جامعة أم القرى

UMM AL-QURA UNIVERSITY

Name	ID
Remas Al-Quthami	444001952
Sara Al-otaibi	444004152

Introduction

In the current digital age, data plays a crucial role in enhancing the understanding of consumer behavior and improving healthcare. This report aims to explore three key analytical areas: predicting diabetes among the Pima Indians, market basket analysis in e-commerce, and sentiment analysis of food reviews on Amazon.

We begin with the diabetes prediction report, where we utilize machine learning algorithms to analyze data and predict the likelihood of disease onset, aiding in the development of tools for early detection of health conditions. Next, we move on to market basket analysis, which reveals the purchasing patterns followed by customers, enabling companies to enhance their marketing strategies and increase customer satisfaction. Finally, we examine sentiment analysis of food reviews on Amazon to gain deeper insights into customer opinions and needs, thereby enriching their experiences and positively influencing sales strategies.

Pima Indians Diabetes Prediction Report

1. Introduction

This report provides a detailed analysis of diabetes prediction using the Pima Indians Diabetes dataset. The primary objective is to predict whether a patient is likely to have diabetes based on diagnostic measurements, using machine learning models. The dataset consists of medical features such as glucose level, BMI, blood pressure, and others, along with a target variable indicating diabetes presence.

2. Dataset Description

The Pima Indians Diabetes dataset contains information for 768 patients, including various health in- dicators such as Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and the outcome variable (Outcome) which indicates the presence (1) or absence (0) of diabetes. The dataset is publicly available on Kaggle and was used to train machine learning models for diabetes prediction.

3. Data Preprocessing

To prepare the dataset for analysis, the following steps were performed:

• Checking for Missing Values: The dataset was inspected for missing values using isnull().sum(), revealing no explicit missing values, but some columns had zero values where they were not possible, such as in Glucose, Blood Pressure, and BMI.

• Replacing Zero Values: Columns such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI had zero values replaced with the median of each respective column, ensuring realistic values and reducing bias.

• Feature and Target Split: The dataset was split into features (X) and the target variable (y). The target variable (Outcome) was used to train the machine learning models.

• Train-Test Split: The data was split into training (80%) and testing (20%) sets using train test split() to ensure proper model evaluation.

4. Model Training and Evaluation

4.1 Gaussian Naive Bayes

A Gaussian Naive Bayes model was used to classify the patients based on their medical features. Gaussian Naive Bayes assumes that the features follow a normal distribution, making it a good choice for continuous features such as Glucose and BMI.

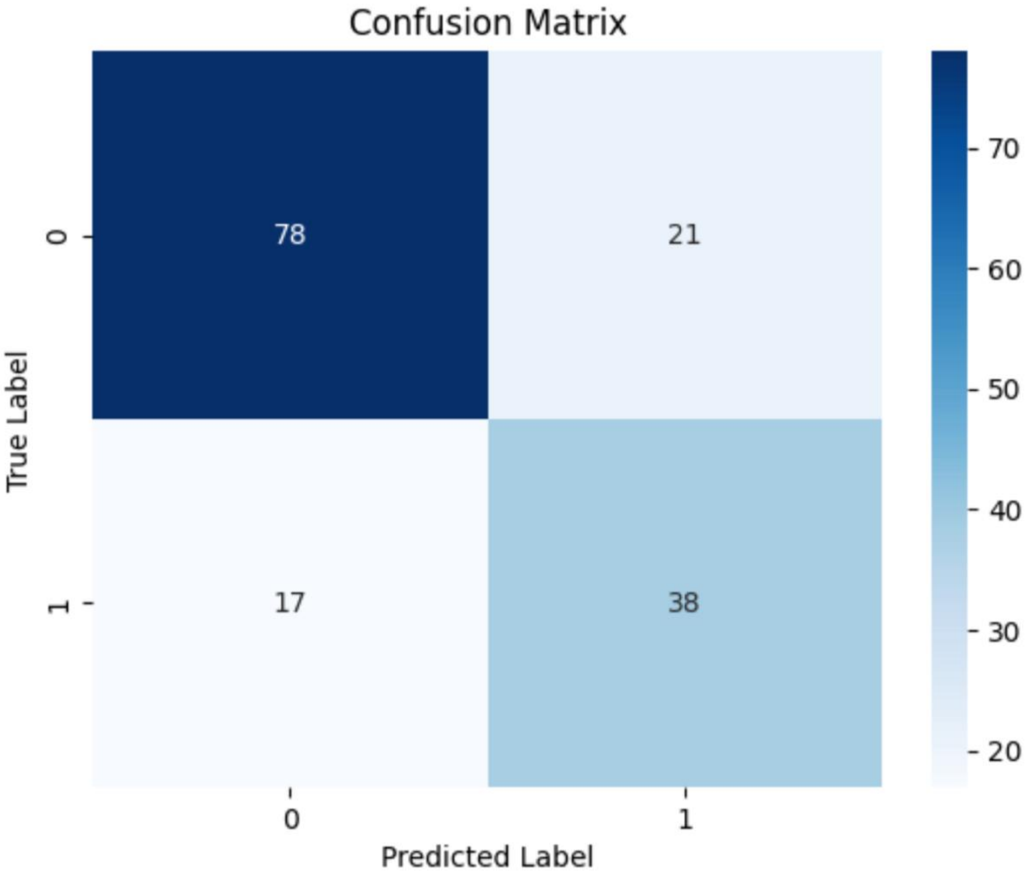
The model was trained on the training data (X train, y train) and evaluated on the test set (X test, y test).

4.2 Evaluation Metrics

•Accuracy Score: The model achieved an accuracy of 74.68% on the test data, which indicates that the model correctly predicted diabetes presence or absence in roughly three-fourths of the cases.

• Confusion Matrix: The confusion matrix was used to analyze the performance of the model in more detail:
– True Positives (TP): Correct predictions of patients with diabetes.
– True Negatives (TN): Correct predictions of patients without diabetes.
– False Positives (FP): Patients incorrectly predicted as having diabetes.
– False Negatives (FN): Patients incorrectly predicted as not having diabetes.

The confusion matrix was visualized using a heatmap, providing an intuitive understanding of the model’s strengths and weaknesses.



• Classification Report: The classification report provided a summary of precision, recall, and F1-score for each class:

– Precision: The model's ability to correctly identify positive predictions.
– Recall: The model's ability to find all relevant instances of diabetes.
– F1-Score: A harmonic mean of precision and recall, providing a balance between the two.

5. Results and Discussion

• The Gaussian Naive Bayes model performed reasonably well, with an accuracy of 74.68%, indicating that it can be useful for initial diabetes prediction. However, the model had some limitations in distinguishing between false positives and false negatives.

• The confusion matrix heatmap indicated that false negatives were more prevalent, which is critical in medical contexts as failing to diagnose diabetes can lead to serious health consequences.

• Precision and Recall: The precision was higher for predicting patients without diabetes, while recall was higher for patients with diabetes, highlighting the trade-off between identifying all pos- itives and avoiding false positives.

6. Conclusion

The analysis of the Pima Indians Diabetes dataset using a Gaussian Naive Bayes classifier showed promising results for diabetes prediction. The accuracy, confusion matrix, and classification report helped evaluate the performance of the model. Although the Gaussian Naive Bayes model was effective in predicting diabetes for many patients, further improvements are necessary to reduce the occurrence of false negatives, which is crucial for medical applications.

Market Basket Analysis on E-commerce Dataset

1. Introduction

This project aims to analyze Brazilian e-commerce data using association rule learning techniques. Algorithms such as Apriori and FP-growth were applied to extract frequent patterns and relationships between products that customers tend to purchase together, with the goal of improving business strategies such as targeted marketing, product bundling, and inventory management. However, neither algorithm produced significant associations or meaningful results.

2. Data Preparatio

2.1 Data Overview

We used three main datasets:

- olist_order_items_dataset.csv: Contains information about the products listed in each order.
- olist_orders_dataset.csv: Contains order details such as order ID, customer ID, order status, and purchase date.
- olist_products_dataset.csv: Contains product details such as category names and dimensions.

2.2 Handling Missing Values

We checked for missing values and found none of significant impact.

2.3 Data Merging

We merged the datasets to create a unified dataset for analysis:

- Merged the orders data with order items data using order_id.

```
total_orders = pd.merge(orders_df, order_items_df, on='order_id')
```

- Then, merged the result with product details to add additional features.

```
product_orders = pd.merge(total_orders, products_df, on='product_id')
```

2.4 Data Filtering

We filtered the data to include the top 1000 most popular products and the top 2000 orders to focus the analysis on the most relevant data.

2.5 Top 10 Most Ordered Products

A bar chart was used to display the top 10 most ordered products:

- Most ordered product: Product with ID 314663af, with 500 orders.
- Least ordered product: Product with ID c1e95ad7, with around 300 orders.

2.6 Transaction Matrix Creation

We created a binary transaction matrix where each row represents an order and each column represents a product: “1” indicates the product was purchased, and “0” indicates it was not.

3. Association Rule Extraction

3.1 Applying Apriori and FP-growth Algorithms

Both algorithms were applied using a minimum support (min_support) value of 0.5%. However, neither algorithm produced any significant results or meaningful associations between products.

4. Model Evaluation

Despite the absence of associations, the algorithms were evaluated based on the following metrics:

- Support: The proportion of orders that contain a specific set of items.
- Confidence: The probability of purchasing an associated product when the base product is purchased.
- Lift: How much purchasing one product increases the likelihood of purchasing another product compared to random chance.

5. Insights

5.1 Top Products and Categories

The most frequently ordered products and categories were identified, helping to understand customer preferences.

5.2 Absence of Associations

Despite correctly applying the Apriori and FP-growth algorithms, no significant associations between products were found. This is attributed to the nature of the data itself, as there appear to be no strong correlations between the products purchased together. This suggests that the dataset may not contain enough frequent purchase patterns to extract meaningful association rules, rather than an issue with the algorithm settings or implementation.

6.Conclusion

In conclusion, this project aimed to discover patterns and associations in Brazilian e-commerce data using the Apriori and FP-growth algorithms. Despite proper data preparation and the correct application of both algorithms, no significant associations were identified. This indicates that the data does not exhibit strong purchase relationships between products, limiting the ability to extract valuable association rules. However, the insights gathered on the most frequently ordered products remain useful for understanding customer preferences. Further analysis with different datasets or approaches may be needed to uncover hidden patterns that can better support business strategies.

Sentiment Analysis Report on Amazon Food Review

1.Introduction

This report aims to analyze customer sentiment for Amazon Fine Food Reviews using natural
language processing (NLP) and machine learning models. Sentiment analysis helps to determine whether a review is positive or negative, providing insights into customer satisfaction and product quality. We used TF-IDF vectorization for feature extraction and
trained multiple classifiers to evaluate their performance in classifying customer sentiment.

2.Dataset Description

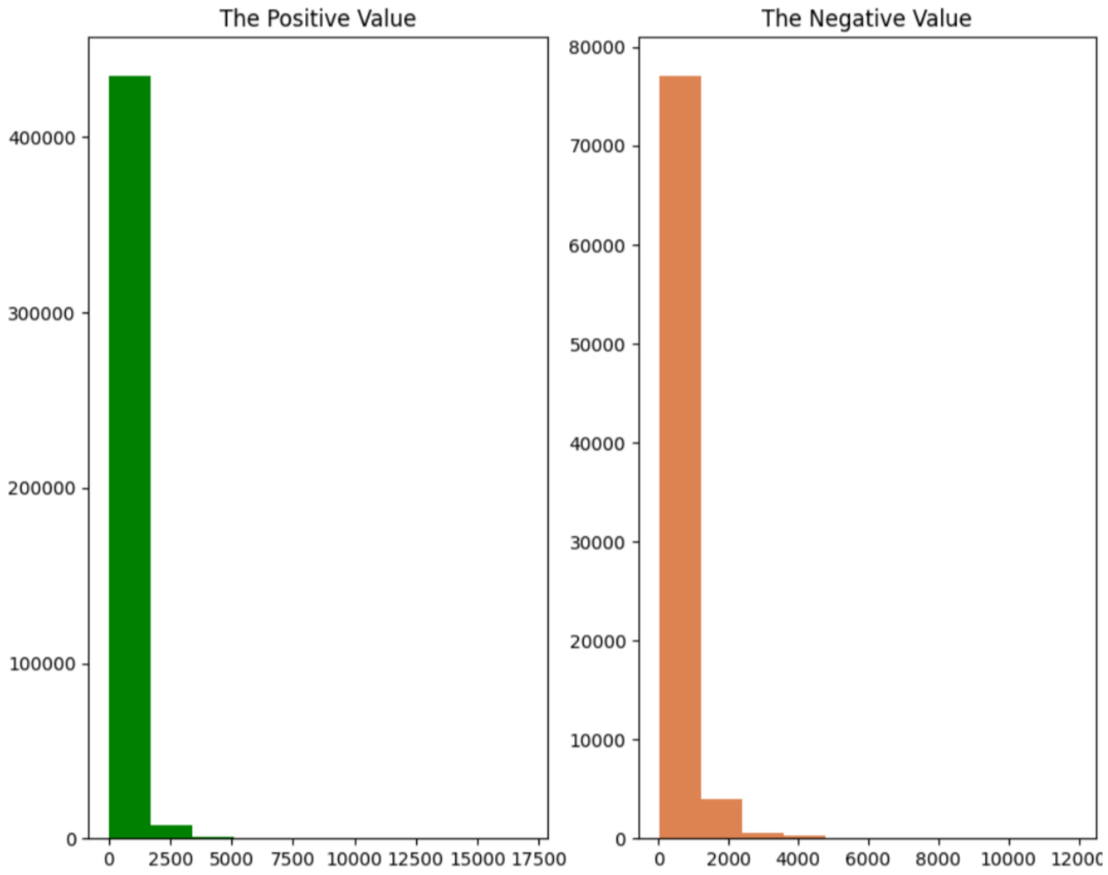
The dataset used for this analysis is the Amazon Fine Food Reviews dataset. The dataset contains reviews, ratings (scores), product information, and user data. Each review includes a Score ranging from 1 to 5, which we used to label reviews as positive or negative. Scores 1 and 2 were considered negative, while scores 4 and 5 were labeled as positive. Reviews with a score of 3 were excluded, as they were considered neutral.

3.Data Preprocessing

To prepare the data for analysis, the following preprocessing steps were performed:

• Text Cleaning: The Text column was converted to lowercase, numbers and punctuation were removed, and whitespace was stripped.

• Sentiment Mapping: The Score column was used to create a binary sentiment label, where scores greater than 3 were labeled as positive (1), and scores less than 3 were labeled as negative (0).



• Combining Summary and Text: The Summary and Text columns were concatenated to create a single text column containing all relevant information for analysis.

4. Feature Extraction

We used TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to transform the text data into numerical features suitable for model training. TF-IDF provides a measure of the importance of each word relative to the entire dataset, giving more weight to less frequent but significant words.

5. Model Training and Evaluation

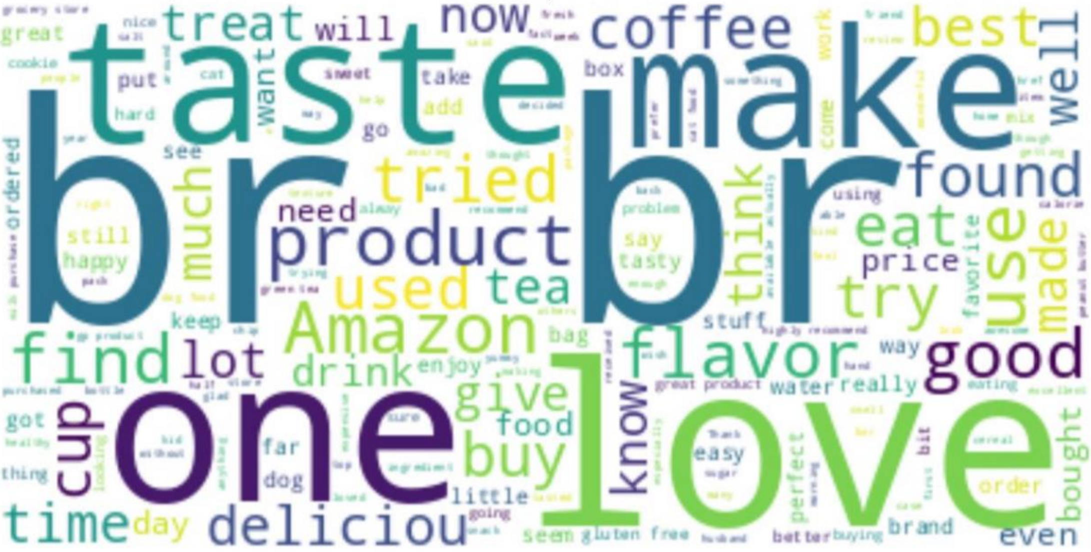
Two machine learning models were trained and evaluated to classify the sentiment:

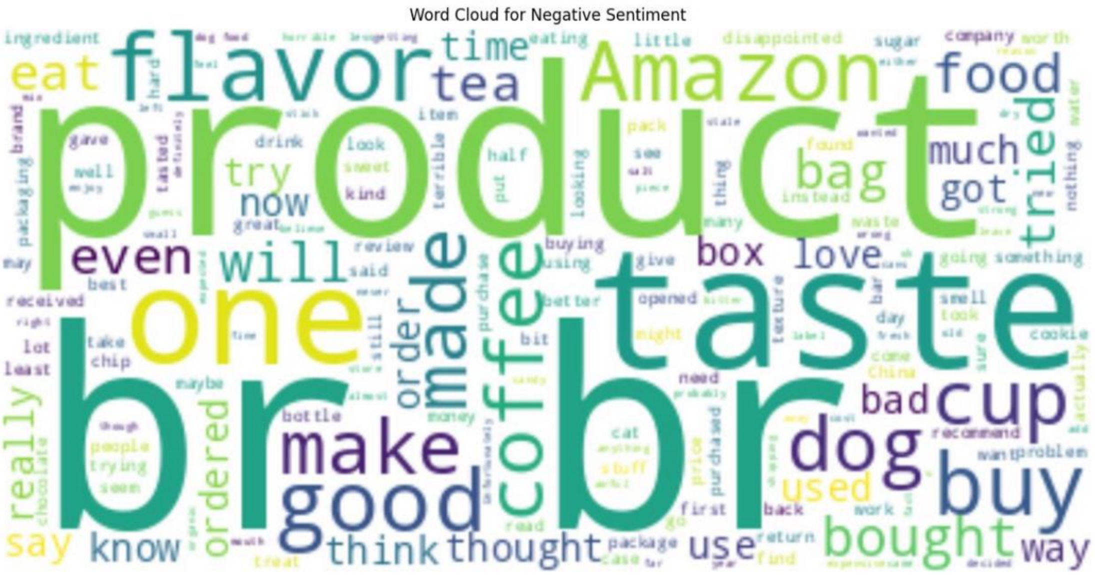
• Logistic Regression: This is a linear model that is widely used for binary classification tasks. It was trained on the TF-IDF vectors and achieved an accuracy of approximately 92%. The classification report showed good precision and recall for both positive and negative sentiment classes.

• Multinomial Naive Bayes (MNB): This model is commonly used for text classification problems. The MNB model was trained on the TF-IDF vectors and achieved an accuracy of approximately 89%. The confusion matrix and classification report indicated that the model performed well, especially with positive sentiment, but slightly less so with negative sentiment compared to Logistic Regression. Visualizatio.

• Word Cloud: Word clouds were generated for both positive and negative reviews to visualize the most common words in each type of sentiment. Words such as "great," "love," and "excellent" were common in positive reviews, whereas words like "disappointed" and "bad" were common in negative reviews.

wordcloud for positive value





Word Cloud for Negative Sentiment

• Review Lengths: Histograms were used to visualize the distribution of review lengths
for positive and negative reviews. This helped understand the patterns in the length of reviews for each sentiment type.
• Common Words: A bar chart was created to display the 10 most common words in the reviews. This provided insight into frequently used terms across all reviews.

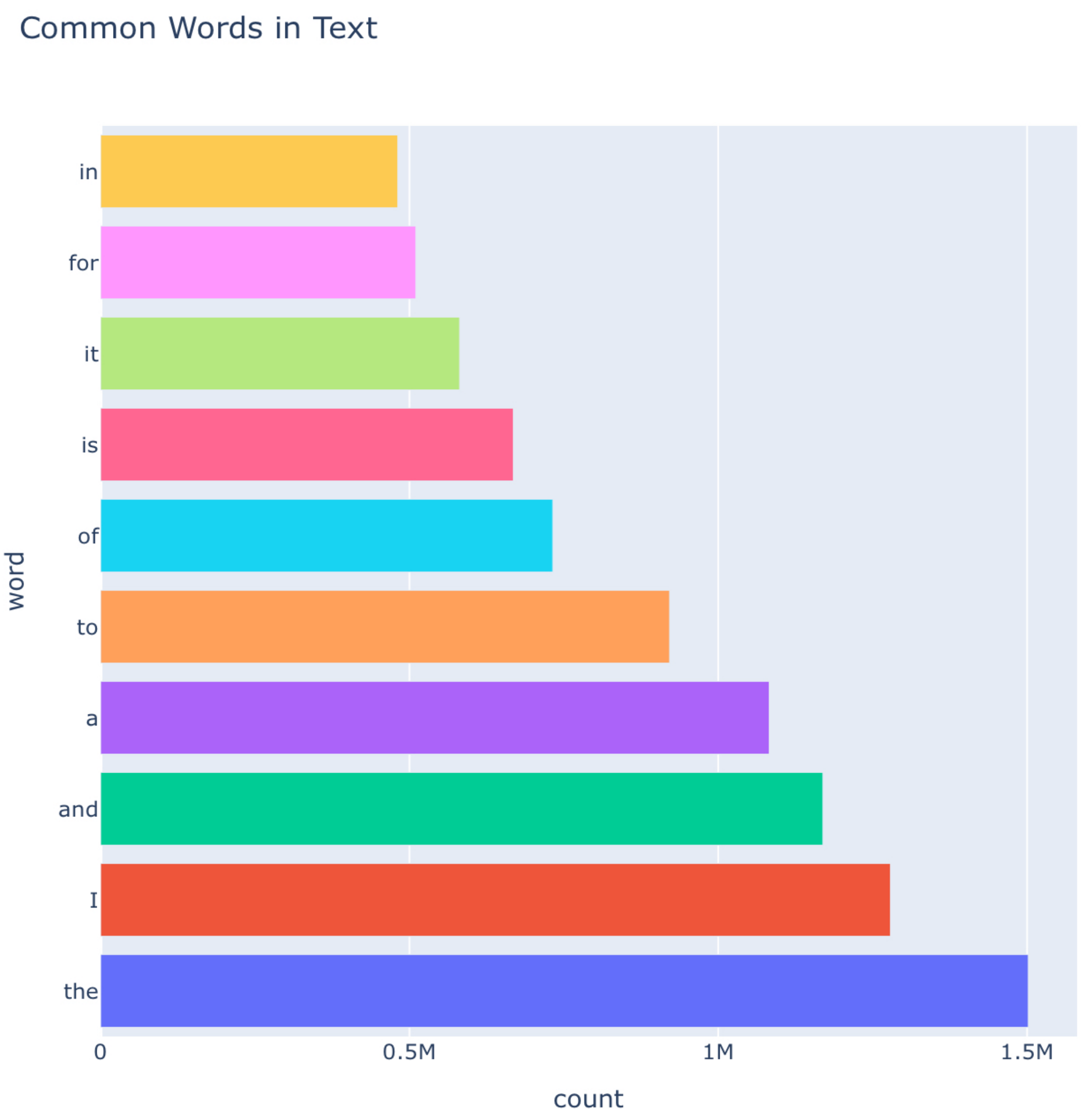
6.visualizing

shows a visualization of the most common words found in a given text. The visualization is created using the Plotly data visualization library in Python.

The chart has the following features:

1. The x-axis shows the count of each word, while the y-axis lists the actual words.
2. The colors used for the bars are defined in a list of hex color codes.
3. The `fig.update_traces()` function is used to update the marker color of the bars based on the predefined color list.
4. The `fig.show()` function is called to display the final visualization.

The resulting visualization shows the frequency of the most common words in the text, with the most frequent words displayed as the longest bars. This type of visualization can be useful for quickly identifying the key terms and themes present in a body of text.



7. Results and Discussion

- The Logistic Regression model performed better than Multinomial Naive Bayes, achieving a higher accuracy score and showing a better balance between precision and recall.
- The word clouds provided a good overview of the language used in different sentiment classes. Positive reviews had more descriptive and enthusiastic words, whereas negative reviews often contained direct complaints or disappointment.
- The review length analysis revealed that positive reviews were generally longer than negative ones, indicating that users tend to elaborate more when they are satisfied.

8. Conclusion

In conclusion, sentiment analysis on Amazon Fine Food Reviews using Logistic Regression and Multinomial Naive Bayes provided valuable insights into customer opinions. Logistic Regression proved to be more effective in classifying the sentiment, with higher accuracy and balanced performance metrics. The visualizations helped us understand the patterns and characteristics of positive and negative reviews. These insights could be useful for companies to identify areas for improvement, enhance customer satisfaction, and gain competitive advantages.

Conclusion: A Comprehensive Vision for Improving Business Decisions and Health Outcomes Through Data.

In this report, we analyzed food reviews on Amazon using sentiment analysis techniques to better understand customer opinions and needs.

We also conducted market basket analysis on an e-commerce dataset to identify products that customers tend to purchase together. This analysis helps companies improve their sales and marketing strategies by offering personalized promotions and complementary products, thereby increasing customer satisfaction and enhancing sales volume.

Furthermore, the analysis included studying diabetes data from the Pima Indians using machine learning algorithms to predict the likelihood of disease onset. This work assists in developing tools that contribute to early disease detection and provide appropriate care for patients.