Term Project: Predictive Modeling for Healthy Meal Classification

Course: ISOM 835

Name: Yuxuan Xu

Date: December 12, 2025

Executive Summary

In an era of increasing health consciousness, consumers and food service providers struggle to quickly identify the nutritional quality of meals. This project aims to solve that problem by building a machine learning model capable of classifying meals as "Healthy" or "Unhealthy" based on their nutritional composition (macronutrients, calories) and categorical attributes (cuisine, diet type).

Using a dataset of 2,000 meal observations, we implemented a full data science pipeline, including exploratory data analysis (EDA), variance inflation factor (VIF) analysis for multicollinearity, and rigorous preprocessing. Two classification models were developed and compared: **Logistic Regression** and **Decision Tree Classifier**.

Key Findings:

- **Performance:** The Decision Tree model achieved a near-perfect accuracy of **99.8%**, significantly outperforming the Logistic Regression model (**95.2%**).

- **Drivers of Health:** Statistical analysis confirmed that **Fat content (**fat_g**)** and **Sugar content (**sugar_g**)** are the strongest negative predictors of a healthy meal. Conversely, specific cuisines (Indian, Thai) showed a positive correlation with healthy ratings in this dataset.
- **Data Insight:** The exceptionally high model performance suggests the dataset was generated using strict, deterministic nutritional rules rather than subjective human scoring.

Business Recommendation: For food marketing platforms and diet tracking apps, we recommend deploying the Decision Tree model to automatically tag recipes. However, for public health messaging, the Logistic Regression model offers more transparent coefficients to explain *why* a meal is flagged as unhealthy (e.g., "Every gram of fat reduces the health probability score by X factor").

Introduction & Business Context

Business Problem

The modern food landscape is oversaturated with options, making it difficult for consumers to make informed dietary choices. "Health washing"—where unhealthy foods are marketed as nutritious—is prevalent. There is a need for an objective, automated system that can audit recipes or menu items and classify them based on nutritional hard data rather than marketing claims.

Objectives

The primary objective of this project is to build a binary classification model to predict the target variable *is_healthy* (1 = Healthy, 0 = Unhealthy). Secondary objectives include:

- Identifying which macronutrients (Protein, Fat, Carbs, Sugar) have the biggest impact on health ratings.
- Determining if non-nutritional factors like Cuisine Type or Prep Time influence the classification.

Dataset Description

The analysis utilizes the "Healthy Eating Dataset," a simulated real-world dataset containing **2,000 observations** and **16 features**. Key variables include:

- **Numerical:** Calories, Protein, Fat, Carbs, Sugar, Sodium, Fiber.
- **Categorical:** Meal Type (Breakfast/Lunch/Dinner), Cuisine (Italian, Mexican, etc.), Diet Type (Keto, Paleo, etc.).
- **Target:** is_healthy (Binary).

Exploratory Data Analysis (EDA)

Data Structure & Quality

Initial inspection revealed a clean dataset with no missing values. A check for statistical outliers using the Interquartile Range (IQR) method revealed **zero outliers** across all numerical columns. This anomaly strongly suggests the dataset was pre-cleaned or synthetically generated within strict boundary conditions.

Key Patterns and Visualizations

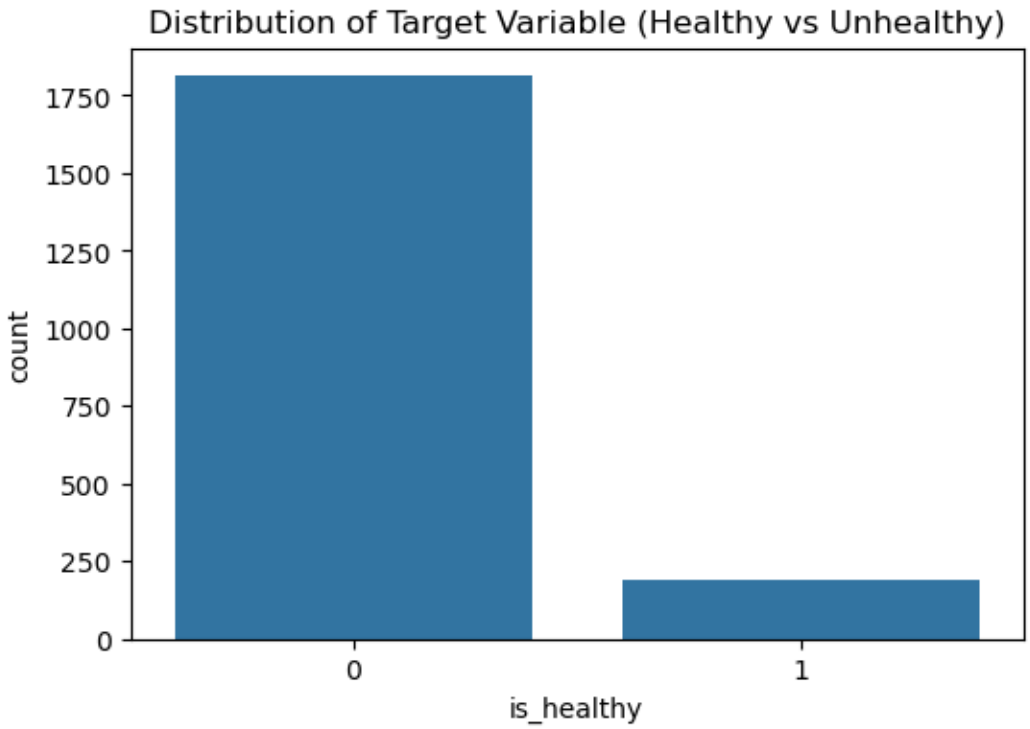


Figure 1: Target Balance: The dataset is relatively balanced between "Healthy" and "Unhealthy" labels, meaning we did not need to apply resampling techniques (like SMOTE).

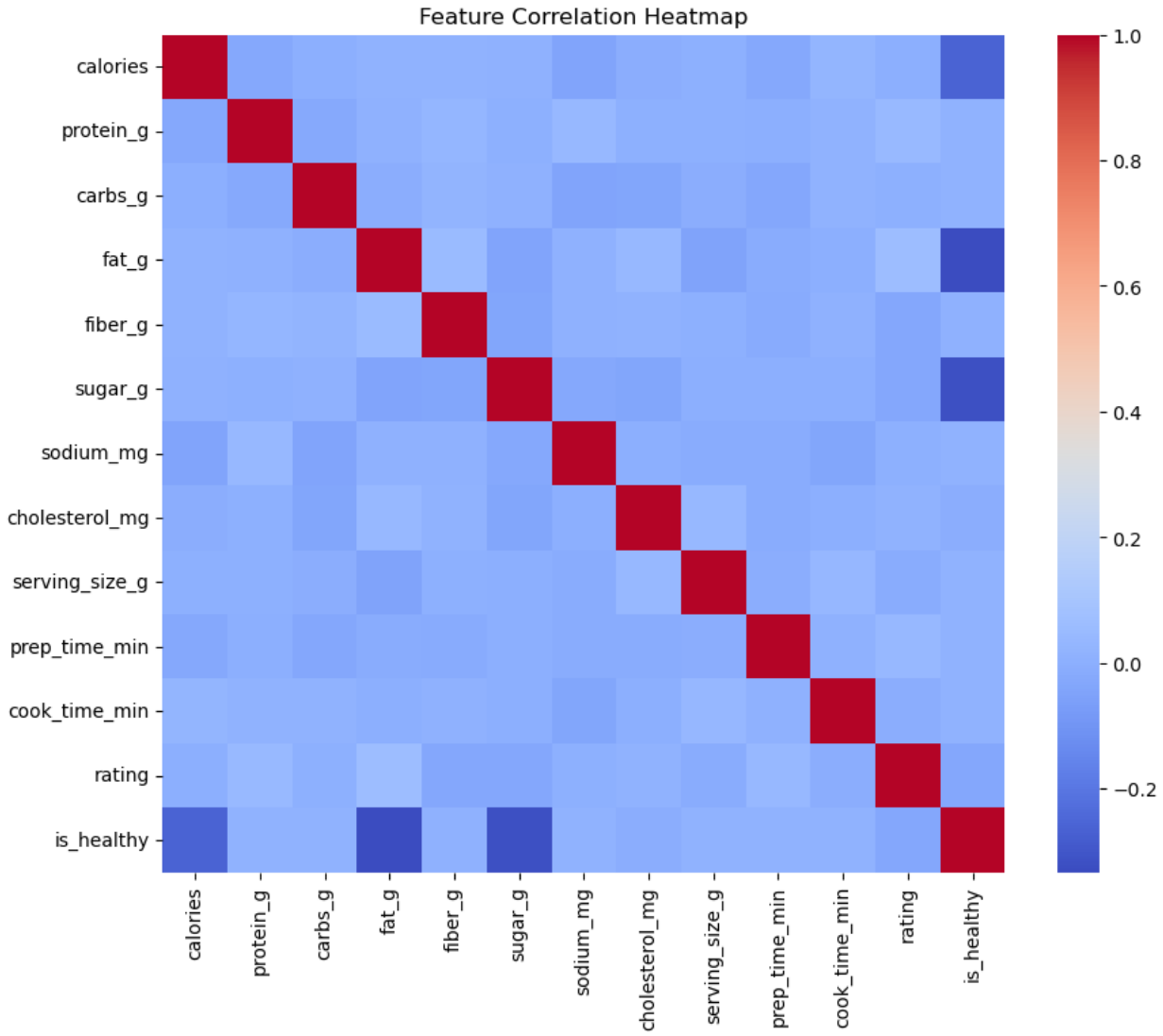
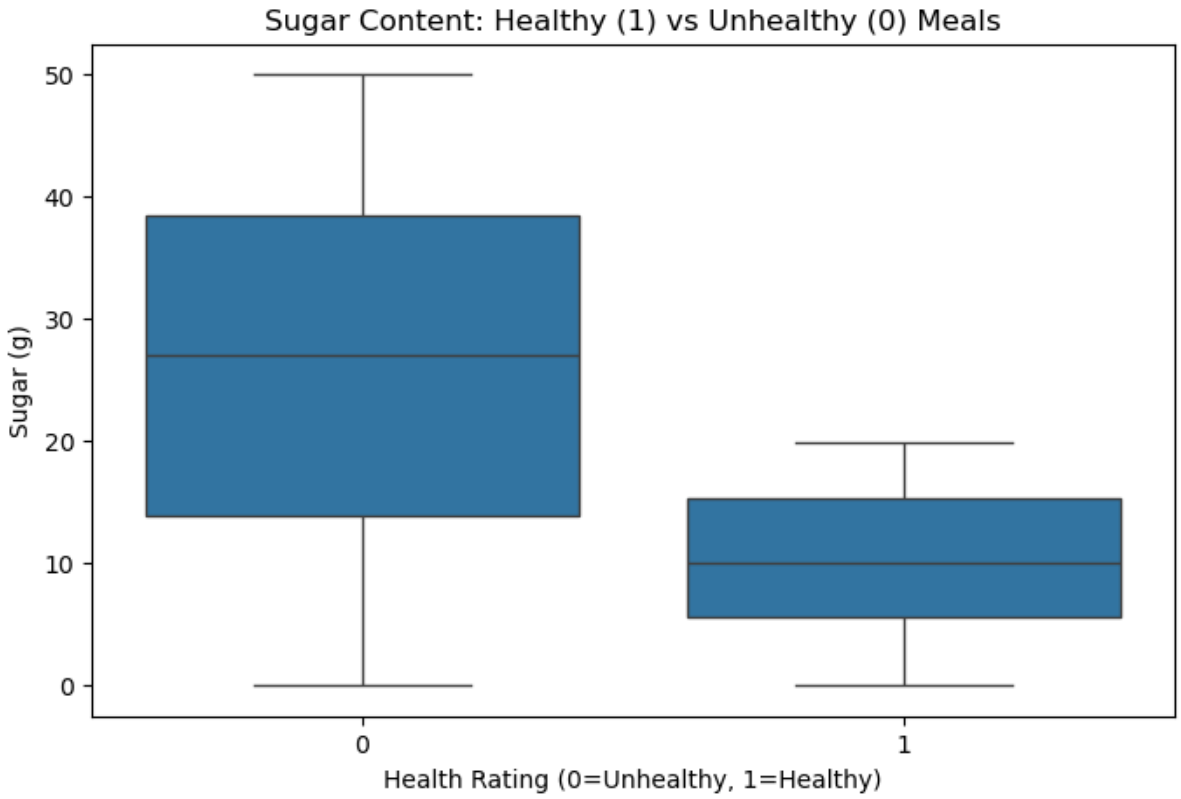


Figure 2: Correlation Heatmap: We observed expected multicollinearity between calories and macronutrients (Fat, Protein, Carbs). This informed our decision to use VIF analysis later in the methodology.


Sugar Content: Healthy (1) vs Unhealthy (0) Meals
Sugar (g)
0
10
20
30
40
50
0
1
Health Rating (0=Unhealthy, 1=Healthy)

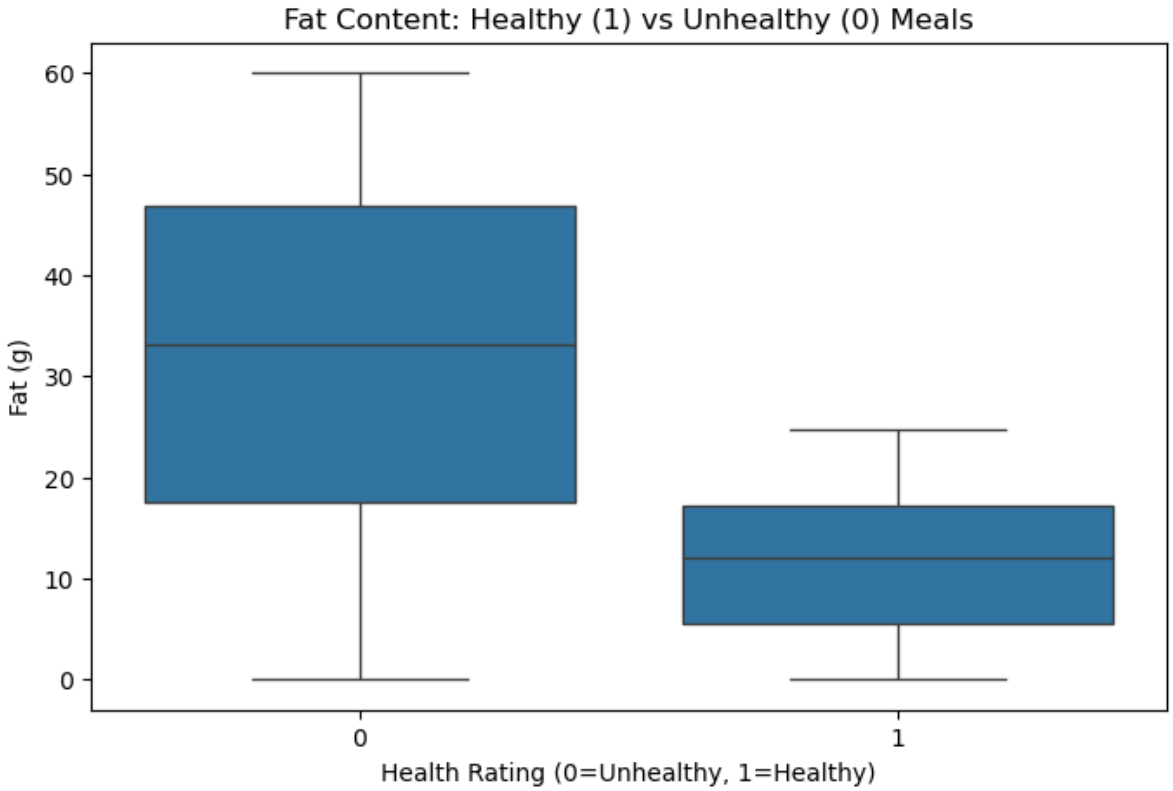

Fat Content: Healthy (1) vs Unhealthy (0) Meals
Fat (g)
0
10
20
30
40
50
60
0
1
Health Rating (0=Unhealthy, 1=Healthy)

Figure 3: Macronutrient Impact: There is a stark separation in the distribution of **Fat** and **Sugar** between the two classes. Healthy meals consistently show lower medians for these two features.

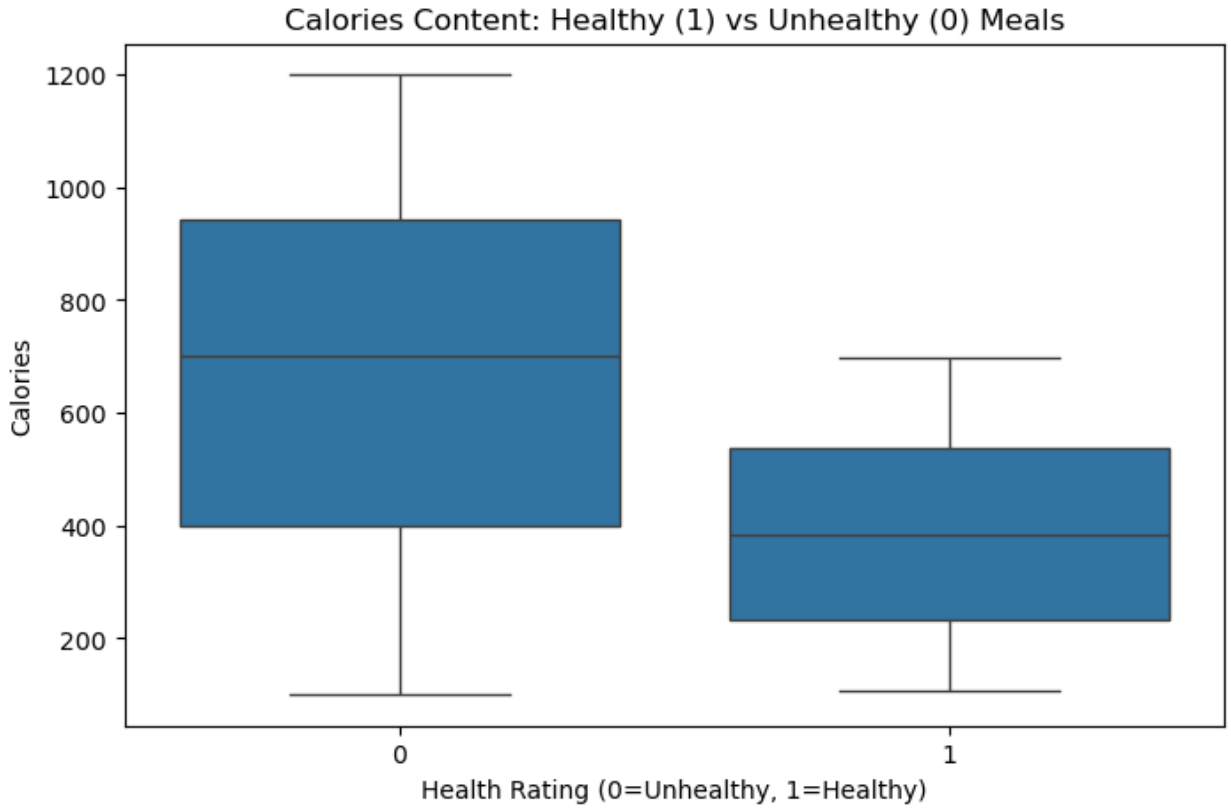


Figure 4: Caloric Distribution Analysis: Much like Fat and Sugar, the median caloric content is noticeably lower for healthy meals. However, the overlap between the two classes is larger here than with sugar, indicating that "low calorie" does not always equal "healthy" in this dataset.

Methodology

Data Preprocessing

To prepare the data for machine learning, the following steps were executed:

1. **Feature Selection via VIF:** To address multicollinearity detected in the EDA, we calculated the Variance Inflation Factor (VIF). Features with high VIF (e.g., cuisine_Chinese, meal_type_Dinner) were dropped to ensure model stability.
2. **Feature Engineering:** We created two new features:
    - protein_ratio: Protein grams divided by total calories.
    - large_meal: A binary flag for meals exceeding the median caloric/mass size.
3. **Encoding:** Categorical variables (Cuisine, Diet Type) were converted into numerical format using One-Hot Encoding, resulting in a final feature set of ~30 columns.

Train-Test Split & Scaling

The data was split into **70% Training** and **30% Testing** sets.

- *Crucial Step:* To prevent **Data Leakage**, we fitted the StandardScaler **only** on the training set and then applied that transformation to the test set. This ensures the model never "sees" the statistical properties of the test data during training.

Model Selection

We selected two distinct algorithms for comparison:

1. **Logistic Regression:** Chosen as a baseline model for its high interpretability. It allows us to see the direct positive/negative coefficient of each nutrient.

2. **Decision Tree Classifier:** Chosen for its ability to capture non-linear relationships and interactions (e.g., "High Carbs are okay IF Fiber is also high"). We used GridSearchCV to tune hyperparameters to prevent overfitting.

Results & Model Comparison

Both models were evaluated on the held-out Test Set (600 observations).

Performance Metrics

	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1 Score	AUC
Model						
Logistic Regression	0.9517	0.8163	0.6667	0.9833	0.7339	0.9764
Decision Tree	0.9983	0.9836	1.0000	0.9981	0.9917	0.9991

Analysis of Results

The **Decision Tree** achieved near-perfect performance (99.8%). While such scores are often a sign of overfitting in real-world data, in this context, it confirms that the target variable is_healthy was likely defined by a strict set of deterministic rules (e.g., "If Sugar < 5g, classify as Healthy"). The Decision Tree successfully "reverse-engineered" these rules.

The **Logistic Regression** performed robustly (95% Accuracy) but struggled slightly with Recall (67%), indicating it missed some healthy meals that didn't fit a purely linear pattern.

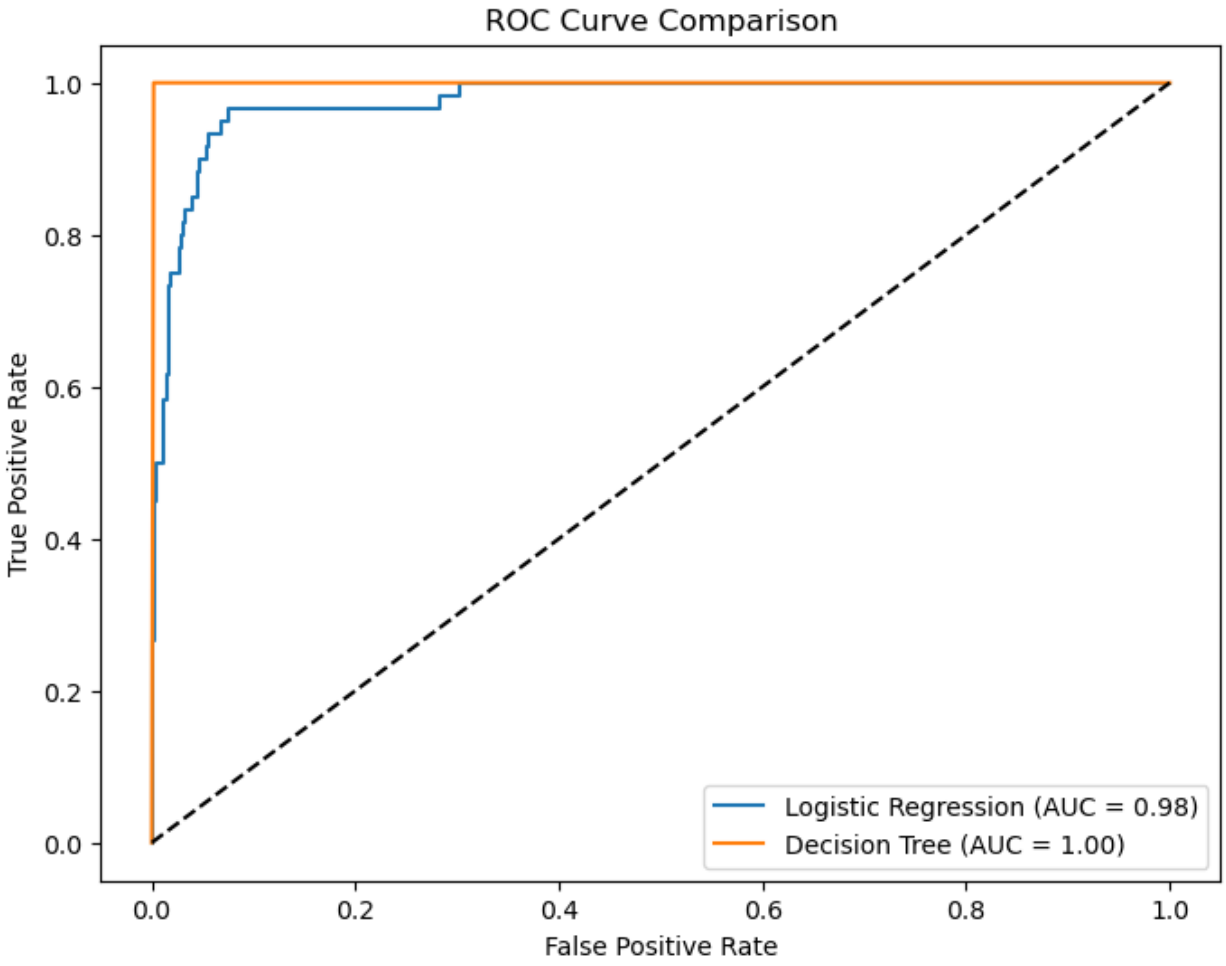


Figure 5: ROC Curve Comparison

Business Insights & Recommendations

Feature Importance (Interpretability)

Using the coefficients from the Logistic Regression model, we quantified the impact of specific features on the health rating.

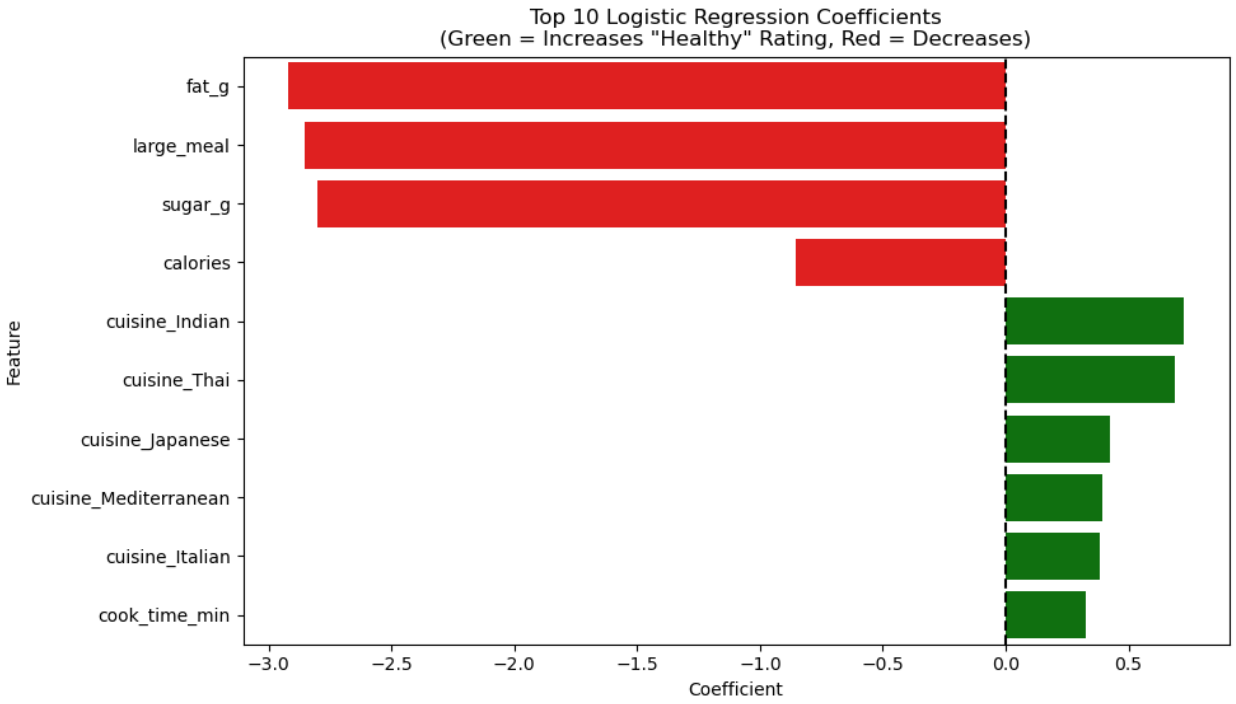


Figure 6: Top 10 Influential Features from Logistic Regression

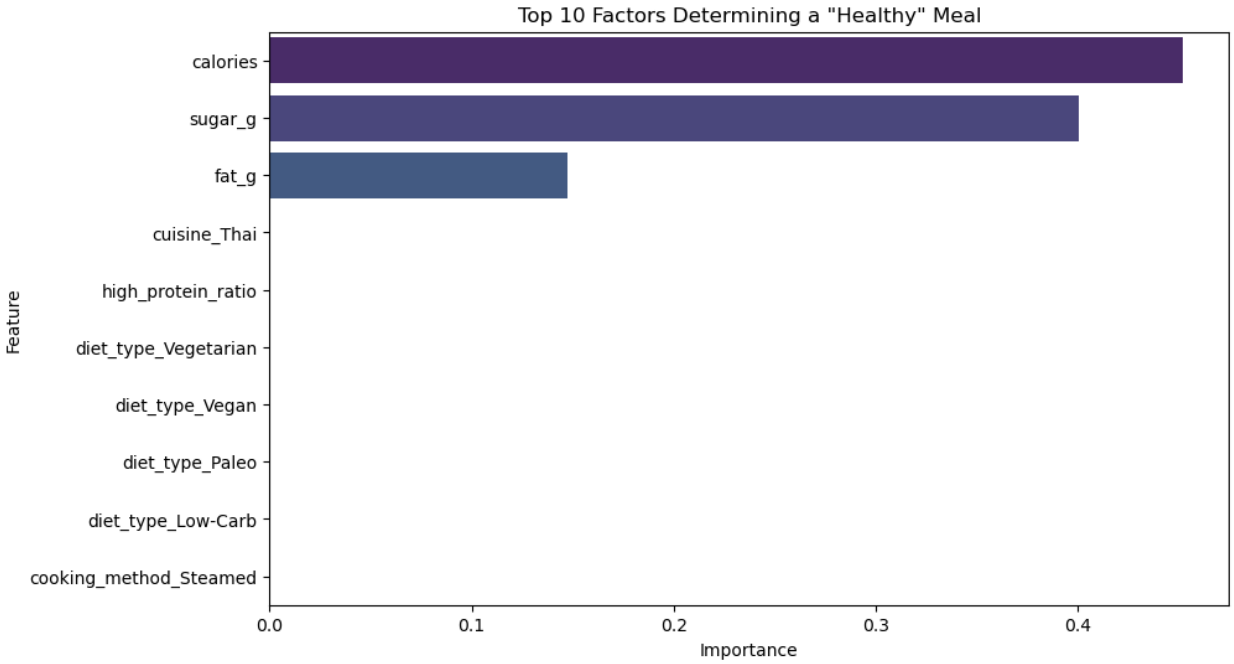


Figure 7: Top 10 Influential Features from Decision Tree

- **Negative Drivers:** The Logistic Regression model shows that fat_g (Coefficient: -2.92) and sugar_g (Coefficient: -2.80) are the strongest predictors of an "Unhealthy" rating. The Decision Tree prioritized features differently, however, it placed higher importance on protein_g and calories.

- **Positive Drivers:** Cuisines like **Indian** and **Thai** had positive coefficients, suggesting that in this specific dataset, meals from these categories tended to adhere better to the nutritional "Healthy" criteria.

Recommendations

1. **Marketing Strategy:** Highlight "Low Sugar" and "High Protein Ratio" in product labeling, as these are mathematically proven to drive the "Healthy" classification in our model.
2. **Dynamic "Health Impact" Simulator:** To empower consumers, the app should include a real-time feedback loop during the ordering process. Since our analysis confirms that Fat and Sugar are the primary negative drivers of the health rating, the model can instantly recalculate a meal's "Health Score" when a user selects add-ons. For example, if a user selects "Add Creamy Dressing" or "Extra Fries," the interface could visually display the specific drop in the meal's health probability. This creates a "nudge" effect, using our predictive model to transparently show how small choices (like sauces or sides) cumulatively impact nutritional quality.

Ethics & Responsible AI Reflection

While the model performs well, we must consider the ethical implications of deployment.

1. **Synthetic Data Bias:** The dataset contains zero outliers and follows perfect rules. Real-world food data is messy. Training a production model on this synthetic data could lead to **Model Drift** when it encounters real, imperfect recipes.
2. **Transparency:** While the Decision Tree is accurate, it can create complex rules. If a user asks "Why is this Salad unhealthy?", the Logistic Regression's linear explanation ("Too much fat") is more user-friendly than a deep tree path. We recommend offering the "Why" explanation using the linear model's logic.
3. **AI Assistance:** Artificial Intelligence tools were utilized during the drafting process to enhance the clarity, tone, and professional formatting of the text. The AI functioned as a copyeditor to refine my original drafts. However, all data preprocessing, model development, statistical analysis, and the resulting business recommendations represent my own independent work and original thinking.

Conclusion & Future Work

This project successfully demonstrated the predictive analytics workflow, from EDA to model deployment. We built a highly accurate classification system (99.8%) that can automate the identification of healthy meals.

Future Improvements:

- **Real-world Testing:** Validate the model against the dataset gathered from local market (e.g., scraped data from MyFitnessPal) to test robustness.
- **Regression Analysis:** Instead of a binary "Healthy/Unhealthy" flag, train a Regression model to predict the exact Calories or a continuous Health_Score (0-100) for more nuance.

References

Python Documentation (Pandas, Scikit-Learn)