MCA 5141 – Machine Learning Lab
Week – 9

EXER 1

Download the "Womens Clothing E-Commerce Reviews.zip" file and answer the following:

1. Preprocessing:

a. Find any null values are present or not, If present remove those data.
b. Remove the data that have less than 5 reviews.
c. Clean the data and remove the special characters and replace the contractions with its expansion. Convert the uppercase character to lower case. Also, remove the punctuations.

2. Separate the columns into dependent and independent variables (or features and labels). Then you split those variables into train and test sets (80:20).
3. Apply the Naïve Bayes Classification Algorithm on Sentiment category to predict if item is recommended
4. Tabulate accuracy in terms of precision, recall and F1 score.

EXER 2

1. Data Preprocessing and Feature Engineering
- Load the dataset and explore its structure.
- Identify and handle missing values appropriately.
- Perform feature selection by calculating correlation coefficients and removing highly correlated features.
- Convert continuous variables into categorical bins where appropriate (e.g., discretizing age-based rates).
- Apply dimensionality reduction techniques such as PCA to optimize feature space.
- Create a binary target variable based on whether the 'Total.Rate' is above or below the third quartile, making classification more challenging.

2. Split the dataset into training and testing sets with an 80-20 ratio.

3. Implementing Naïve Bayes
- Select the following features for classification:
  - Rates.Age.< 18
  - Rates.Age.18-45
  - Rates.Age.45-64
  - Rates.Age.> 64
  - Types.Lung.Race.White
  - Types.Lung.Race.Black
  - Types.Lung.Race.Hispanic
- Train multiple Naïve Bayes models (GaussianNB, MultinomialNB, and BernoulliNB) using only the selected features.
- Compare the models based on precision, recall, F1-score, and AUC-ROC curve.
- Analyze the assumptions of each Naïve Bayes variant and determine which one fits the dataset best.