**File Overview**

**File Name: MultipleFiles/1 (1).ipynb**

Purpose: The notebook is designed for data analysis and sentiment prediction based on product reviews, ratings, and other attributes.

Table of Contents

Problem Definition

Data

Features

Preparing the Tools

Load Data

Data Exploration (EDA)

Data Visualization

Data Frequency Analysis

Data Cleaning and Preprocessing

Modeling and Predictions

Conclusion

1. Problem Definition

Statement: Given the textual reviews, ratings, and other product-related attributes, can we predict the sentiment labels of reviews as positive, negative, or neutral?

2. Data

Description: The dataset contains product reviews and related metadata, including product prices, ratings, and textual review content. Each review is classified into a sentiment label: positive, negative, or neutral.

3. Features

Data Dictionary:

product\_name: Name of the product.

product\_price: Price of the product.

Rate: Numerical rating (1 to 5).

Review: Short summary of the user’s opinion.

text: Detailed textual review.

labels: Sentiment classification (positive, negative, neutral).

sentiment\_code: Numerical encoding of the labels (2: Positive, 1: Neutral, 0: Negative).

4. Preparing the Tools

Libraries Used:

pandas: For data manipulation and analysis.

numpy: For numerical operations.

matplotlib: For data visualization.

seaborn: For statistical data visualization.

sklearn: For machine learning models and evaluation.

5. Load Data

Data Loading Code:

python

Verify

Open In Editor

Run

Copy code

df = pd.read\_csv("modified\_cleaned\_datathon.csv")

6. Data Exploration (EDA)

Goals:

Understand the dataset.

Identify missing values and outliers.

Determine how to treat different data types.

7. Data Visualization

Visualizations:

Bar plots for sentiment distribution.

Histograms for product price distribution.

Correlation heatmaps.

8. Data Frequency Analysis

Sentiment Code Distribution:

Count of each sentiment label in the dataset.

9. Data Cleaning and Preprocessing

Data Cleaning Steps:

Dropping unnecessary columns.

Handling missing values.

Encoding categorical variables.

10. Modeling and Predictions

Machine Learning Models Used:

Logistic Regression

K-Nearest Neighbors

Random Forest Classifier

Model Evaluation Metrics:

Confusion Matrix

Classification Report

Precision, Recall, F1 Score

11. Conclusion

Summary of Findings: The notebook concludes with insights gained from the data analysis and the performance of the models used for sentiment prediction.

Additional Notes

Error Handling: The notebook includes error handling for data type conversions and missing values.

Visualization Enhancements: Various visualizations are included to better understand the data distributions and relationships.

This documentation provides a structured overview of the Jupyter Notebook, detailing its purpose, contents, and methodologies used throughout the analysis.

The provided text is a structured overview of a Jupyter Notebook designed for data analysis and sentiment prediction based on product reviews. The notebook follows a standard data science workflow:  
The provided Jupyter Notebook is designed to predict the sentiment (positive, negative, or neutral) of product reviews using data analysis and machine learning techniques.

**1. Problem Definition**

The goal is to classify sentiments based on textual reviews, ratings, and other product data.

**2. Data**

The dataset contains product reviews along with metadata like prices, ratings, and review text. Each review is labeled with a sentiment.

**3. Features**

The data dictionary includes:

* Product Name
* Product Price
* Rating
* Review Summary
* Detailed Textual Review
* Sentiment Classification (Positive, Negative, Neutral)
* Numerical Sentiment Code

**4. Tools**

Libraries used:

* pandas
* numpy
* matplotlib
* seaborn
* sklearn

**5. Data Loading**

Code to load the data is provided.

**6. Data Exploration (EDA)**

This stage involves:

* Understanding the dataset
* Identifying missing values and outliers
* Determining how to handle different data types

**7. Data Visualization**

Includes:

* Bar plots for sentiment distribution
* Histograms for product price distribution
* Correlation heatmaps

**8. Data Frequency Analysis**

Analyzes the count of each sentiment label in the dataset.

**9. Data Cleaning and Preprocessing**

Includes:

* Dropping unnecessary columns
* Handling missing values
* Encoding categorical variables

**10. Modeling and Predictions**

Uses machine learning models:

* Logistic Regression
* K-Nearest Neighbors
* Random Forest Classifier

Model evaluation metrics:

* Confusion Matrix
* Classification Report
* Precision
* Recall
* F1 Score

**11. Conclusion**

Summarizes:

* Findings from data analysis
* Model performance for sentiment prediction

**12. Additional Features**

* Error handling for data type conversions and missing values

Various visualizations to understand data distributions and relationships

**Summary**

* **Problem Definition**: Predict sentiment labels (positive, negative, neutral) from product reviews and metadata.
* **Data**: Product reviews, ratings, and other attributes. Features include product name, price, rating, review summary, detailed text review, sentiment label, and sentiment code.
* **Tools**: Libraries like pandas, numpy, matplotlib, seaborn, and sklearn.
* **Data Loading**: Loading data from a CSV file.
* **Data Exploration (EDA)**: Understanding the dataset, identifying missing values and outliers, and determining data treatment.
* **Data Visualization**: Bar plots, histograms, and correlation heatmaps to visualize data.
* **Data Frequency Analysis**: Analyzing the frequency of sentiment codes.
* **Data Cleaning and Preprocessing**: Dropping unnecessary columns, handling missing values, and encoding categorical variables.
* **Modeling and Predictions**: Using machine learning models like Logistic Regression, K-Nearest Neighbors, and Random Forest Classifier. Evaluation using Confusion Matrix, Classification Report, Precision, Recall, and F1 Score.
* **Conclusion**: Summarizing findings and model performance.

The notebook also includes error handling and visualization enhancements. The documentation provides a clear outline of the notebook's purpose, content, and methodologies.