AML PROJECT DELIVERABLE - 1

Group 23 Members: Shashwat Kumar(sk5520), Saatvik Saradhi Inampudi(si2464), Tanner Hillison(tkh2119), Mengxi Liu(mi5189), Gaifan Zhang(gz2360)

1. Problem statement: The aim of this project is to analyze customer reviews from Amazon's Fine Food category and predict the sentiment of a given review (Negative/Neutral/Positive) using various machine learning (ML) techniques. Sentiment analysis on such large-scale data can help businesses better understand customer preferences, improve product recommendations, and adjust marketing strategies. This problem falls under supervised learning with multi-class text classification as the core task where we determine the performance of our model by analyzing the accuracy as well as the confusion matrix.

2. Identification and description of the dataset

Dataset link: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

Description: The dataset consists of **568,454 reviews** from Amazon's Fine Food category, collected over a period of 10 years. It contains detailed review information, including product and user metadata, which provides a rich context for sentiment analysis. Each review includes the following fields:

- **Id**: Row identifier for each review entry.
- **ProductId**: Unique identifier for the product being reviewed.
- **UserId**: Unique identifier for the user who submitted the review.
- **ProfileName**: Profile name of the user submitting the review.
- **HelpfulnessNumerator**: Number of users who found the review helpful.
- **HelpfulnessDenominator**: Total number of users who found the review helpful or not.
- **Score**: Rating assigned to the product (an integer between 1 and 5).
- **Time**: Unix timestamp representing when the review was submitted.
- **Summary**: Summary provided by the reviewer.
- **Text**: The main content of the review describing what the customer felt about the product.

3. Proposed ML techniques

At the start of the project, exploratory data analysis (EDA) and data preprocessing will be performed to clean and prepare the dataset for model training. Following this, machine learning techniques such as Logistic Regression, Naive Bayes with Bag of Words (BoW) and TF-IDF, Random Forest, and Support Vector Classifier (SVC) will be applied on the dataset. Additionally, advanced tools like Word2Vec, GloVe, and XGBoost will be explored. The performance will be evaluated using metrics like accuracy, confusion matrix, precision, recall, F1-score, and ROC-AUC.