جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis (2) Report

Dr. Omaima Fallatah

By Students:

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.

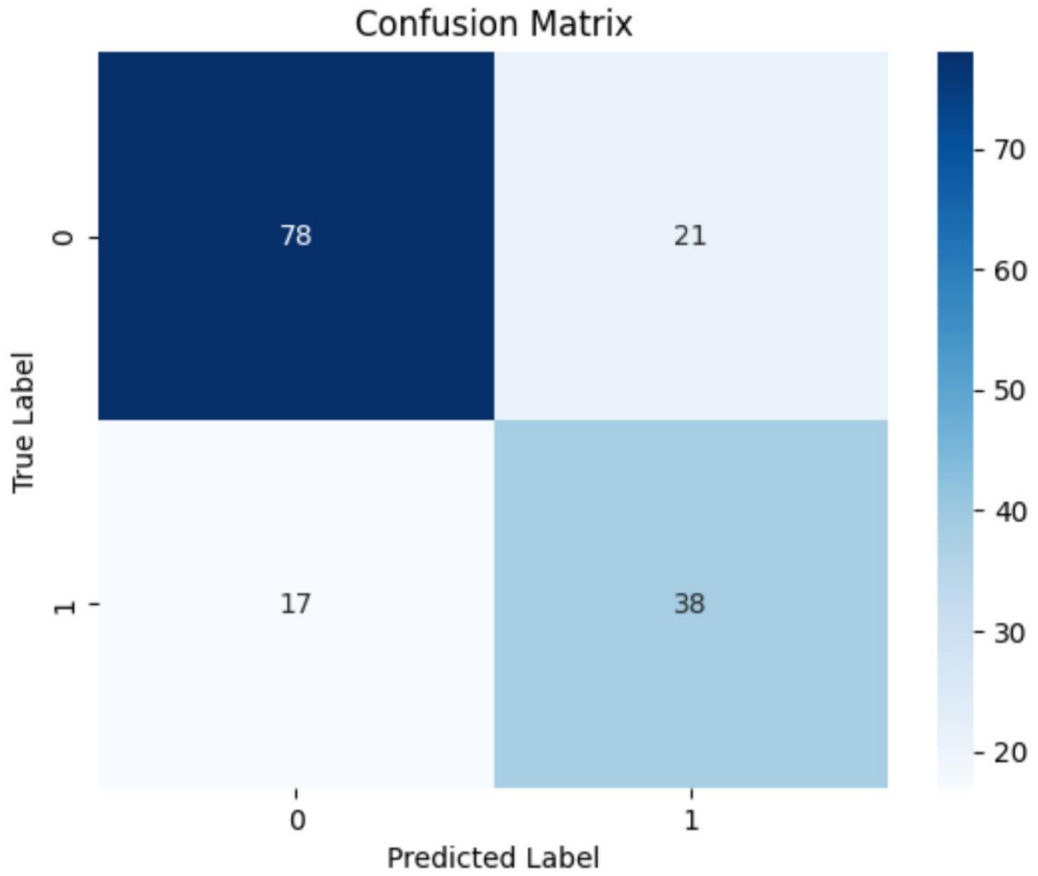
The model was trained on the training data (X train, y train) and evaluated on the test set (X test, y test).

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

The goal of this project is to analyze the Brazilian E-Commerce dataset using association rule learning techniques. Association rule mining, such as the Apriori algorithm, is a powerful tool that helps uncover frequent patterns, correlations, or associations between products that customers tend to purchase together. By identifying these relationships, businesses can gain valuable insights into consumer behavior. These insights can then be applied to enhance various business strategies, such as cross-selling, product bundling, personalized marketing campaigns, and inventory management, ultimately improving customer satisfaction and increasing sales.

2. Data Preprocessing

2.1 Dataset Overview

The analysis used three key datasets from the Brazilian E-Commerce Dataset:

•olist order items dataset.csv: Contains information on individual products included in each order.

•olist orders dataset.csv: Contains details about orders, such as order id, customer id, order status, and order purchase timestamp.

•olist products dataset.csv: Contains product details such as product id, product category name, and dimensions of products.

2.2 Missing Values

Missing values were checked in all datasets:
Missing values in Orders dataset: 0
Missing values in Order Items dataset: 0
Missing values in Products dataset: 0
There were no significant missing values

2.3 Data Merging

The datasets have been merged to create a unified and comprehensive dataset for deeper analysis. First, the orders_df dataset, which contains order information, was merged with the order_items_df dataset to link each order with the specific products it contains:

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

Next, this data was merged with the products_df dataset, which includes specific details about the products, such as product categories and dimensions. This allowed us to enrich the data with additional attributes like product names and categories:

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

The resulting merged dataset now provides a comprehensive view of each transaction, including details about the purchased products and their associated categories. This unified dataset forms the basis for subsequent analyses, enabling us to explore purchasing patterns, product associations, and extract potential marketing insights.

2.4 Filtering Data

To focus on the most frequently purchased products and orders, we filtered the dataset .to include only the top 1,000 products and the top 2,000 orders

First, we identified the most common products by counting the occurrences of each product ID. We selected the top 1,000 products based on these counts. Similarly, we counted the number of occurrences for each order ID and selected the top 2,000 .orders

Using these selections, we created a filtered dataset that includes only the transactions associated with the identified top products and orders. This filtering process helps concentrate our analysis on the most relevant data, allowing us to gain insights into .popular products and high-volume orders

2.5 Visualizing the Top 10 Products

The bar chart displays the top 10 most ordered products from the order data. The horizontal axis represents the shortened product IDs, while the vertical axis shows the number of times each product was ordered.

Key Observations:

Most Ordered Product: The product with ID 314663af is the most ordered, with approximately 500 orders.

Least Ordered Product: The product with ID c1e95ad7 was ordered around 300 times.

Variation in Orders: Orders range from 300 to 500, indicating clear differences in product popularity.

This data highlights the significance of certain products compared to others, which can aid in making marketing and strategic decisions.

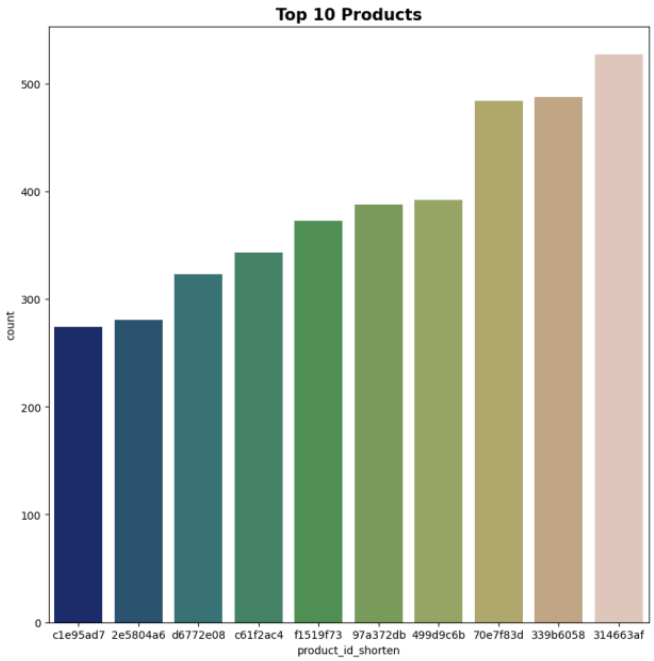


Figure 1: Top 10 Most Ordered Products

2.6 Transaction Matrix Creation

A binary transaction matrix was created where the rows represent orders and the columns represent products. Each cell in the matrix indicates whether a specific .product was purchased in a particular order

."If the product was purchased in that order, the value is set to "1
."If the product was not purchased, the value is set to "0
This matrix allows for a clear representation of product purchases across different orders, facilitating further analysis of purchasing patterns and relationships between products.

3. Association Rule Mining

3.1 Apriori Algorithm

The Apriori algorithm was applied to generate frequent itemsets from the basket data. A minimum support threshold of 0.5% was used, meaning that items appearing in more than 0.5% of all baskets would be considered part of the frequent itemsets.

The Apriori algorithm is a widely used data mining technique aimed at identifying relationships between items in large datasets. By identifying frequently occurring itemsets, this information can be utilized in applications such as product recommendations, customer behavior analysis, and improving marketing strategies.

By using the parameter use_colnames=True, it ensures that the column names are included in the results, making it easier to identify the frequent items in the resulting datasets**.**

3.2 Generating Association Rules

The frequent itemsets were used to generate association rules using the lift metric:
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)

3.3 Filtering Rules

Rules were filtered based on confidence and lift:
filtered_rules = rules[(rules['confidence'] > 0.6) & (rules['lift'] > 1)]

3.4 Visualizing Association Rules

A scatter plot was used to visualize support vs confidence with lift as the bubble size:

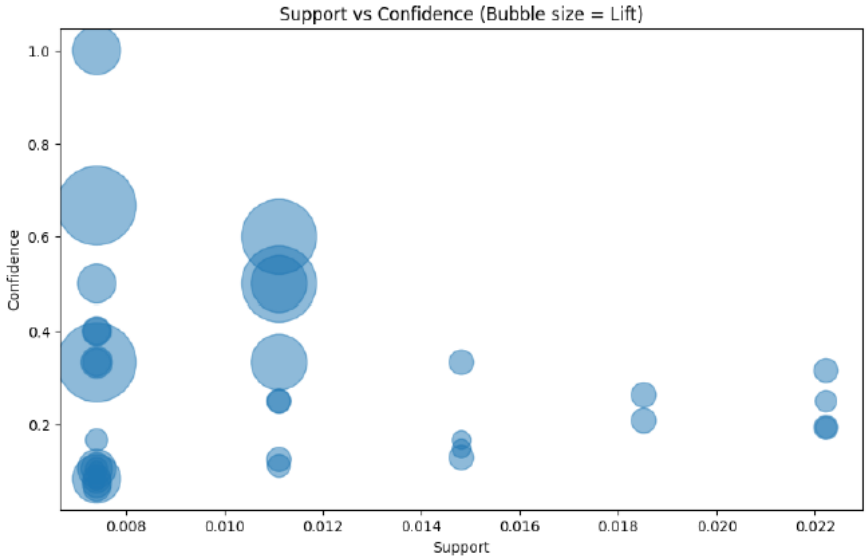


Figure 2: Support vs Confidence (Bubble size = Lift)

4. Model Performance

The model's performance is based on the following key metrics:

• Support: Proportion of orders that contain the itemset.
• Confidence: Likelihood that the consequent will be purchased if the antecedent is purchased.

• Lift: Measures how much more likely the consequent is to be purchased given the antecedent, compared to random chance.

The top 10 rules were selected based on confidence:
rules = rules.sort_values(by='confidence', ascending=False)
rules.head(10)

4.1 Example Rule Interpretation

- Antecedent: [’product A’] - Consequent: [’product B’] - Confidence: 0.75

- Lift: 1.5

This means that 75% of the time, when product A is purchased, product B is also purchased, and the likelihood of buying product B increases by 1.5 times compared to random chance.

5. Insights Gained

5.1 Top Products and Categories

The top 10 most purchased products and categories were identified. This helps in understanding customer preferences and highlights the most popular products.

5.2 Strong Product Associations

The Apriori algorithm revealed strong relationships between products, which can be utilized for:

Cross-selling Opportunities: Suggesting related products.

Product Bundling: Grouping products together to increase sales.

5.3 Specific Product Rules

Rules associated with specific products were identified, enabling targeted marketing and offering recommendations for products related to those specific items.

6. Conclusion

By applying the Apriori algorithm for association rule learning, significant insights into customer purchasing habits were uncovered. The extracted rules present opportunities to enhance cross-selling strategies, allowing for the suggestion of complementary or related products when a specific item is purchased. Additionally, these rules can be leveraged to bundle products in promotions and deals, increasing the likelihood of joint purchases and boosting revenue.

Furthermore, the analysis helped identify the most frequently purchased and preferred products by customers, revealing strong associations between certain items. This information is highly valuable for better targeting marketing campaigns, as it enables the promotion of relevant products to customers or the creation of personalized recommendations that enhance the shopping experience and drive sales.

7. Future Work

In the future, several directions can be explored to expand the analysis and achieve deeper insights:

Analyzing Product Associations Over Time: It would be beneficial to study product associations across specific time periods, such as different seasons or promotional periods. This could help understand how purchasing patterns change throughout the year and identify products that see increased sales during certain times. Such insights could enhance seasonal marketing strategies and improve the effectiveness of promotional offers.

Utilizing Hierarchical Product Categories: Analyzing associations between products at more detailed levels using a hierarchical structure of product categories can provide deeper insights. This approach helps uncover relationships not only between individual products but also among related product groups. It enhances the understanding of customer preferences at the category level and can aid in developing integrated sales strategies, promoting groups of products together based on their shared classifications.
By employing these advanced analyses, personalized recommendations can be offered, further enhancing the customer experience and increasing growth and profitability opportunities.

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

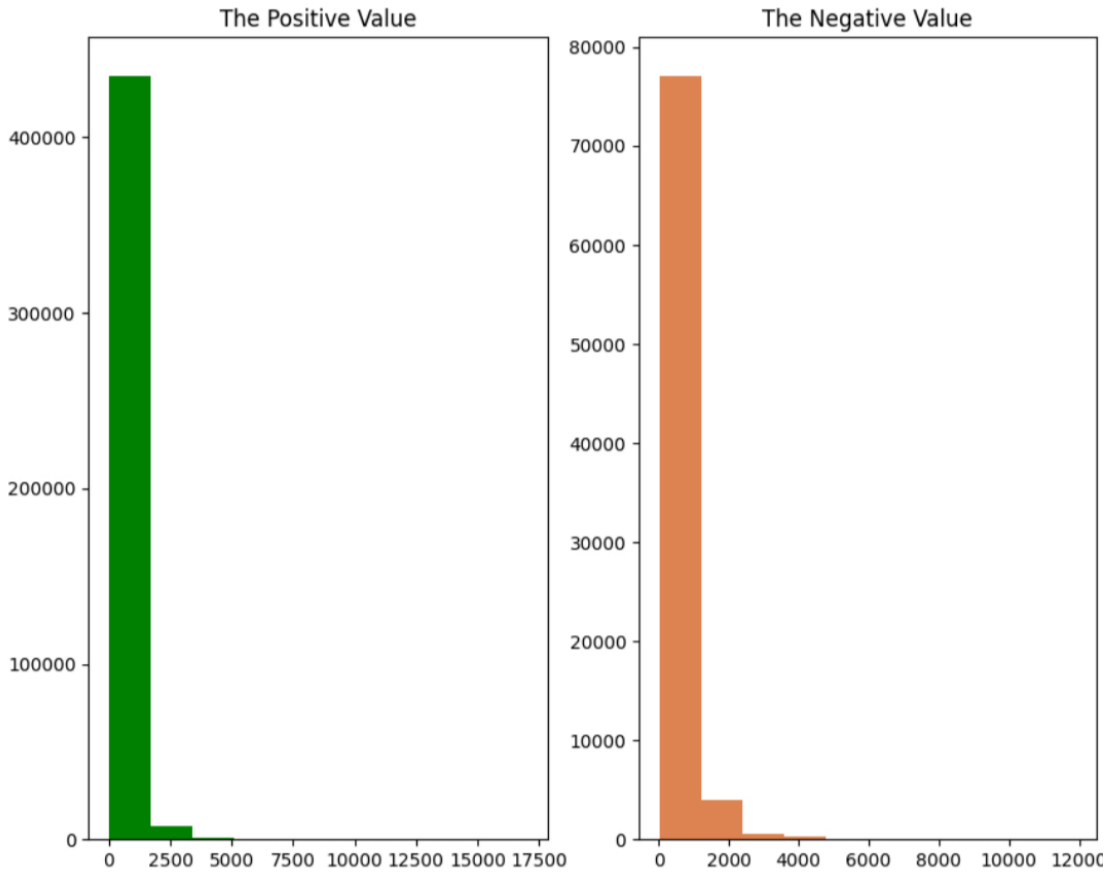
The dataset used for this analysis is the Amazon Fine Food Reviews dataset. The dataset contains reviews, ratings (scores), product information, and user data. Each review includes a Score ranging from 1 to 5, which we used to label reviews as positive or negative. Scores 1 and 2 were considered negative,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

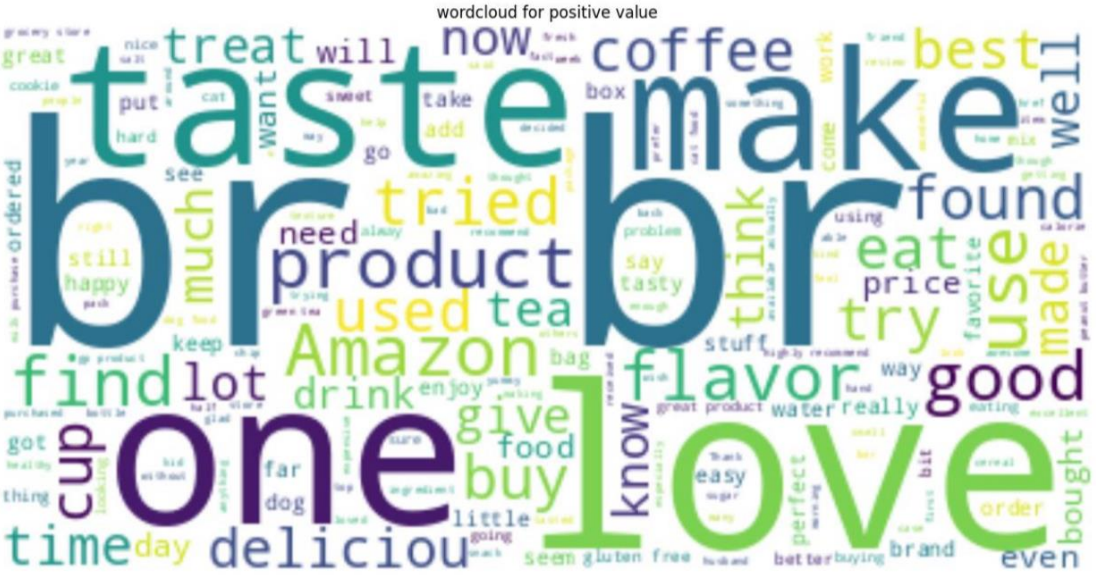
5. Model Training and Evaluation

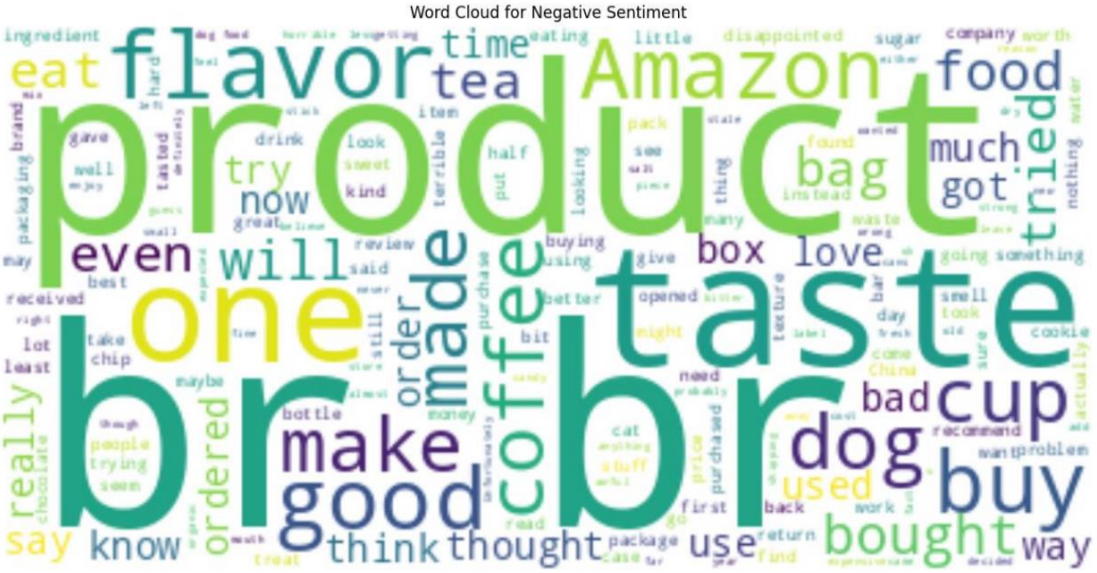
Two machine learning models were trained and evaluated to classify the sentiment:

• 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.





• 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.

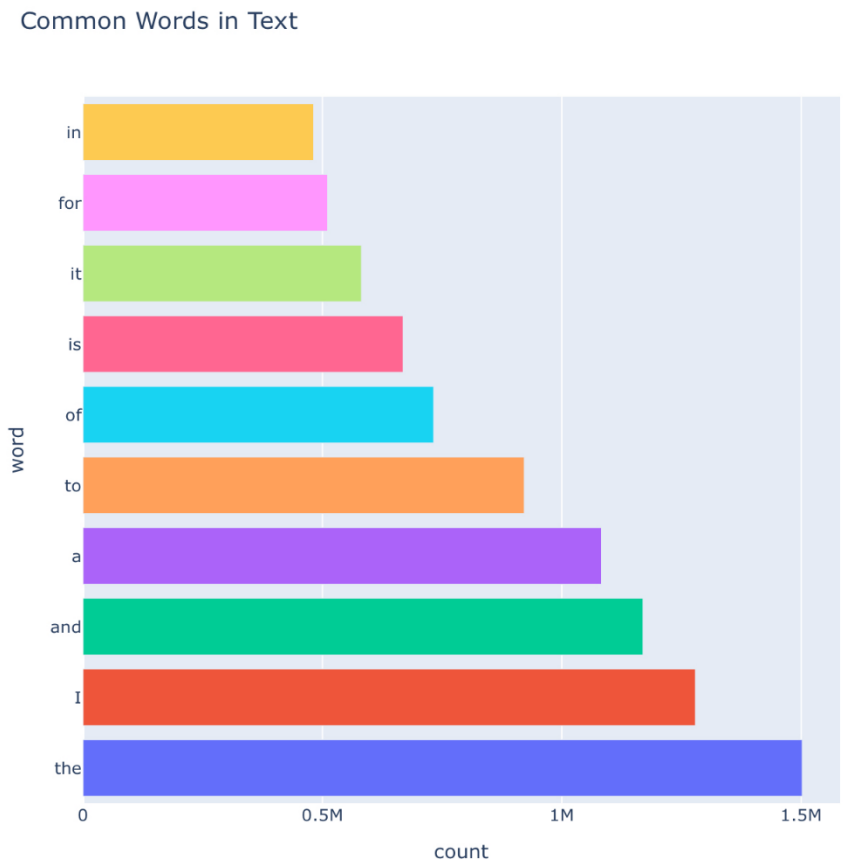
6.visualizing

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.