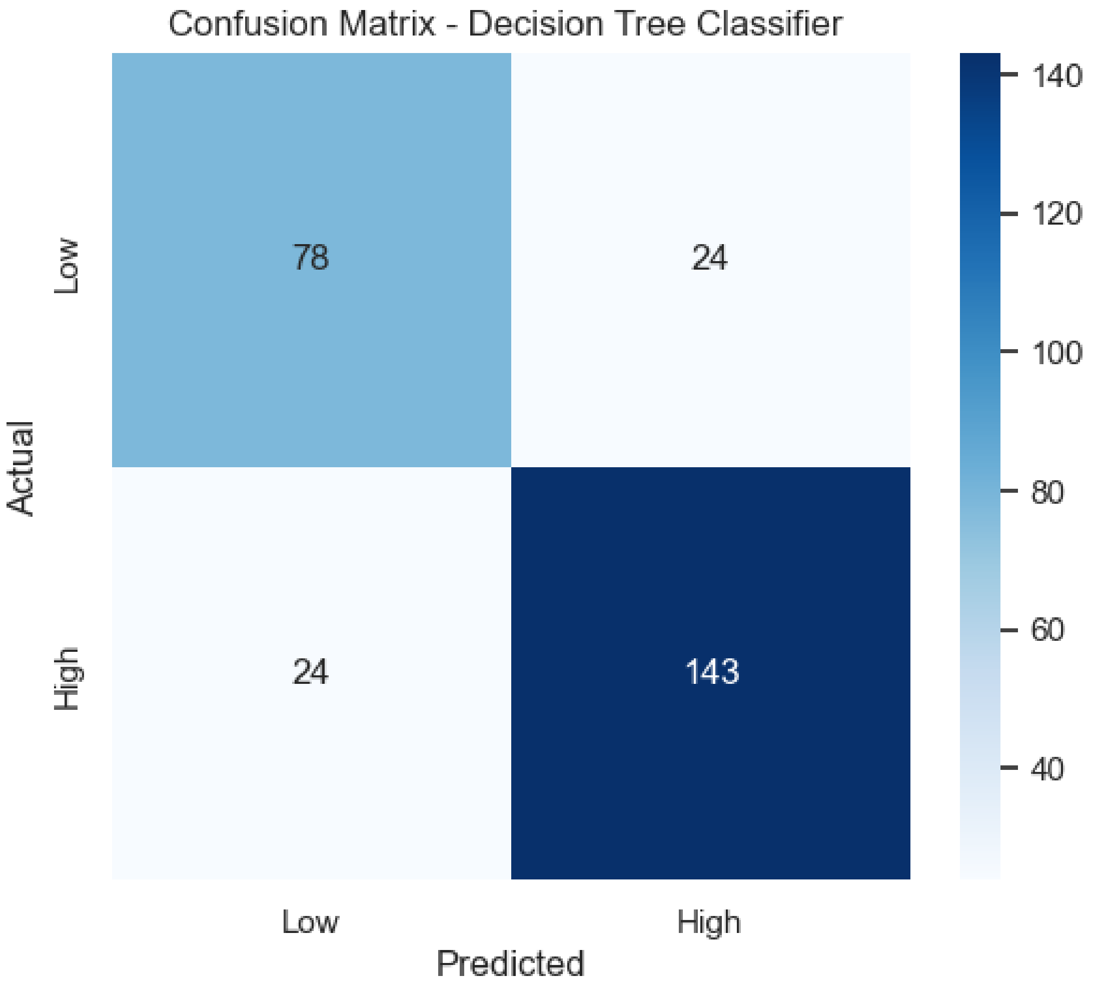
Correlation Heatmap between Numeric Variables

	calories	carbohydrate	sugar	protein	category	servings	high_traffic
calories	1.00	-0.00	-0.08	0.20	0.08	-0.04	0.07
carbohydrate	-0.00	1.00	0.01	-0.03	0.07	-0.04	0.07
sugar	-0.08	0.01	1.00	-0.10	-0.18	0.00	-0.09
protein	0.20	-0.03	-0.10	1.00	0.02	-0.06	0.04
category	0.08	0.07	-0.18	0.02	1.00	0.07	0.56
servings	-0.04	-0.04	0.00	-0.06	0.07	1.00	0.04
high_traffic	0.07	0.07	-0.09	0.04	0.56	0.04	1.00

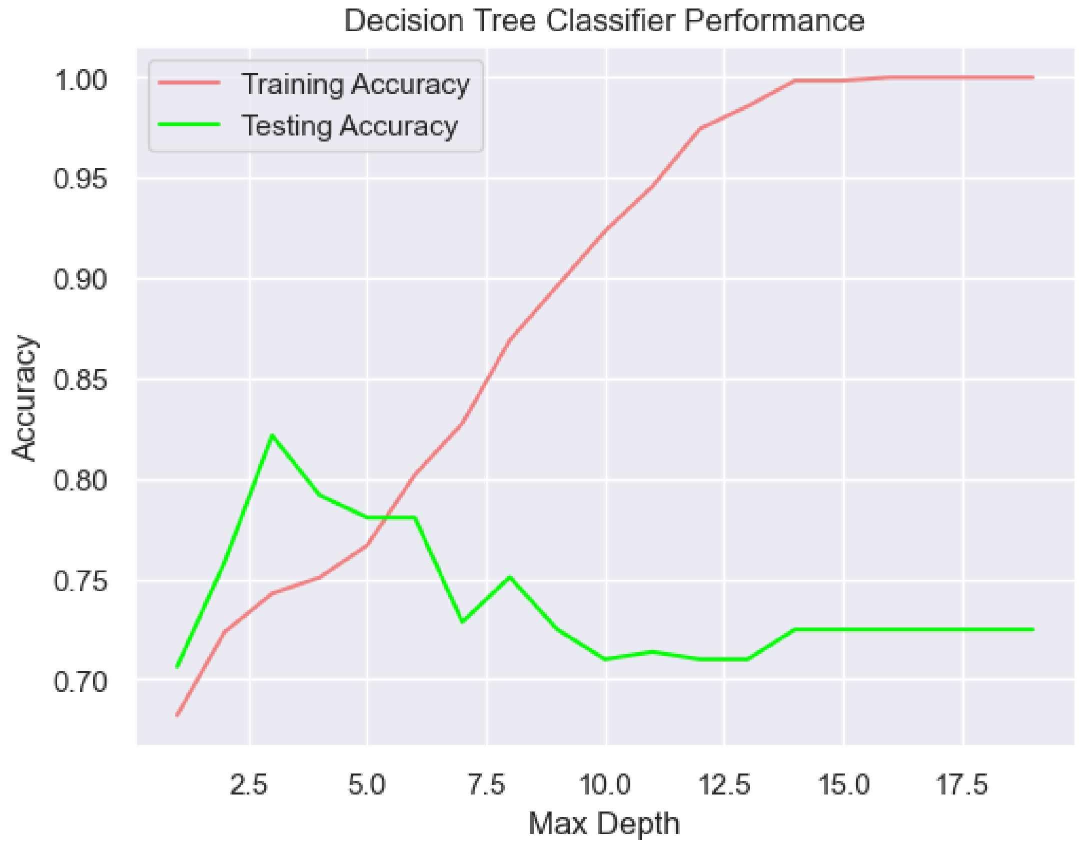
3.Model Development



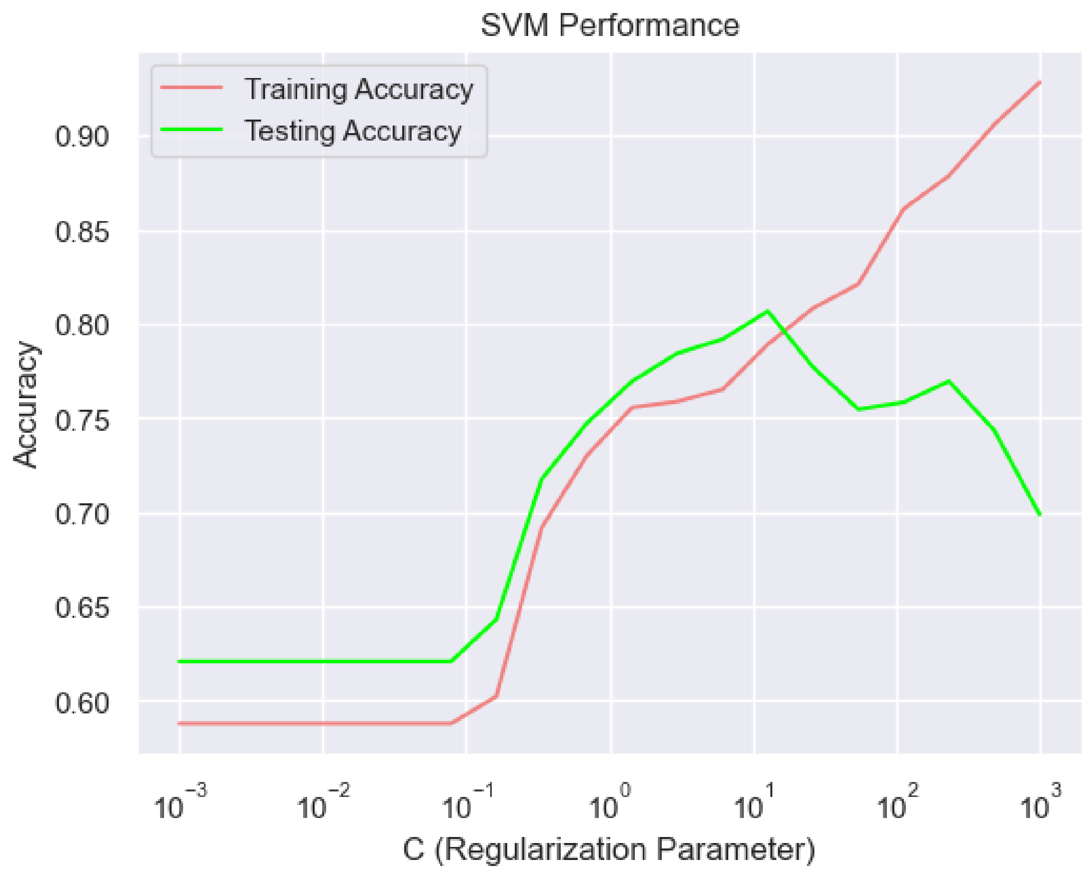




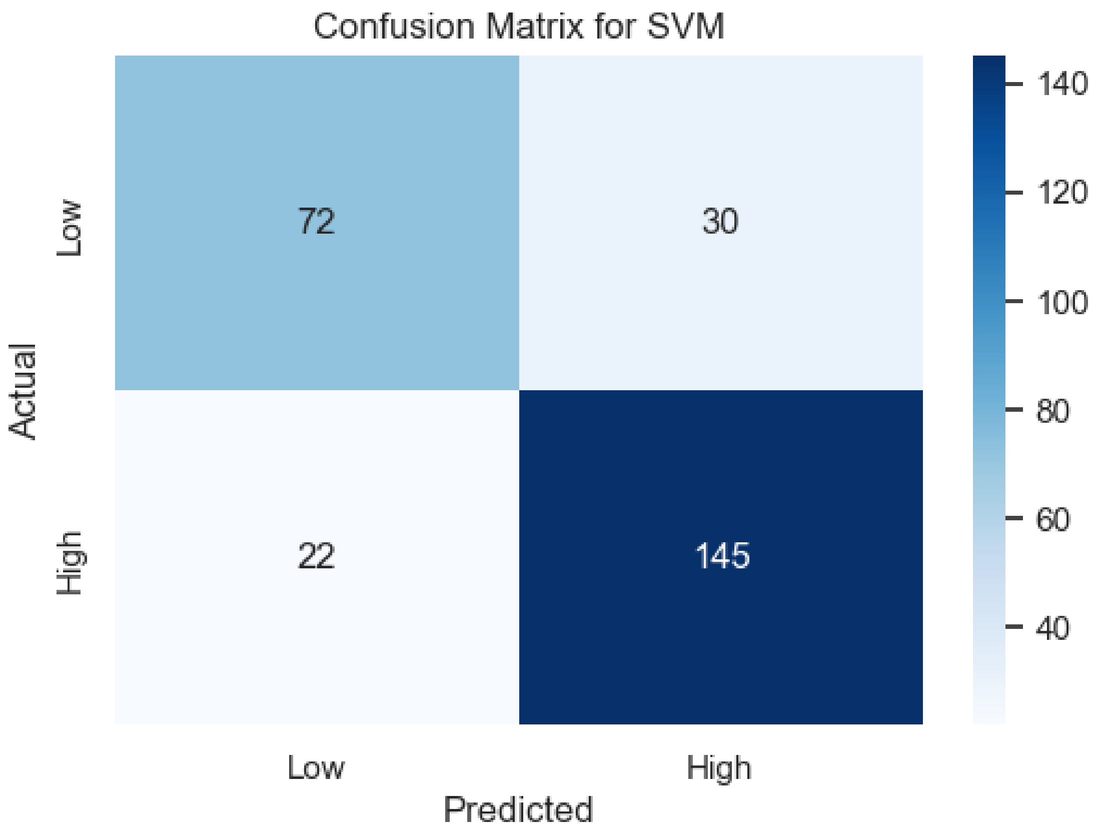
Confusion Matrix - Decision Tree Classifier



Decision Tree Classifier Performance



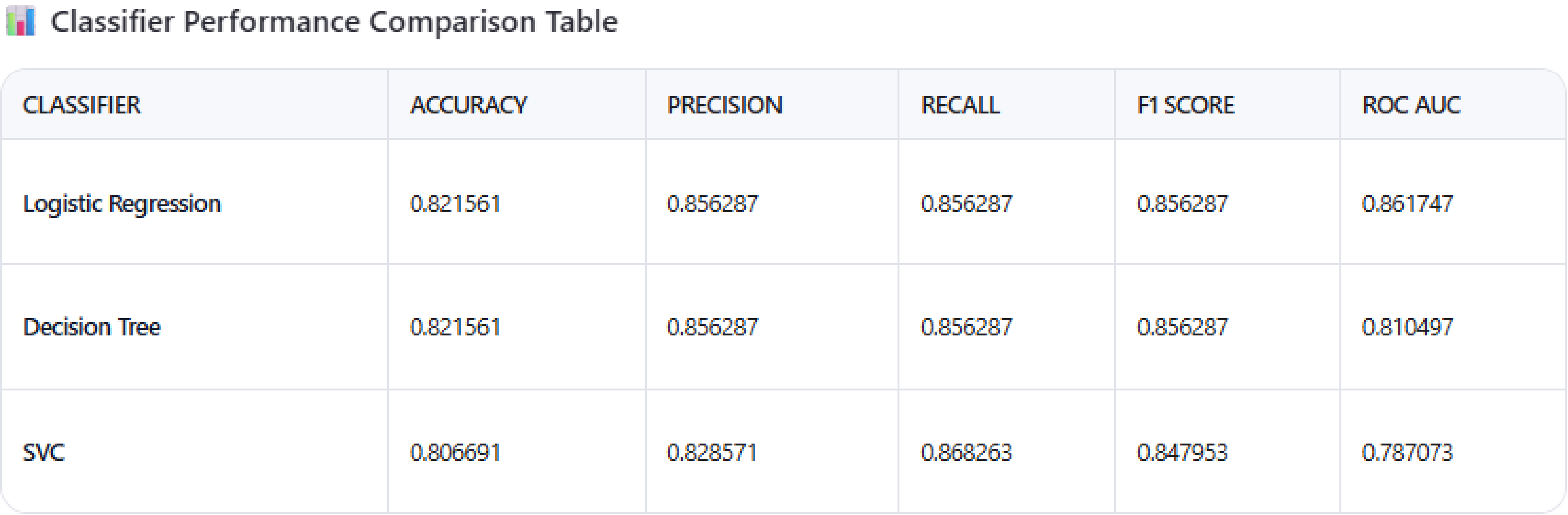
SVM Performance



Confusion Matrix for SVM

Actual \ Predicted	Low	High
Low	72	30
High	22	145

4. Model Evaluation



📊 Classifier Performance Comparison Table

CLASSIFIER	ACCURACY	PRECISION	RECALL	F1 SCORE	ROC AUC
Logistic Regression	0.821561	0.856287	0.856287	0.856287	0.861747
Decision Tree	0.821561	0.856287	0.856287	0.856287	0.810497
SVC	0.806691	0.828571	0.868263	0.847953	0.787073

5. Business Metric

Metric Definition: to align with the business goal of predicting popular recipes 80% of the time while minimizing unpopular recommendations, we propose the following metric:
Weighted Accuracy : A custom metric that weights recall higher than precision to prioritize identifying popular recipes (high traffic = "High").

- The formula Weighted Accuracy =0.7×Recall+0.3×Precision
- Model Performance Using Business Metric Using the weighted accuracy metric:
  - Logistic Regression: 0.7×0.856+0.3×0.856=0.856
  - Decision Tree: 0.7×0.856+0.3×0.856=0.856
  - Support Vector Classifier: 0.7×0.868+0.3×0.829=0.857

- While the tuned Support Vector Classifier slightly outperforms logistic regression in this metric, logistic regression remains the most balanced and interpretable choice.

6. Result & Recommendations

- Visualizations revealed significant differences in recipe categories between high-traffic and other recipes.
- High-traffic recipes favored categories like Potato , Beverages , and Pork , while lower-traffic recipes leaned toward Breakfast , Chicken , and Dessert .
- Nutritional attributes (calories, carbohydrates, sugar, protein) showed weak correlations with traffic levels, indicating that these factors alone are not strong predictors of popularity.

- **Recommendations**
  - Deploy Logistic Regression Model : Given its superior performance, interpretability, and alignment with business goals, the tuned Logistic Regression model should be deployed to predict high-traffic recipes. This model will help the product team identify and display recipes that are likely to drive higher traffic and subscriptions.
  - Monitor Weighted Accuracy : Use the Weighted Accuracy metric to evaluate model performance in production. Set a minimum threshold of 0.8 to ensure the model meets the business requirement of correctly predicting popular recipes 80% of the time.
  - Leverage Recipe Categories : Focus on promoting recipes from categories identified as high-traffic drivers: Potato , Beverages , and Pork . Avoid over-representing less popular categories like Chicken , Dessert , and Vegetables unless they align with specific customer preferences or seasonal trends.

- Feature Engineering : Explore additional features such as seasonality , user demographics , and recipe metadata (e.g., keywords, tags) to further improve model performance. Incorporate feedback loops to capture user interactions (e.g., clicks, shares) and refine predictions over time.
- Iterative Improvement : Continuously monitor model performance and retrain periodically with new data to adapt to changing user preferences. Conduct A/B testing to validate the impact of recommended recipes on website traffic and subscription rates.
- Enhance Data Collection : Address gaps in the current dataset, particularly missing nutritional values and ambiguous categorization. Standardize data collection processes to ensure consistency and completeness for future analyses.

THE END