YASHIKA KHURANA

TASK 3: NLP

AIM:

The primary objective of the task is to predict the product category for a given directory of products using their description.

APPROACH:

Firstly, I conducted an Exploratory Data Analysis to find out huseful information. The directory of products consisted of a category tree which listed many sub categories for the products. Since, the question mentioned that primary category was to be used, primary category was acquired from the tree by slicing to obtain its root. The category tree column was dropped and primary category column was inserted. Next,

unique category counts were found to know the number of records corresponding to each category. The data was visualised as represented in the Jupyter notebook. The analysis revealed that certain categories had very less records against them. These categories would not have adequate training set available. Hence, they were dropped by setting a threshold value of 100 records. Also, the dataset comprises of product rating and overall rating columns, which were removed after analysing that only 1797 records had ratings. Moreover, this data was not important to predict the category. So, it was discarded after visualising. A pie chart was made to understand the products of FK_advantage, which depicted that only 3.9 % are of advantage. EDA also showed that 5153 products were unbranded.

After EDA, categories were predicted. For this purpose, the descriptions were pre-processed.

Pre-processing steps included:

- Converting to lower-case
- Tokenizing
- If token was found in ['key', 'features','specifications', 'rs', 'flipkart', 'com'] , it was not considered as these words are found in a lot of decriptions but they are not deciphering any information.

- Removing stop-words
- Stemming
- Removing digits

The descriptions were then converted into vectors using CountVectorizer.

Then, the data was split into training and test set in the ratio of 80:20. The training data was trained using Multinomial Naïve Bayes, Bernoulli Naïve Bayes and Gaussian Naïve Bayes.

Then their respective scores and test accuracies were reported as below:

METHOD	**TEST ACCURACY**	**SCORE**
Multinomial Naïve Bayes	95.36	0.96
Binomial Naïve Bayes	75.01	0.76
Gaussian Naïve Bayes	89.8	0.97