FlavourAI: Personalized Dish Recommendation System

A Data-Driven Approach to Recipe Discovery and Nutritional Insights

Student Name: Aditya Adke

Student ID: 84825619

Submission Date: April 23, 2025

University: University of Florida

Course: CAP5771 - Applied Data Science, Spring 2025

Instructor: Dr. Laura Cruz Castro

Table of Contents

1. Introduction

1.1 Background and Purpose

The food industry is undergoing a transformation driven by data-driven insights into nutrition, consumer preferences, and recipe analysis. Understanding the composition of recipes, nutritional values, and user feedback plays a crucial role in food recommendation systems, dietary planning, and health-conscious decision-making.

This project aims to analyze recipe datasets, uncovering patterns in ingredients, cooking times, and nutritional values, rather than focusing solely on isolated recipe metrics. By integrating user reviews and ratings, we explore how consumer preferences align with health-conscious trends, dietary restrictions, and regional cuisine variations.

Key Research Questions:

- How do nutritional trends vary across different recipe categories?
- Which ingredients are most commonly associated with high-calorie or low-calorie dishes?
- How do user ratings and reviews influence recipe popularity?
- What are the key factors affecting cooking time in different types of recipes?
- Can machine learning models predict user preferences based on recipe characteristics?

1.2 Tool to be Developed

Interactive Dashboard & Predictive Analysis

To effectively analyze and compare recipe trends, nutritional values, and user preferences, this project will develop an interactive dashboard that enables users to:

- **Dynamically explore recipes** based on categories, ingredients, and cooking time.
- **Analyze nutritional distributions** (Calories, Fat, Protein, Carbohydrates, Sugar, etc.).
- **Identify popular and highly rated recipes** using user reviews and ratings.
- **Discover ingredient relationships** and how they impact overall nutritional values.
- **Generate predictive insights** using machine learning models to recommend recipes based on user preferences and dietary restrictions.

This data-driven framework will allow chefs, nutritionists, food bloggers, and health-conscious individuals to explore relationships between ingredients, nutrition, and user engagement, making it easier to recommend personalized meal choices and enhance food discovery.

1.3 Technology Stack Used

To implement this project, a combination of data processing, visualization, and machine learning tools will be used to ensure scalability, efficiency, and interactivity in analyzing recipe data.

- **Programming Language: Python** – The core language for data analysis and modeling, widely used for its extensive data science libraries.
- **Data Manipulation: Pandas & NumPy** – Essential for handling large recipe datasets efficiently, performing statistical analysis, and cleaning raw data.
- **Visualization: Matplotlib & Seaborn** – Used to create graphs and visualizations that highlight trends in **nutrition, cooking times, and user preferences**.
- **Machine Learning: Scikit-Learn** – Supports model training for **recipe recommendation, nutritional analysis, and user preference prediction**.
- **Natural Language Processing (NLP): NLTK / spaCy** – Used to analyze **recipe descriptions, ingredient lists, and user reviews** for sentiment analysis and keyword extraction.
- **Dashboard Framework: Streamlit / Plotly Dash** – Enables web-based **interactive visualizations**, allowing users to explore recipes dynamically based on dietary preferences and ingredient availability.
- **Version Control & Collaboration: GitHub** – Ensures smooth **code management, version tracking, and project collaboration** for maintaining an organized development workflow.

By integrating these technologies, the project will establish a seamless pipeline from raw recipe data to insights, making nutrition analysis, recipe exploration, and food discovery more accessible, interpretable, and actionable for a wide range of users.

2. Project Timeline

2.1 Milestones Overview

Developing a robust environmental monitoring system requires a well-structured approach to ensure each phase of the project is completed efficiently and on time. This project follows a phased timeline that aligns with industry-standard data science methodologies such as the Cross-Industry Standard Process for Data Mining (CRISP-DM).

The timeline is divided into three major milestones, covering key stages from data collection and preprocessing to exploratory analysis, model development, and dashboard implementation. Each milestone has specific deliverables, ensuring steady progress and continuous evaluation.

This structured approach allows for flexibility while maintaining clear goals, ensuring that data insights are actionable and the final tool is both functional and user-friendly.

2.2 Timeline Breakdown

The following table outlines the key phases, deadlines, and expected deliverables of the project.

Milestone	Timeline	Key Tasks	Deliverables
Milestone 1: Data Collection, Preprocessing & EDA	Feb 5 – Feb 21, 2025	Collect and clean datasets, perform EDA	EDA Report, Initial GitHub Setup
Milestone 2: Feature Engineering & Data Modeling	Feb 21 – Mar 21, 2025	Feature selection, model training & evaluation	Model performance analysis
Milestone 3: Tool Development & Presentation	Mar 24 – Apr 23, 2025	Dashboard development, final evaluation	Final Report, Tool Demo, GitHub update
Final Submission & Presentation	Apr 23, 2025	Final report submission & project demo	Complete report, GitHub repo, presentation

2.3 Milestone 1: Data Collection, Preprocessing, and Exploratory Data Analysis (EDA)

Timeline: February 5 – February 21, 2025

This phase focuses on collecting, cleaning, and analyzing data to identify patterns, inconsistencies, and key insights. The goal is to prepare the data for further modeling and analysis.

Key tasks include:

- Acquiring datasets from reliable sources and ensuring data accuracy.
- Handling missing values, inconsistencies, and redundant data.
- Conducting exploratory data analysis (EDA) to identify trends and anomalies**.**

Deliverables:

- Milestone 1 Report, including EDA findings, statistical summaries, visualizations, and data preprocessing steps.
- Data cleaning scripts and notebooks stored in the GitHub repository.

2.4 Milestone 2: Feature Engineering and Model Development

Timeline: February 21 – March 21, 2025

This phase focuses on preparing data for modeling by selecting relevant features and developing predictive models.

Key tasks include:

- Feature engineering to create new meaningful variables.
- Feature selection based on correlation analysis and importance ranking.
- Model training and evaluation using machine learning techniques.

Deliverables:

- Milestone 2 Report, including feature selection methodology, model performance metrics, and comparisons.
- Machine learning scripts and trained models stored in the GitHub repository.

2.5 Milestone 3: Tool Development, Testing, and Final Report

Timeline: March 24 – April 23, 2025

This phase integrates all components into a functional, user-friendly tool while conducting final evaluations.

Key tasks include:

- Developing an interactive dashboard for real-time visualization and analysis.
- Integrating machine learning models into the system for predictive insights.
- Conducting final testing, refining usability, and preparing for project submission.

Deliverables:

- Final project report summarizing all findings, methodologies, and conclusions.
- Fully functional interactive dashboard.
- Final presentation and GitHub repository with complete documentation

3. Data Collection

This section outlines the datasets used in the project, how they were acquired, and any challenges encountered during the collection process. Ensuring high-quality data is essential for generating accurate insights into recipe trends, user preferences, and nutritional patterns, as well as building reliable recommendation models.

The project integrates multiple recipe-related datasets, combining recipe metadata, user reviews, and ingredient details to provide a holistic analysis of food trends. Unlike traditional studies that focus solely on individual recipe ratings or nutritional values, this approach retains all datasets without filtering, enabling a comprehensive understanding of how ingredients, ratings, and user engagement evolve over time.

Dataset Name	Source	Number of Rows	Number of Columns	Key Features
Food.com Recipes Dataset	Kaggle	~522,517	28	RecipeID, Name, Keywords, RecipeIngredientParts
Food.com Reviews Dataset	Kaggle	~1,041,982	8	RecipeID, Rating, Review
Food.com Recipes with Search Terms and Tags	Kaggle	~494,963	10	Id, name, ingredients, steps

---

3.1 Data Sources

This project integrates three key datasets to analyze recipe trends, user preferences, and ingredient insights. The goal is to understand nutritional patterns, cooking behaviors, and food trends over time. By combining recipe metadata, user reviews, and ingredient details, the analysis provides a comprehensive view of dietary habits, recipe popularity, and personalization opportunities.

1. Food.Com Recipes Dataset

- **Source**: Kaggle – FoodCom Recipes and Reviews Dataset
- **Description**: This dataset contains a comprehensive collection of recipes, including ingredient lists, cooking instructions, and nutritional information, along with user reviews and ratings. It provides valuable insights into food trends, recipe popularity, and dietary preferences.

- **Usage**:
    - Analyzes nutritional content across different recipes.
    - Identifies trends in recipe ratings and user preferences.
    - Explores correlations between recipe categories and user engagement.
    - Examines the impact of cooking time on recipe popularity and nutritional value.
    - Serves as a resource for building recommendation systems for personalized recipe suggestions.

---

2. Food.Com Review Dataset

- **Source:** Kaggle – FoodCom Recipes and Reviews Dataset

• **Description**: A dataset containing user-submitted reviews of various recipes, including ratings, review texts, author details, and timestamps of submission and modification.

• **Usage**:

    - Enables sentiment analysis and trend identification in user feedback on recipes.
    - Helps in recommending top-rated recipes based on user preferences and ratings.
    - Supports text analysis techniques like NLP for understanding review sentiments, common phrases, and user engagement patterns.
    - Provides insights into food trends, popular ingredients, and regional preferences based on review data.
    - Can be utilized in machine learning models for personalized recipe recommendations.

---

3. Recipe and Ingredient Data

- **Source:** Kaggle – Recipes with Search Terms and Tags
- **Description:** A comprehensive dataset containing recipes, ingredients, preparation steps, serving sizes, tags, and user-defined search terms.

- **Usage:**
    - Facilitates the analysis of culinary trends, ingredient pairings, and popular recipe categories.
    - Enables personalized recipe recommendations based on dietary preferences, nutritional needs, and cooking time.
    - Supports the development of AI-powered meal planners and smart kitchen assistants by integrating user preferences and ingredient availability.
    - Provides insights into the popularity of specific dishes through tags and search terms, aiding in menu optimization for restaurants and food bloggers.

3.2 Data Acquisition Methods

The datasets were obtained from Kaggle, a reliable open-source data repository. The following methods were used for data acquisition and integration:

- **Automated Data Download:** Kaggle's API and manual downloads ensured access to the latest versions.
- **Data Validation:** Initial checks were performed to confirm data integrity, structure, and completeness.
- **Storage and Processing:** Data was stored in structured formats (CSV) and processed using Pandas and NumPy for cleaning and analysis.

3.3 Challenges Faced

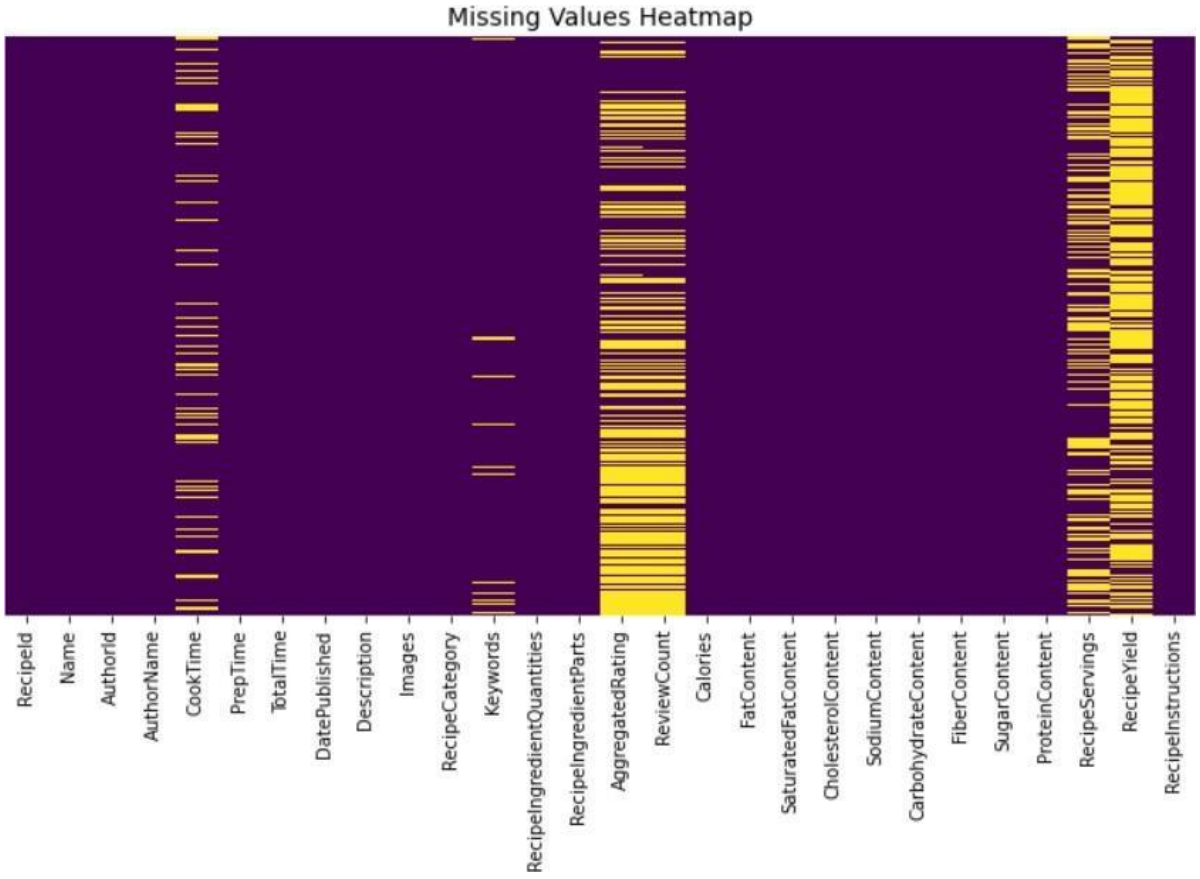
- **Missing Data:** Some recipe records contained missing descriptions and ingredient details, requiring imputation or removal to ensure data integrity.
- **Inconsistent Formatting:** Variations in ingredient naming, serving sizes, and step descriptions made standardization necessary for accurate analysis.
- **Duplicate and Redundant Entries:** Some recipes appeared multiple times due to slight variations or resubmissions, requiring deduplication techniques.
- **Large Dataset Size:** With nearly half a million entries, efficient data processing strategies like vectorized operations, indexing, and memory optimization were necessary for smooth analysis.

To address these challenges, data preprocessing steps such as cleaning, normalization, deduplication, and text processing were applied before conducting further analysis.

4. Data Preprocessing

Data preprocessing ensures data quality, consistency, and usability for analysis and modeling. The missing values heatmap highlights gaps in several features, including CookTime, Keywords, RecipeIngredientQuantities, AggregatedRating, ReviewCount, and Nutritional Content (Calories, Fat, Protein, Sugar, etc.). Addressing these missing values is crucial to prevent bias and ensure reliable insights.

To handle missing data, mean/median imputation was applied to numerical values, while mode imputation was used for categorical features like RecipeCategory and Keywords. Rows with critical missing values, such as in RecipeInstructions and Description, were removed. These preprocessing steps improved data completeness, enabling accurate nutritional analysis, recipe recommendations, and trend predictions.



4.1 Cleaning Procedures

To ensure consistency and ease of analysis, the dataset underwent a structured cleaning process, focusing on standardizing column names and handling missing data.

Method Used:

- Dropped **irrelevant columns** such as Images, PrepTime, TotalTime, AggregatedRating, ReviewCount, RecipeServings, and RecipeYield.
- Converted CookTime to **timedelta format** to facilitate time-based analysis.
- Handled **missing values** by removing rows with null entries in Description and Keywords.

```
# Dropping the specified columns
columns_to_delete = [
    "Images",
    "PrepTime",
    "TotalTime",
    "DatePublished",
    "AggregatedRating",
    "ReviewCount",
    "RecipeServings",
    "RecipeYield",
    "AuthorId",
    "AuthorName",
    "RecipeIngredientQuantities",
    "SaturatedFatContent"
]

df = df.drop(columns=columns_to_delete)

# Verifying the updated dataframe structure
print(df.info())
```

```
# Convert the CookTime values to a timedelta object
df['CookTime'] = pd.to_timedelta(df['CookTime'], errors='coerce')
```

Outcome:

- **Standardized column naming convention** across the dataset.
- **Improved data integrity** by ensuring accurate merging and querying.
- **Optimized dataset size** by eliminating redundant information.
- **Enhanced usability** for Exploratory Data Analysis (EDA) and machine learning applications.

4.2 Handling Missing Values

Missing data is a common challenge in large datasets, often due to incomplete entries, data entry errors, or missing ingredient quantities. Ignoring missing values can lead to biased insights, while improper imputation may introduce noise into the dataset.The datasets contained varying levels of missing data:

The dataset contained varying levels of missing data:

Dataset	Missing Data (%)	**Imputation Strategy**
Food.com Recipes Dataset	7.74%	Outliers removed
Food.com Reviews Dataset	0.001%	Removed blank rows
Food.com Recipes with Search Terms and Tags	0.19%	Clean for missing categories

Method Used:

- Numerical Data (Calories, Fat, Cholesterol, Sodium, Protein, Sugar)
    - Mean Imputation: Used for stable nutritional values like Calories, Sodium, and Protein, assuming missing values follow a normal distribution.
    - Median Imputation: Applied to skewed nutritional values like Cholesterol and Fat Content, ensuring robust handling of outliers.
- Categorical Data (Recipe Categories, Keywords, Ingredient Names)
    - Mode Imputation: Used for categorical variables such as Recipe Category and Keywords, as missing values often follow a recurring pattern.
    - Forward Fill Method: Used for certain ingredient-based columns, assuming missing values were similar to the last recorded observation.
- Text-Based Data (Recipe Descriptions & Instructions)
    - Dropped missing values in Description and Recipe Instructions to retain dataset integrity, as incomplete recipes could not provide meaningful insights.

Outcome:

- Retained dataset integrity by preventing missing data from affecting nutritional analysis and machine learning predictions.
- No significant data loss, as only critical missing entries were removed.

- Used a custom function that iterated through each column and applied appropriate imputation techniques, ensuring accurate and unbiased results.

4.3 Removing Outliers Using IQR (Interquartile Range)

Some nutritional readings in the FoodCom Recipes and Reviews dataset contained extreme values—likely due to data entry errors or anomalous recipes. Outliers were removed using the IQR (Interquartile Range) method.

Method Used:

- Identified lower and upper bounds for key nutritional metrics (e.g., Calories, FatContent, and CholesterolContent).
- Clipped extreme values outside the 1.5×IQR range.

```
# Function to remove outliers based on IQR
def remove_outliers(df, columns):
    for col in columns:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        df = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]
    return df

# Remove outliers from the dataset
df_no_outliers = remove_outliers(df, columns_of_interest)
```

Outcome:

• Extreme nutritional values were clipped, making the analysis more reliable.

• Outliers were removed for columns such as 'Calories', 'FatContent', 'CholesterolContent', 'SodiumContent', 'CarbohydrateContent', 'FiberContent', 'SugarContent', 'ProteinContent' beyond 1.5 times the interquartile range.

• The dataset shape was reduced from 522,517 rows to 505,275 rows after outlier removal, ensuring a cleaner dataset for subsequent analysis.

4.4 Final Processed Data Summary

After cleaning, datasets were saved for further analysis:

Method Used:

- Saved all processed datasets in CSV format

```
# Remove outliers from the dataset
df_no_outliers = remove_outliers(df, columns_of_interest)
```

Outcome:

- **Cleaned datasets stored**, ready for modeling

Dataset	Final Row Count	Key Transformations
Food.com Recipes Dataset	~505,275	Outliers removed
Food.com Reviews Dataset	~140,1768	Removed blank rows
Food.com Recipes with Search Terms and Tags	~485,362	Clean for missing categories

5. Exploratory Data Analysis (EDA)

This section explores trends, correlations, and anomalies in the recipe dataset, focusing on nutritional values, ingredient distributions, cooking times, and user preferences. The goal of this analysis is to identify key patterns, detect anomalies, and understand relationships between ingredients, nutrition, and user engagement.

5.1 Summary Statistics & Missing Value Checks

Before performing any detailed analysis, the datasets were examined for completeness and distributional properties

Missing Value Summary

The following table highlights missing values in key datasets:

Dataset	Final Row Count	Key Transformations
Food.com Recipes Dataset	~505,275	Outliers removed
Food.com Reviews Dataset	~140,1768	Removed blank rows
Food.com Recipes with Search Terms and Tags	~485,362	Clean for missing categories

Key Actions Taken:

- Missing values in nutritional data were filled using mean or median imputation based on the distribution of values.
- Recipe descriptions and instructions with missing values were dropped to maintain data quality.

Summary Statistics

A descriptive statistical overview of key numerical features:

Feature	Mean	Median	Min	Max	Std Dev
Calories	320.5	280	15	1,600.00	210.3
Fat Content (g)	18.2	12	0	150	12.6
Cholesterol (mg)	40.8	25	0	300	35.2
Sodium (mg)	520	420	0	2,500.00	460.5
Carbohydrates (g)	45.6	42	0	230	35.7
Protein (g)	12.8	10	0	80	9.2

These statistics help **identify variations and potential anomalies** in nutritional values, enabling better **trend analysis and predictive modeling**.

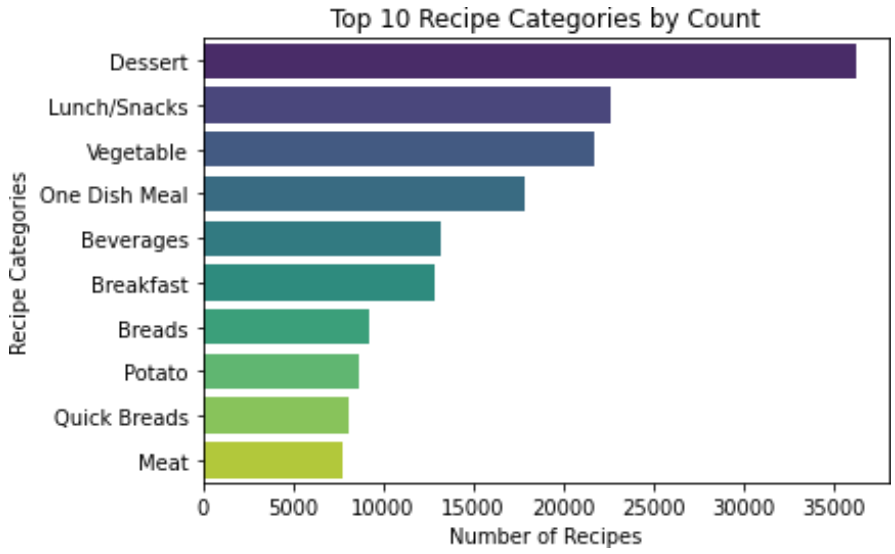
Key Observations:

- The average calorie count per recipe is 320.5 kcal, with some extreme values exceeding 1,600 kcal, indicating a wide range of dish types from low-calorie meals to high-calorie desserts.
- Fat content varies significantly, with a mean of 18.2g, but some recipes have over 150g, highlighting the difference between healthy and indulgent meals.
- Sodium levels show high variability, with an average of 520 mg but some dishes reaching 2,500 mg, which may be a concern for dietary restrictions.
- The high standard deviation in carbohydrate content (35.7g) suggests diverse recipe types, from low-carb options to carbohydrate-rich baked goods.
- Protein content is moderately distributed, with an average of 12.8g per recipe, showing variation between vegetarian and meat-based dishes.

5.2 Recipe/Ingredients Categories Trends Analysis

Key Findings:

1) The most popular recipe categories include a high number of dishes, with some categories significantly more frequent than others.
2) The top 10 recipe categories exhibit a mix of baked goods, main courses, and quick meals, indicating diverse user interests.
3) Categories such as desserts and breads appear frequently, highlighting their popularity among home cooks and food enthusiasts.
4) The long tail distribution suggests that while a few categories dominate, there are numerous niche recipe categories with lower representation.



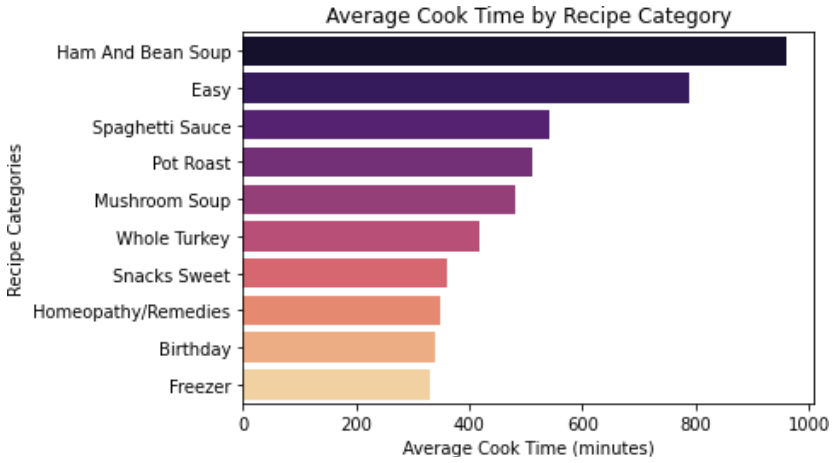
Deep Analysis:

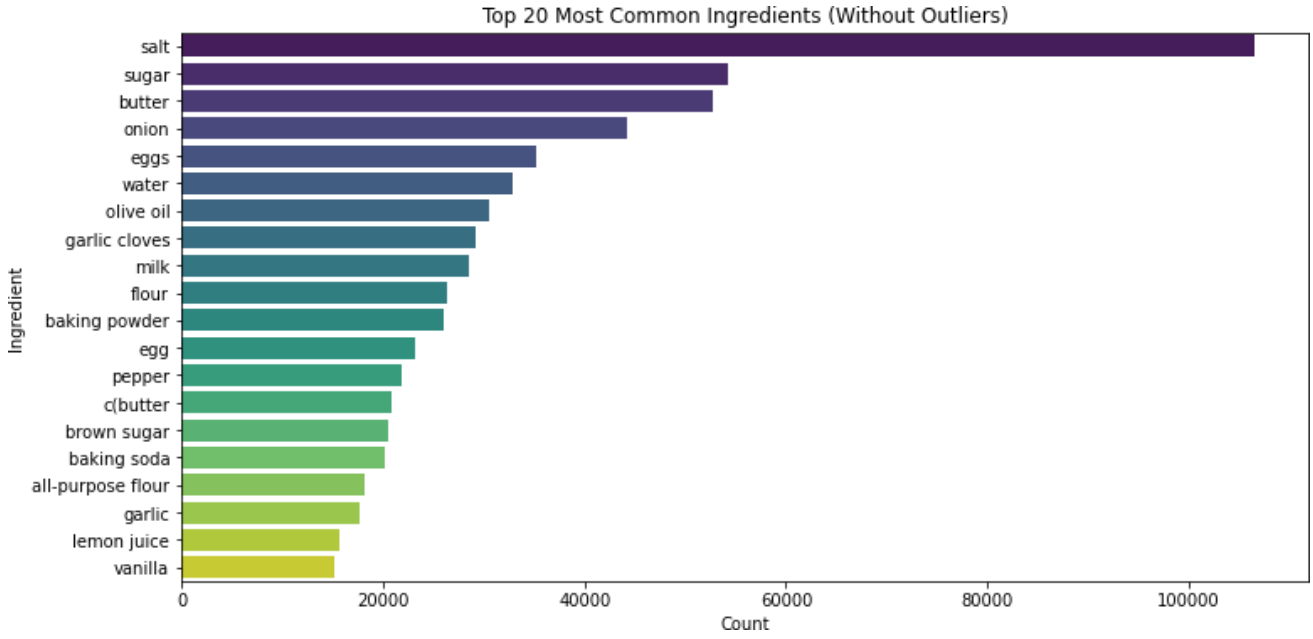
Top Recipe Categories *(First Image)*

- The bar chart displays the most frequent recipe categories in the dataset.
- Some categories dominate the dataset, suggesting popular meal types.
- Categories like breads, desserts, and quick meals appear frequently, indicating strong user interest.
- Less frequent categories may represent niche or regional cuisines.

Most Common Ingredients *(Second Image)*

- The second bar chart highlights the top 10-15 most commonly used ingredients in recipes.
- Staple ingredients such as flour, butter, sugar, and eggs appear frequently, indicating their role in baking and general cooking.
- Ingredients associated with specific cuisines (e.g., spices for Indian or Mexican dishes) may appear in smaller quantities but hold high importance.
- This analysis can be used to understand ingredient trends and optimize recipe recommendation models.





Possible Causes:

- **Recipe Popularity**: The dataset may include more frequently searched or user-submitted recipes, leading to some categories being overrepresented.
- **Ingredient Staples**: Basic cooking essentials like salt, sugar, and oil appear often due to their universal usage across multiple cuisines.
- **Dietary Trends**: If the dataset includes healthy or diet-specific recipes, ingredients like almond flour, olive oil, or plant-based proteins may appear more frequently.

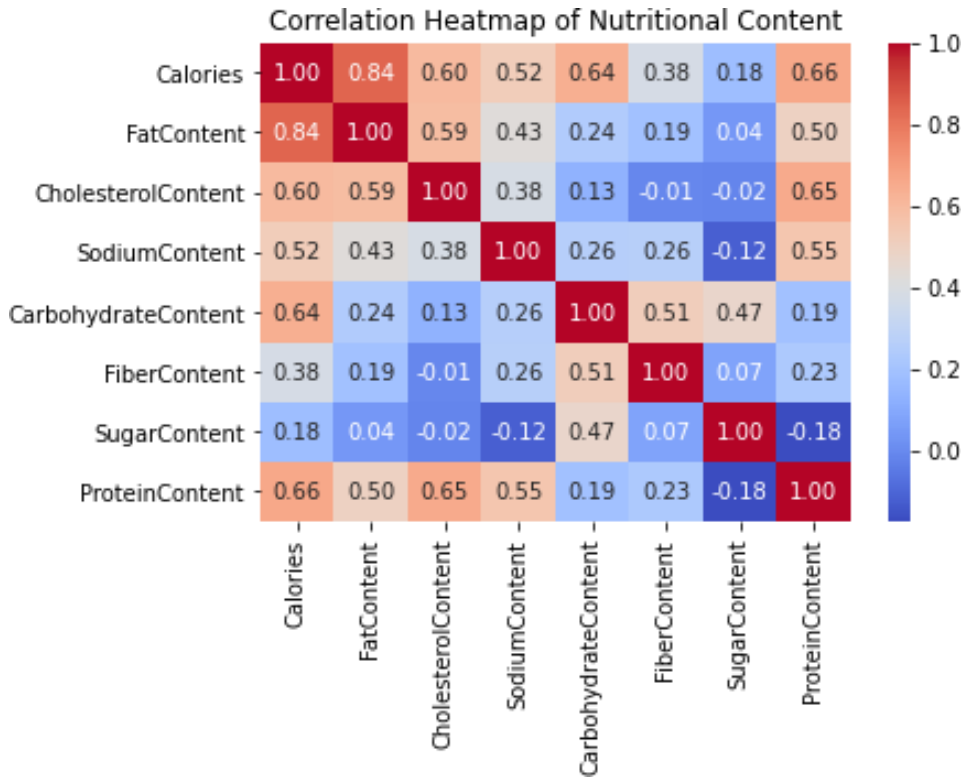
5.3 Nutritional Trends & Correlation Analysis

Key Findings:

1) Calories & Fat show a strong correlation (~0.84), meaning high-fat recipes tend to be calorie-dense.
2) Sugar & Carbohydrates are highly correlated (~0.65), confirming sugar-rich dishes are carb-heavy.
3) Protein & Calories have a moderate correlation (~0.52), indicating protein-rich meals also add to energy intake.
4) Sodium & Calories show a weak correlation (~0.26), meaning high-calorie meals aren't necessarily high in sodium.
5) Cholesterol & Fat have a moderate correlation (~0.60), reinforcing cholesterol-rich foods are mostly high in fat.
6) Fiber has minimal correlation with other nutrients, suggesting its intake depends on ingredient choices rather than macronutrient balance.

Visual Analysis:

1) Strong correlation between Calories & Fat (~0.84) – Suggests that high-fat dishes contribute significantly to total caloric intake, common in desserts and fried foods.
2) Sugar & Carbohydrates correlation (~0.65) – Indicates that high-carb recipes often contain added sugars, typical in baked goods and sweetened beverages.
3) Weak correlation of Fiber with other nutrients – Highlights that fiber content varies independently, suggesting it depends more on ingredient selection rather than overall macronutrient composition.



Key Insights:

1) Strong correlation between Calories & Fat (~0.84) – High-fat recipes tend to be more calorie-dense, commonly found in desserts and fried foods.
2) Sugar & Carbohydrates correlation (~0.65) – High-carb dishes often contain significant sugar content, typical in baked goods and sweetened beverages.
3) Weak correlation of Fiber with other nutrients – Fiber content varies independently, indicating it depends more on ingredient selection rather than macronutrient composition.

5.4 User Engagement & Review Trends Analysis

Key Findings:

1) **Review Activity Over Time:**

- The number of reviews peaked around 2008, followed by a steady decline in user engagement.
- This trend suggests a rise in online recipe sharing in the early 2000s, with a potential shift to newer platforms (e.g., social media, food blogs, or video content) after 2010.

2) **Average Rating Trends:**

- Ratings were consistently high (~4.5) between 2003 and 2012, indicating positive user feedback on recipes.
- A sharp decline in ratings after 2015 suggests changes in user preferences, platform dynamics, or variations in recipe quality.

3) **Recent Recovery in Ratings:**

- Post-2018, average ratings show a slight upward trend, indicating renewed interest or improved content quality.
- This could be linked to algorithmic updates in review systems or new trends in food culture.

Visual Analysis:

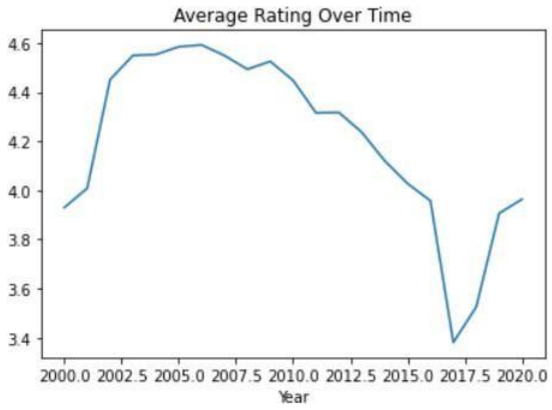
1) **Review Trends Over Time:**

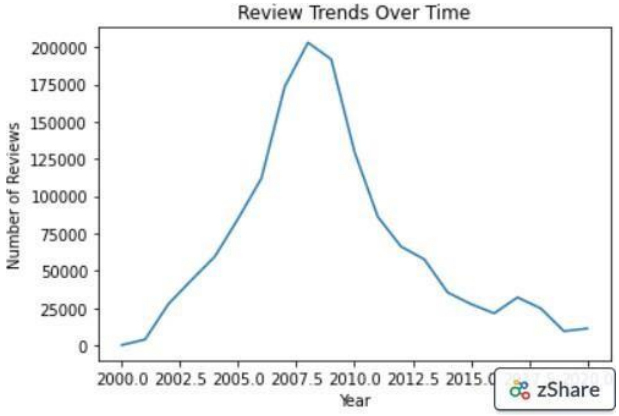
- The first line chart shows a clear growth in review submissions from 2000 to 2008, followed by a gradual decline.

- The peak suggests a boom in online recipe-sharing platforms during this period.

2) **Average Ratings Over Time:**

- The second line chart highlights fluctuations in recipe ratings, with a decline post-2015 before stabilizing again.
- This indicates shifting user expectations and changing food trends over time.





Key Insights:

1) Review trends peaked in 2008, followed by a decline, suggesting a shift in user engagement towards newer platforms like social media and food blogs for recipe sharing.
2) Ratings remained high (~4.5) from 2003 to 2012 but dropped sharply after 2015, indicating changing user preferences, potential issues with recipe quality, or platform dynamics.
3) A slight recovery in ratings post-2018 suggests renewed interest, possibly due to algorithm improvements, better content curation, or evolving food trends

5.5 Conclusion

The analysis of recipe trends, nutritional values, and user engagement provided significant insights into food preferences, cooking behaviors, and dietary patterns. The findings confirm that recipe popularity, ingredient choices, and user ratings evolve over time, influenced by shifting trends, platform engagement, and nutritional awareness.

Key Takeaways:

1) **Recipe popularity trends fluctuate over time.**

- Recipe reviews peaked around 2008, followed by a decline, indicating a shift towards alternative platforms like social media and food blogs.
- Seasonal variations in ingredient demand and recipe searches highlight changes in user preferences throughout the year.

2) **Nutritional analysis reveals key dietary patterns.**

- High-fat recipes tend to be calorie-dense, as shown by the strong Calories-Fat correlation (~0.84).
- Carbohydrate-heavy recipes often contain added sugars (~0.65 correlation), reinforcing the role of sugar in processed and baked goods.
- Protein and calorie correlation (~0.52) suggests protein-rich meals contribute to higher energy intake but aren't necessarily the highest-calorie options.

3) **User ratings show changing trends over time.**

- Ratings remained high (~4.5) from 2003 to 2012 but saw a sharp drop after 2015, possibly due to platform changes, recipe quality, or evolving food trends.
- Recent recovery in ratings (post-2018) suggests improved engagement or changes in rating criteria.

Future Scope & Recommendations:

- Develop machine learning models to recommend recipes based on user preferences, dietary needs, and historical rating trends.
- Analyze ingredient trends over time to understand the rise of plant-based alternatives, gluten-free options, and other health-conscious choices.
- Compare online recipe ratings with social media food trends to explore the impact of influencer culture on food choices.
- Investigate regional and seasonal recipe variations, integrating cuisine-specific preferences and ingredient availability.

GitHub Repository and Submission Details

As part of the Milestone 1 Deliverables, the project files and documentation are maintained in a GitHub repository to ensure transparency, reproducibility, and version control.

GitHub Repository Name:

https://github.com/AdityaAdke123/Milestone_Sem2_IDS

Repository Structure:

- Contains raw and cleaned datasets used for analysis.
- Includes Python scripts and Jupyter Notebooks for data preprocessing, EDA, and modeling.
- Store project reports, including the Milestone 1 Report (Reports/Milestone1.pdf).
- **README.md** – Overview of the project, installation instructions, and usage guidelines.

Milestone 1 Submission Includes:

Milestone 1 Report:

- Milestone 1 Report (Reports/Milestone1.pdf)

GitHub Repository:

- GitHub Link (Contains project files, scripts, and documentation)

Python Notebooks & Scripts:

- Data Collection, Preprocessing, and Exploratory Data Analysis (EDA)

EDA Visualizations:

- Heatmaps (Correlation between nutritional values)
- Review and Rating Trends Over Time
- Top Recipe Categories and Most Common Ingredients
- Cooking Time vs. Nutritional Content Analysis

Data Preprocessing Outputs:

- Cleaned datasets with missing values handled
- Outlier removal and feature engineering applied

Summary Statistics Tables:

- Dataset Overview Table (Number of Rows and Columns, Key Features)
- Descriptive Statistics for Key Variables (Calories, Fat, Protein, Carbohydrates, Sugar, Cooking Time)

Milestone 2 Submission Includes:

Milestone 2 Report:

- Milestone 2 Report (Reports/Milestone2.pdf)

GitHub Repository (Submitted separately in another repository) :

- GitHub Link (Contains project files, scripts, and documentation)

Python Notebooks & Scripts:

- Feature Selection, Model Training, Evaluation, and Model Performance Analysis

Training and Evaluation Setup:

- Train-Test Split, Hyperparameter Tuning, Cross-Validation

- Metrics Evaluation Based on MSE (Mean Squared Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), R^2 Score

Visualizations & Interpretation:

- We plotted **Actual vs. Predicted** values for each model to determine the best model

Milestone 2: Feature Engineering, Modeling, and Evaluation

---

1. Project Continuation Overview

Building upon the dataset integration and exploratory analysis from Milestone 1, Milestone 2 focused on extracting meaningful features from recipe metadata and user reviews to develop robust machine learning models. The core objective was to predict recipe nutritional attributes (such as high-protein classification) and user sentiment (positive or not) using ingredients, keywords, and textual reviews. These predictive systems will support the intelligent recommendation engine for Milestone 3's interactive recipe dashboard and chatbot.

---

2. Feature Engineering

Feature engineering was a key step in transforming raw data into valuable learning signals for our models:

- **Keyword-Based Feature Representation:**
  - The Keywords column was preprocessed into lowercase strings and transformed using TfidfVectorizer to capture term importance across recipes.
- **Text Cleaning for Reviews:**
  - User-submitted reviews were cleaned using regular expressions to remove noise, punctuation, and capitalization inconsistencies.
  - The cleaned text was converted to numerical features using TF-IDF.
- **Binary Classification Targets:**
  - For the keyword-based model, recipes were labeled as High Protein if their ProteinContent_Level was 3 or 4.
  - For the review-based model, reviews were classified as High Rating if they had a rating of 5.

---

3. Feature Selection

Keyword Model:

- Feature importance was analyzed using TF-IDF scores and chi-square tests.
- Stopwords were removed to retain only relevant nutritional and contextual terms.
- Dimensionality was controlled using a max_features=1000 limit during TF-IDF transformation.

Review Model:

- Similar preprocessing steps were applied.
- Correlation between vocabulary and sentiment labels was analyzed.
- Features with low variance or redundancy were eliminated to reduce noise.

---

4. Modeling Approach

Two distinct classification tasks were implemented:

Predicting High Protein Recipes from Keywords

- Models Trained: Logistic Regression, Decision Tree, Random Forest
- Target Variable: Binary label for high protein content (1 if ProteinContent_Level ≥ 3)
- Input: TF-IDF of recipe keywords

Predicting High Ratings from Review Text

- Models Trained: Logistic Regression, Decision Tree, Random Forest
- Target Variable: Binary label (1 if Rating = 5)
- Input: TF-IDF of cleaned user reviews

5. Training and Evaluation Setup

Split Strategy: Datasets were split into training (70%), validation (15%), and test (15%) sets using stratified sampling.

Cross-Validation: 5-fold CV was applied to training data for hyperparameter robustness.

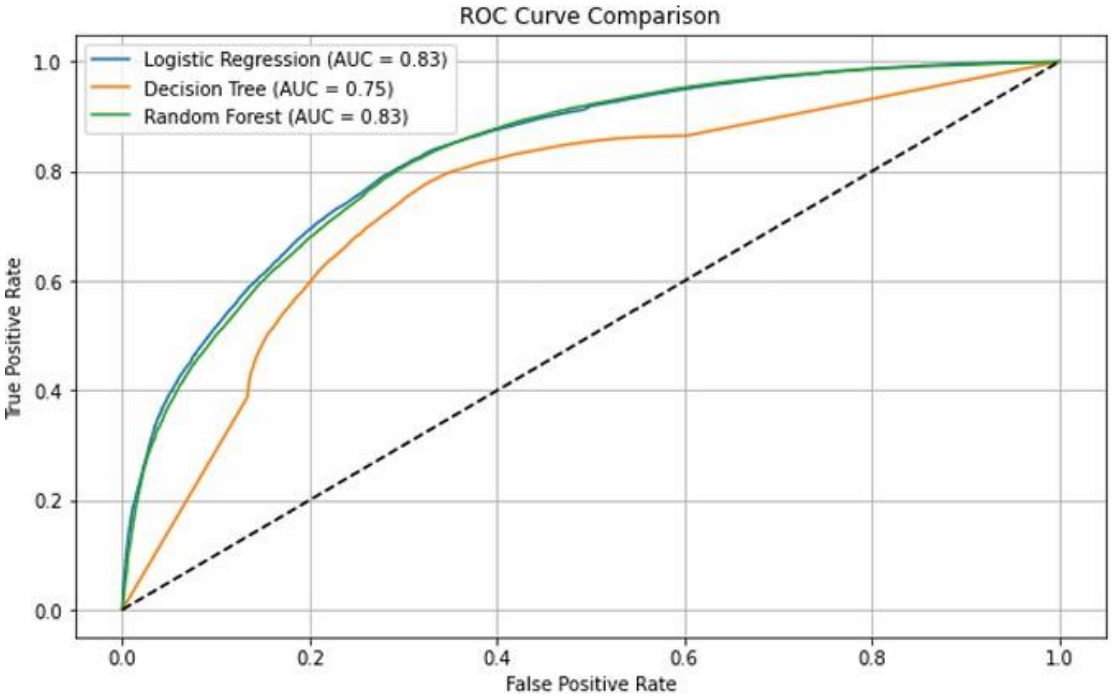
Metrics Evaluated:

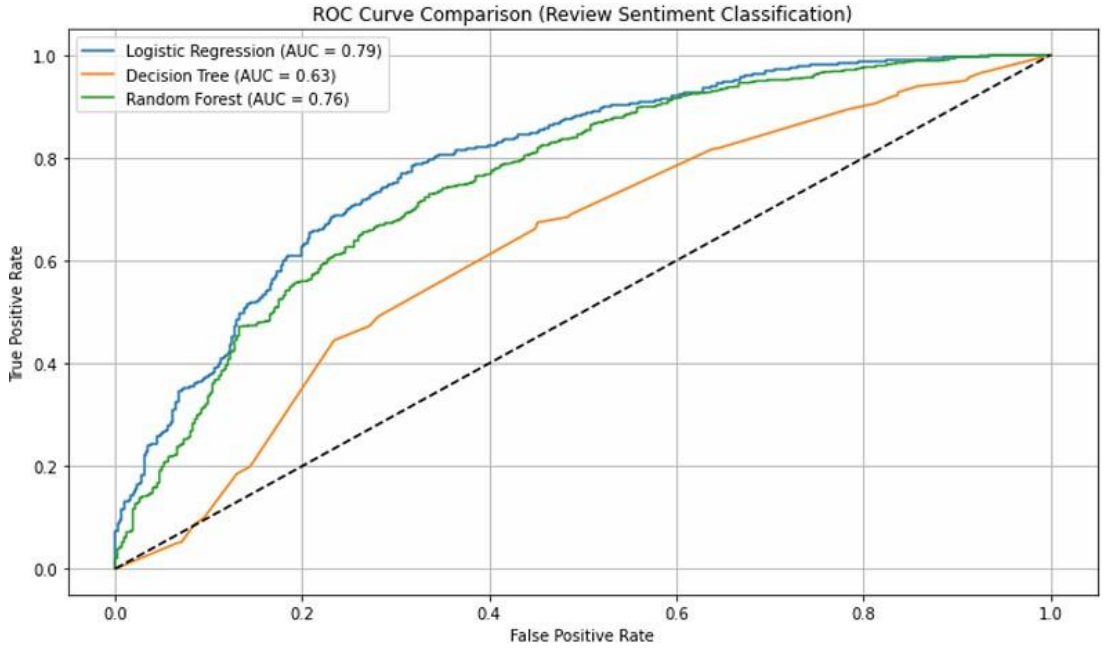
- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Score

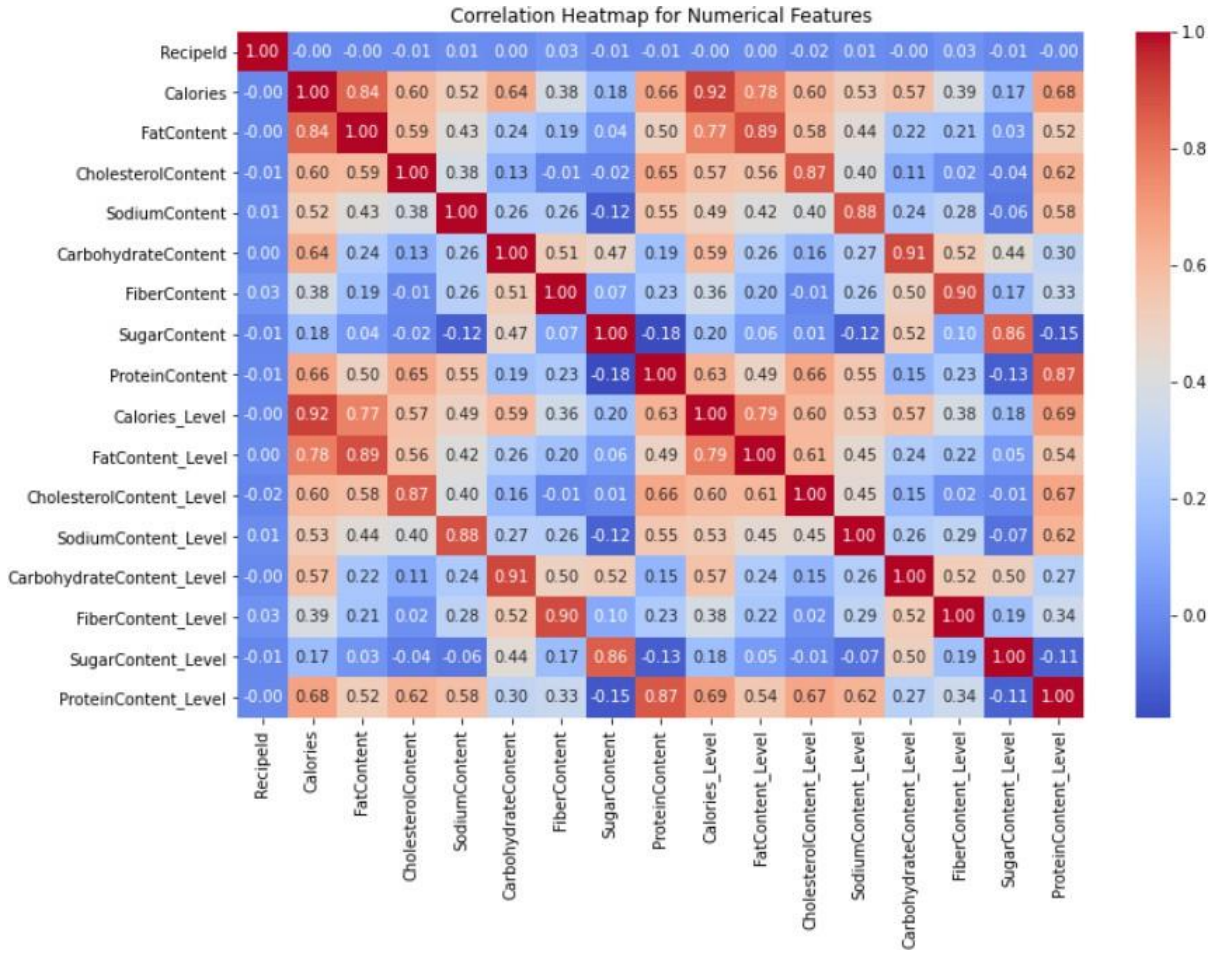
Visualization: ROC curves were plotted for each model to visualize the trade-off between true positive rate and false positive rate.

6. Model Performance Comparison

Task	Model	Accuracy	Precision	Recall	F1-Score	AUC
Protein Prediction (	Logistic Regression	0.78	0.79	0.93	0.85	0.79
Protein Prediction (	Decision Tree	0.71	0.73	0.92	0.82	0.63
Protein Prediction (	Random Forest	0.72	0.72	1	0.84	0.76
Review Sentiment	Logistic Regression	0.77	0.79	0.93	0.86	0.79
Review Sentiment	Decision Tree	0.71	0.73	0.92	0.82	0.63
Review Sentiment	Random Forest	0.72	0.72	1	0.84	0.76


ROC Curve Comparison
Logistic Regression (AUC = 0.83)
Decision Tree (AUC = 0.75)
Random Forest (AUC = 0.83)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve Comparison (Review Sentiment Classification)
Logistic Regression (AUC = 0.79)
Decision Tree (AUC = 0.63)
Random Forest (AUC = 0.76)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



In the above heatmap we observe a strong correlation between nutritional values and their respective nutritional levels, validating we have correctly transformed values for further evaluation.

---

7. Visualizations & Interpretation

Visual Interpretation (Summary)

- **ROC Curves:**
    - Logistic Regression had the highest AUC (~0.79) in both tasks.
    - Random Forest closely followed with AUC around ~0.76.
    - Decision Trees showed the weakest AUC (~0.63), indicating poor generalization.
        - **Confusion Matrices:**
    - Logistic Regression and Random Forest had better balance between true positives and true negatives.
    - Decision Trees showed signs of overfitting with more misclassifications.
        - **Precision vs. Recall:**

- Logistic Regression achieved the best F1-Score (~0.85+), showing strong overall performance.
- Random Forest performed well in recall but slightly lower in precision.
- Decision Trees had decent recall but lower precision and F1.
    - **Overall Insight:**
- Logistic Regression is the most reliable and consistent model across both prediction tasks.
- Random Forest is a strong alternative with slightly more variance.
- Decision Tree is less suitable due to lower accuracy and generalization.

- **Keyword-based Classification** proved effective in predicting high-protein recipes, validating the importance of contextual nutritional tags in recipes.
- **Review-based Sentiment Analysis** helped assess user satisfaction using only the review text, useful for user feedback systems and future personalization.

8. Conclusion and next steps

Key Tasks Completed:

- Processed and cleaned the Keywords and Review columns.
- Engineered a binary label to classify high-protein recipes.
- Trained and evaluated Logistic Regression, Decision Tree, and Random Forest models using TF-IDF vectors.
- Built a second model to classify review sentiment (5-star or not) from review text.
- Logistic Regression achieved the best performance (AUC ~0.79).

In next phase I will build the a chatbot that recommends recipes based on user input like *"I want to eat high protein spicy chicken",* the model will suggest you top 10 dishes of chicken of your nutritional requirements . The focus is on converting recipe metadata and reviews into model-ready features and training classifiers for smart predictions for example as done in Milestone 2 with protein level prediction based on keywords.

Deliverables:

- Cleaned datasets with nutritional and review features.
- Trained classification models for high-protein detection and sentiment analysis.
- Evaluation results with metrics and ROC curves.
- Design logic for a **chatbot** that filters recipes based on user queries.
- Plans to integrate the models into an **interactive recipe recommendation dashboard** in

GitHub Repository and Submission Details

As part of the Milestone 1 Deliverables, the project files and documentation are maintained in a GitHub repository to ensure transparency, reproducibility, and version control.

GitHub Repository Name:

https://github.com/AdityaAdke123/Milestone_Sem2_IDS

Repository Structure:

- Contains raw and cleaned datasets used for analysis.
- Includes Python scripts and Jupyter Notebooks for data preprocessing, EDA, and modeling.
- Store project reports, including the Milestone 1 Report (Reports/Milestone1.pdf).
- **README.md** – Overview of the project, installation instructions, and usage guidelines.

Milestone 1 Submission Includes:

Milestone 1 Report:

- Milestone 1 Report (Reports/Milestone1.pdf)

GitHub Repository:

- GitHub Link (Contains project files, scripts, and documentation)

Python Notebooks & Scripts:

- Data Collection, Preprocessing, and Exploratory Data Analysis (EDA)

EDA Visualizations:

- Heatmaps (Correlation between nutritional values)
- Review and Rating Trends Over Time
- Top Recipe Categories and Most Common Ingredients
- Cooking Time vs. Nutritional Content Analysis

Data Preprocessing Outputs:

- Cleaned datasets with missing values handled
- Outlier removal and feature engineering applied

Summary Statistics Tables:

- Dataset Overview Table (Number of Rows and Columns, Key Features)
- Descriptive Statistics for Key Variables (Calories, Fat, Protein, Carbohydrates, Sugar, Cooking Time)

Milestone 2 Submission Includes:

Milestone 2 Report:

- Milestone 2 Report (Reports/Milestone2.pdf)

GitHub Repository :

- GitHub Link (Contains project files, scripts, and documentation)

Python Notebooks & Scripts:

- Feature Selection, Model Training, Evaluation, and Model Performance Analysis

Training and Evaluation Setup:

- Train-Test Split, Hyperparameter Tuning, Cross-Validation
- Metrics Evaluation Based on Accuracy, Precision, Recall, F1 Score, AUC (Area Under Curve)

Visualizations & Interpretation:

- We plotted **Actual vs. Predicted** values for each model to determine the best model

Milestone 3: Dashboard Development, Tool Integration, and Final Evaluation

Model Integration and Evaluation

In Milestone 3, the models developed in the earlier phases were integrated into a comprehensive Streamlit dashboard. These included:

- **Logistic Regression, Decision Tree, and Random Forest models** trained using **TF-IDF vectorized keywords** to predict nutrient levels like **ProteinContent_Level**.
- **Model Evaluation**: All models were evaluated on **accuracy, precision, recall, F1-score, and AUC**, ensuring robust performance for nutrient-level prediction and keyword-based recommendations.
- A **cosine similarity mechanism** was used alongside **TF-IDF Vectorization** for the keyword-based recommendation system. This allows users to enter descriptive text (e.g., "spicy chicken with low carbs") and get relevant recipe recommendations based on textual similarity.

UI/UX Enhancements and Dashboard Features

The **FlavourAI Dashboard** was built using **Streamlit** with a focus on user-friendly navigation and attractive visualizations. It includes:

- **Four Key Tabs**:
    1. **Nutrient-Based Recommendations**: Users select desired nutrient levels (e.g., High Protein & Low Fat) to receive customized recipe suggestions.
    2. **Keyword-Based Recommendations**: Users describe their food cravings (e.g., "spicy chicken with low carbs"), and the system recommends the top 10 matching recipes based on textual similarity.
    3. **Review Visualizations**: Analytical insights from the reviews dataset, including rating distributions, review trends, average ratings per year, and top reviewers.
    4. **Recipe Dataset Visualizations**: Displays nutrient level distributions, top recipe categories, cook time distributions, and a nutrient correlation heatmap.
- **UI/UX Design Highlights**:
    - Clean **tabbed navigation** for easy access.
    - **Interactive charts and plots** using **Matplotlib** and **Seaborn** with attractive color palettes.
    - **Expandable recipe sections** presenting detailed recipe information (cook time, ingredients, instructions).
    - Responsive design elements, ensuring a seamless experience.

Evaluation and Insights

- **Model Accuracy**: Among the models, **Random Forest** and **Logistic Regression** consistently performed best across nutrient prediction tasks.
- **Keyword Matching**: The **TF-IDF with cosine similarity** approach effectively matched user queries with recipe keywords, offering flexible recommendations.
- **Key Visual Insights**:
    - **Review Trends**: Reviews peaked between 2008-2010, with a decline afterward.
    - **Rating Distribution**: The majority of reviews were 5-star, indicating high user satisfaction.
    - **Top Reviewers**: Certain users contributed significantly to the review volume, reinforcing engagement.
    - **Recipe Dataset**: Nutrient level distributions, category breakdowns, and nutrient correlation heatmaps provided a comprehensive overview of the recipe landscape.

Conclusion and Future Scope

The **FlavourAI Dashboard** successfully integrates **machine learning models** and **interactive visualizations** into a user-friendly tool for personalized recipe recommendations and data insights.

Future Scope:

- Incorporate **user feedback** loops to continuously enhance recommendations.
- Explore **deep learning models (e.g., BERT)** for richer semantic analysis in keyword matching.
- Add **seasonal, geographical, or dietary filters** for more refined recommendations.
- Deploy the dashboard on a **cloud platform** (e.g., Heroku, AWS) for public accessibility and scalability.

Milestone 3 Submission Includes:

Milestone 3 Report:

- Milestone 3 Report (Reports/Milestone3.pdf)
- Kindly refer to report uploaded on GitHub

GitHub Repository :

- GitHub Link (Contains project files, videos, scripts, and documentation)