Assignment 2 Report
Decision Tree and Naïve Bayes
Data Mining

CSE 5334 Section 001

Student Name:
Ahir Kunjal (1001877263)
Sheth Devanshi (1001959019)
Gope Pankaj (1001990547)

Professor: Dr Elizabeth D Diaz

Decision Trees :-

The decision tree approach develops the tree based on a distribution of accessible characteristics. To determine the mixture contained in each feature that may be partitioned, several approaches such as Gini and entropy can be utilized. These metrics provide information about the quality of the distribution obtained when specific data attributes are used to segment records and forecast class values. Controlling the depth at which trees develop allows you to manage growth and anticipate values fast. Accepts a property value and divides the record appropriately. It then takes the next value from the distribution and repeats this process until all elements have been estimated.

.

Naïve Bayes :-

Naïve Bayes employs the well-known Bayes rule approach for probability. It starts with the premise that the characteristics are independent of one another, then calculates the probability of each attribute value for each class value and compares the two for a given data point to forecast that class. One method to use these probabilities is to apply Bayes' theorem in Probability and Statistics to get the dependent probabilities and explain the predictors.

Data Set :-

The dataset uses people's Twitter data. There are several columns such as number of retweets by person, username, number of tweets, etc.

Pre-Processing :-

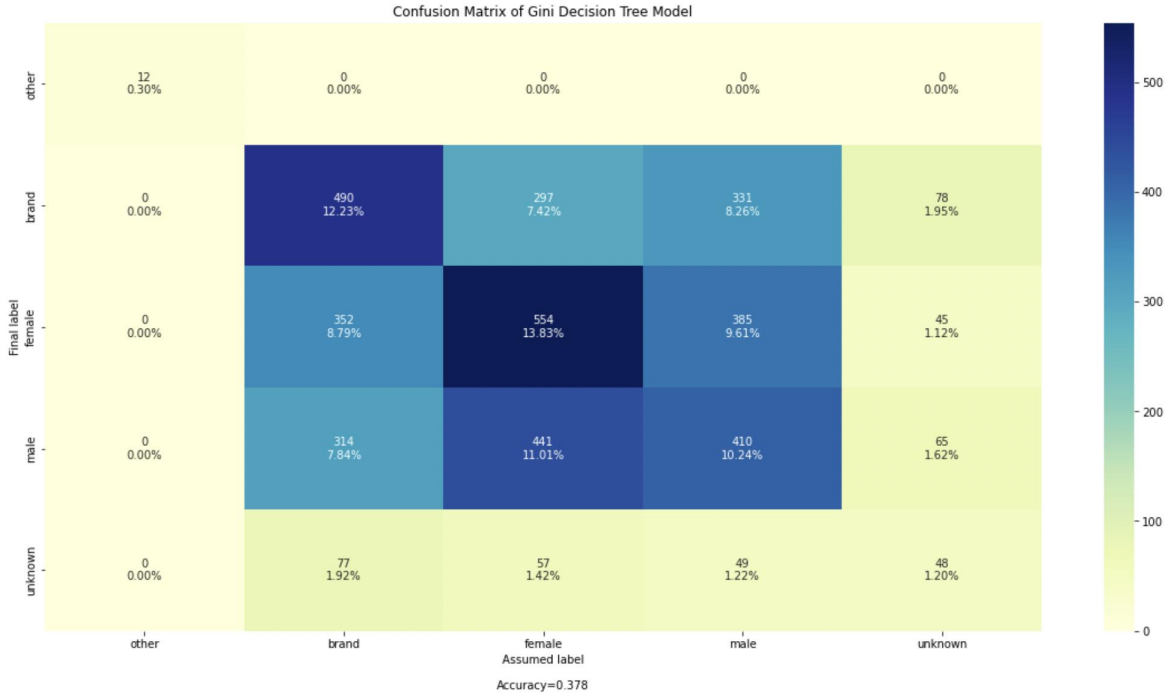
```
#Preprocessing of the data
#Here we have changed to the numbers from text feature which has categories
classifying_gender_read['_unit_state'] = classifying_gender_read['_unit_state'].map({'finalized':0,'golden':1})
classifying_gender_read['profile_yn'] = classifying_gender_read['profile_yn'].map({'yes':0,'no':1})
#Adding the columns and their rows that are not null and not text feature into a new dataset
new_dataframe = classifying_gender_read[['_golden','_unit_state','profile_yn','tweet_count',
                                         '_trusted_judgments','retweet_count','gender:confidence']][~(classifying_gende
new_dataframe
#refs:
#https://www.machinelearningplus.com/pandas/pandas-dropna-how-to-drop-missing-values/
```

First and foremost, we determined the integer and category data types accessible in the columns. Because decision trees are incapable of processing text, we turned it to numbers using dictionaries. Following that, we removed the rows with null values in any of the columns, chose the integer and converted category characteristics, and put them in a separate data frame.

We preserved just the traits that we thought may be beneficial in determining a person's gender.

- **Results for the Gini Model :-**

```
The accuracy is:  0.3780274656679151

The classification report is:
              precision    recall  f1-score   support

       Other       1.00      1.00      1.00        12
       brand       0.40      0.41      0.40      1196
      female       0.41      0.41      0.41      1336
        male       0.35      0.33      0.34      1230
     unknown       0.20      0.21      0.21       231

    accuracy                           0.38      4005
   macro avg       0.47      0.47      0.47      4005
weighted avg       0.38      0.38      0.38      4005
```



The categorization ratio has an accuracy of roughly 37%. The confusion matrix further emphasizes this fact. Each diagonal element adds to the precision, and you can see that there are values farther away from these spots.

The other gender groups were accurately predicted, as can be seen.

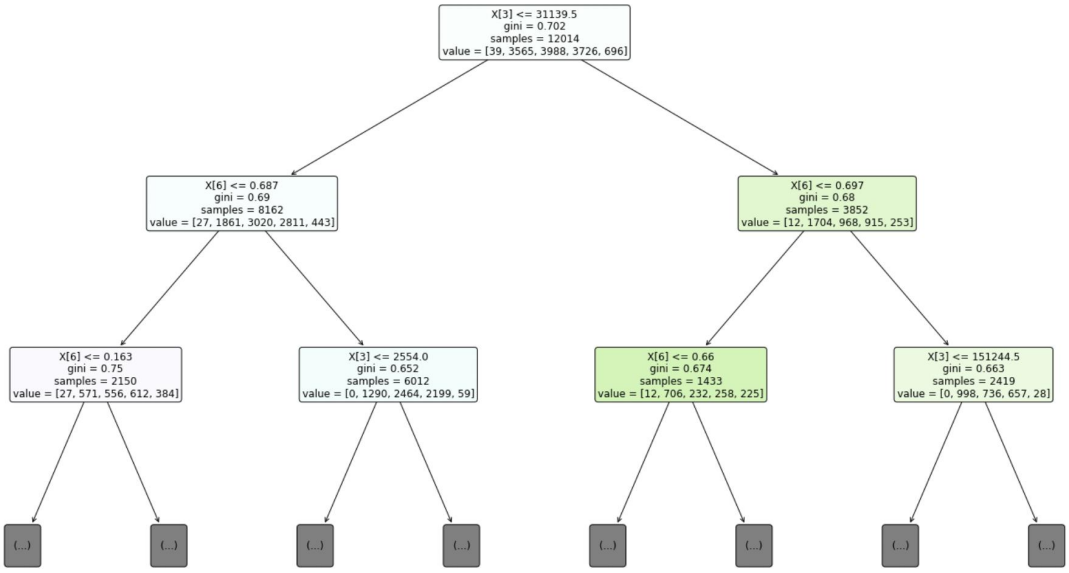
- **Results for the Entropy Model :-**

```
The accuracy is:  0.38352059925093634

The classification report is:
              precision    recall  f1-score   support

       Other       1.00      1.00      1.00        12
       brand       0.40      0.41      0.40      1196
      female       0.41      0.41      0.41      1336
        male       0.35      0.33      0.34      1230
     unknown       0.20      0.21      0.21       231

    accuracy                           0.38      4005
   macro avg       0.47      0.47      0.47      4005
weighted avg       0.38      0.38      0.38      4005
```

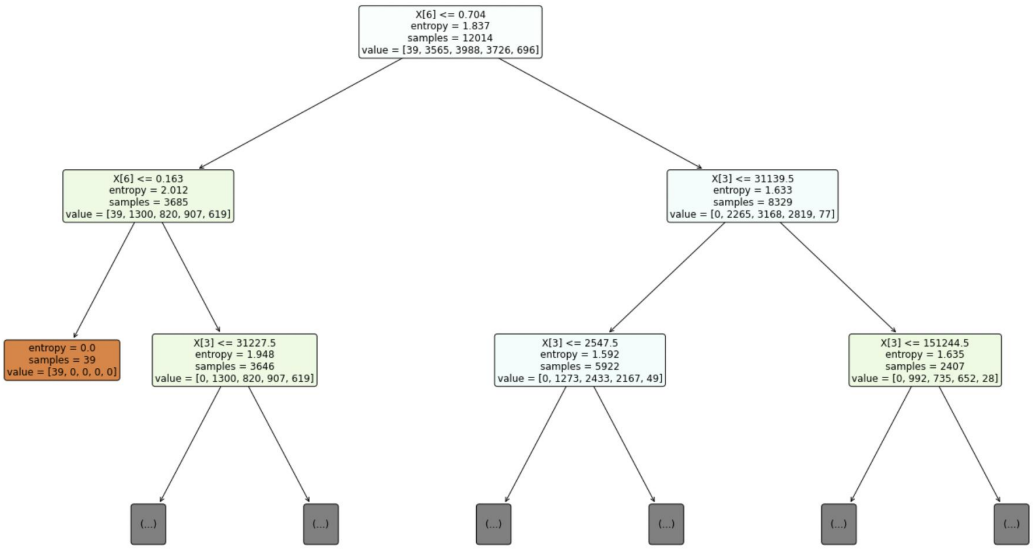
The confusion matrix and classification report show that the entropy model has an accuracy value of roughly 38.4 percent. The values of the diagonal boxes for the entropy model are greater when compared to the Gini model. As a result, the entropy model is superior for this sort of pre-processing and data collection.

- **Decision Trees for Gini and Entropy Model Depth :-**

Gini :-

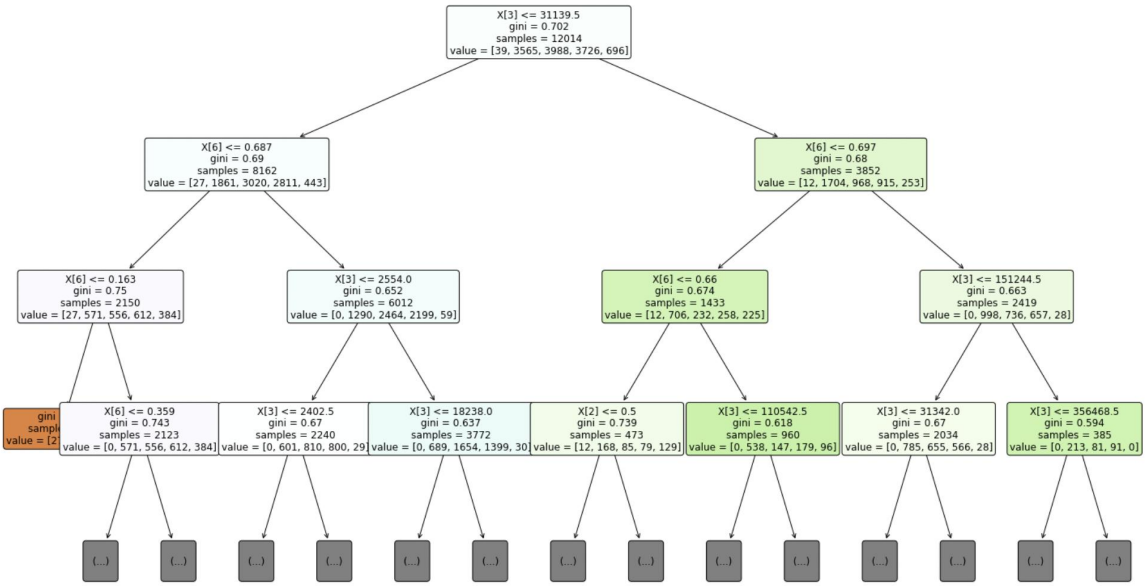


Entropy :-

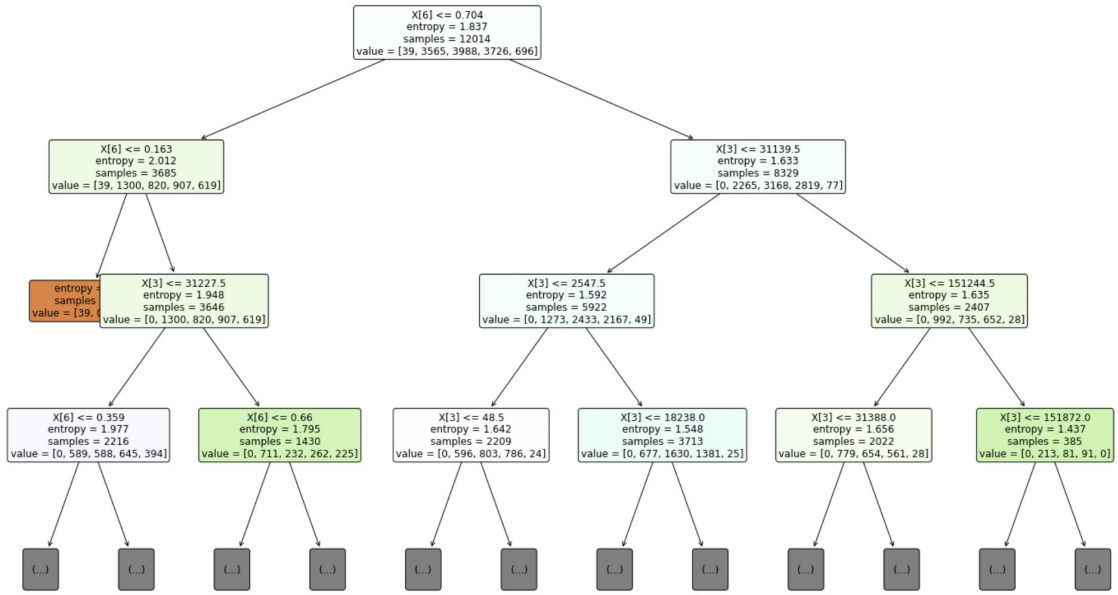


- **Decision Trees for Gini and Entropy Model Depth :-**

Gini :-

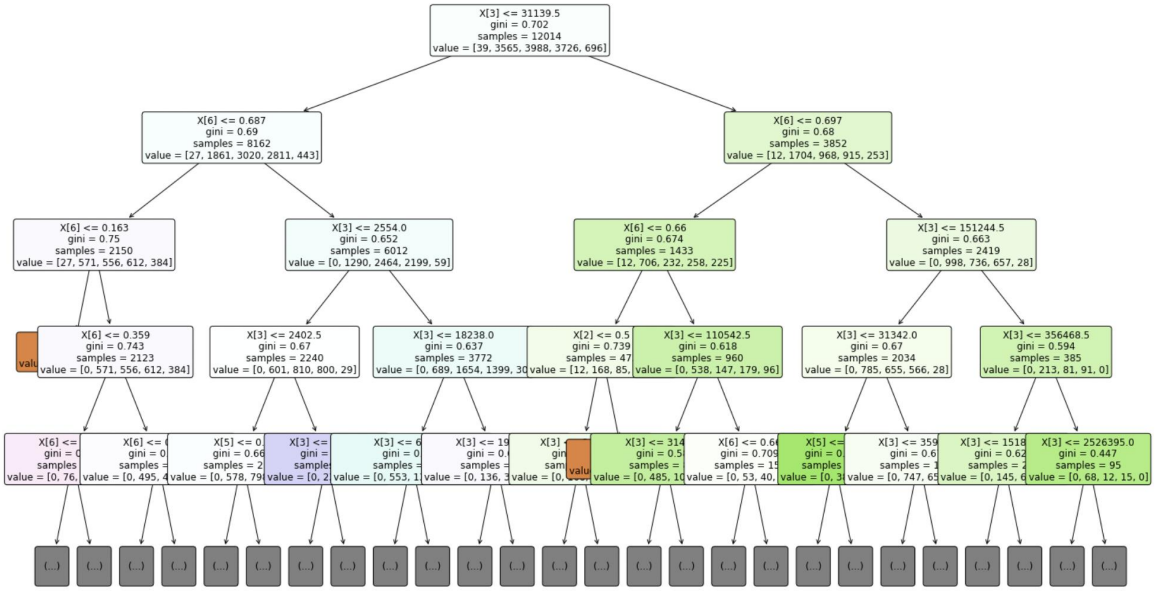


Entropy :-

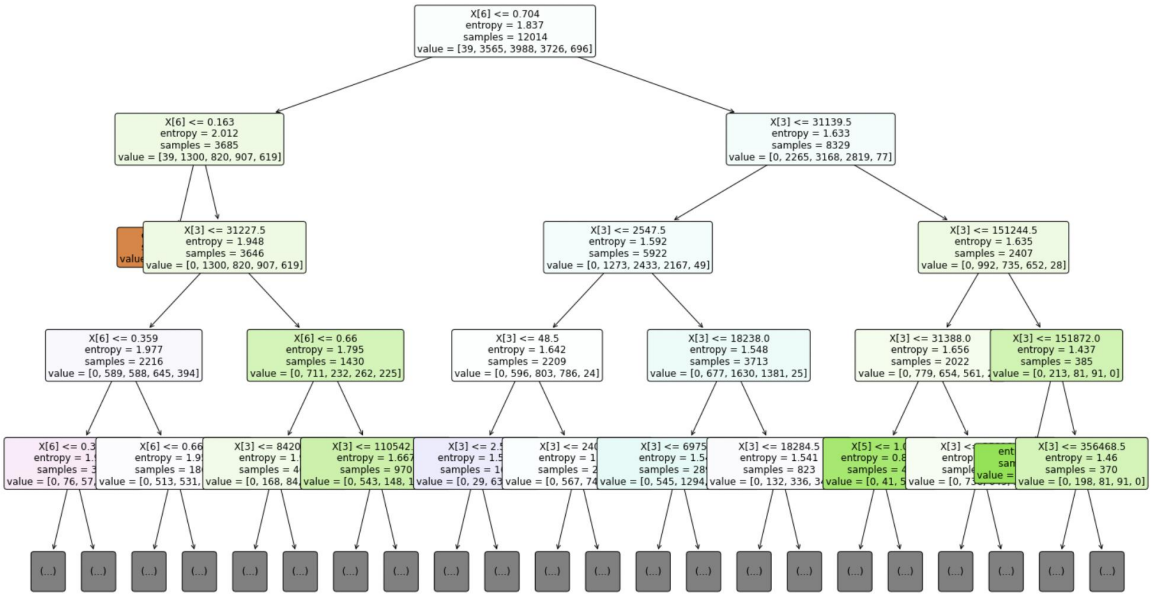


- **Decision Trees for Gini and Entropy Model Depth :-**

Gini :-

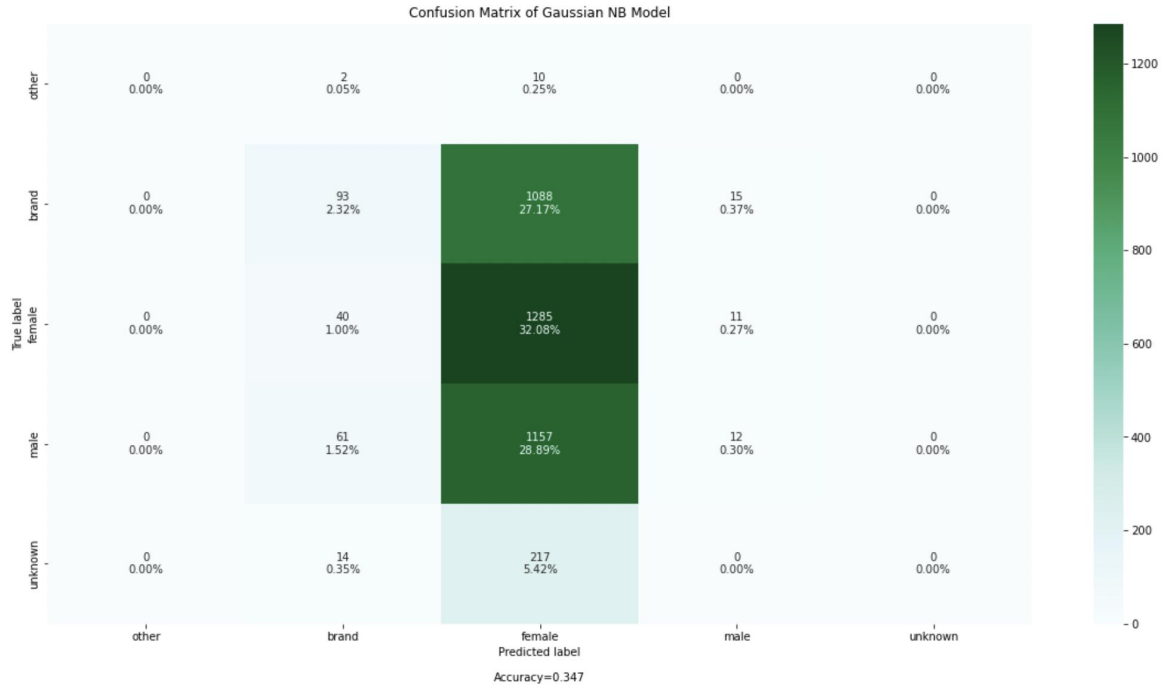


Entropy :-



- **Results for Gaussian Naïve Bayes Model :-**

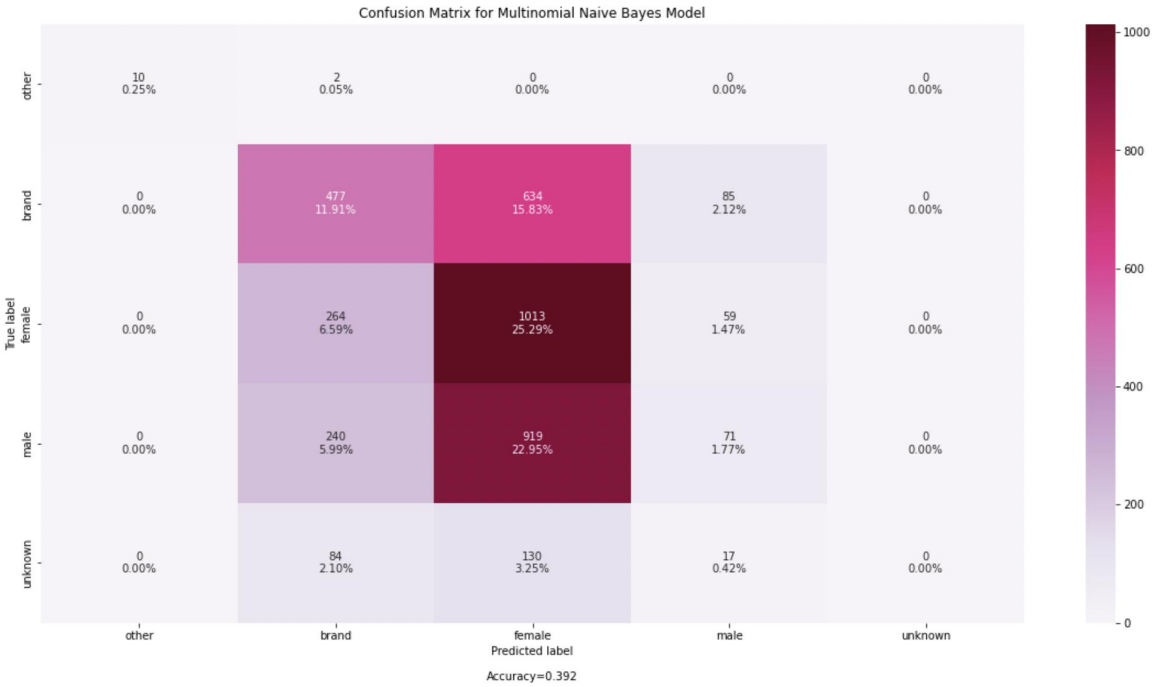
```
The accuracy of the model is:  0.3470661672908864

The classification report is:
                precision    recall  f1-score   support

         Other       0.00      0.00      0.00        12
         brand       0.44      0.08      0.13      1196
        female       0.34      0.96      0.50      1336
          male       0.32      0.01      0.02      1230
       unknown       0.00      0.00      0.00       231

      accuracy                           0.35      4005
     macro avg       0.22      0.21      0.13      4005
  weighted avg       0.34      0.35      0.21      4005
```



The accuracy is **34.7%**

- **Results for the Multinomial Naïve Bayes Model :-**

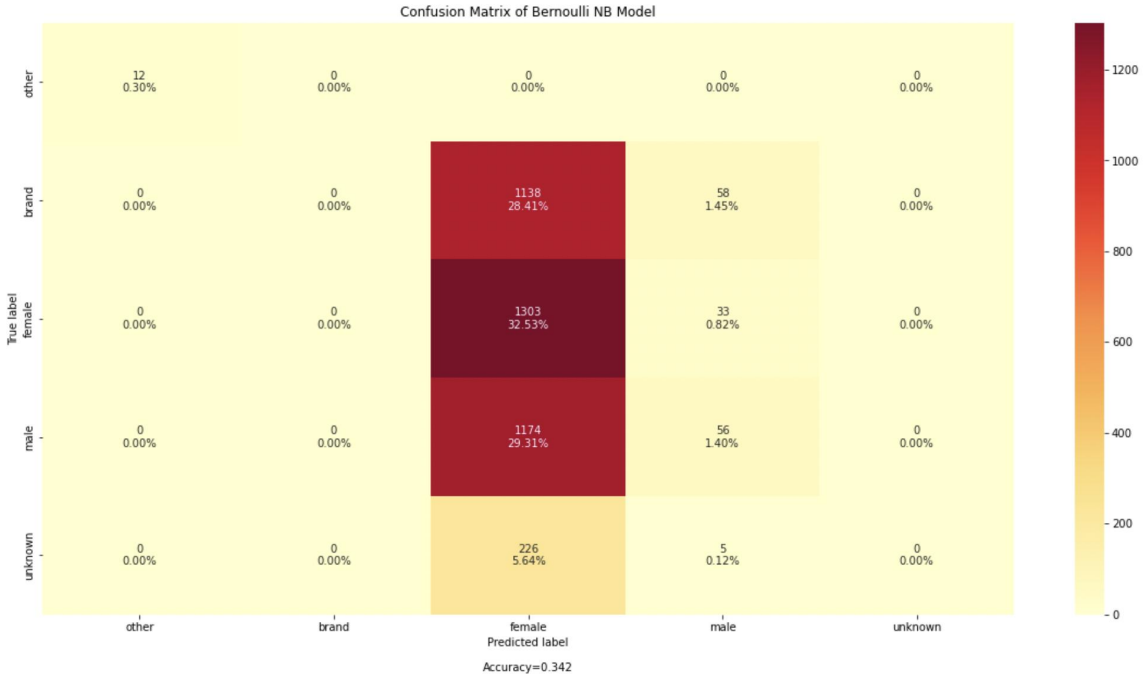
```
The accuracy of the model is:  0.3922596754057428

The classification report is as follows:
              precision    recall  f1-score   support

       Other       1.00      0.83      0.91        12
       brand       0.45      0.40      0.42      1196
      female       0.38      0.76      0.50      1336
        male       0.31      0.06      0.10      1230
     unknown       0.00      0.00      0.00       231

    accuracy                           0.39      4005
   macro avg       0.43      0.41      0.39      4005
weighted avg       0.36      0.39      0.33      4005
```



The accuracy is **39.2%**

- **Results for the Bernoulli Naïve Bayes Model :-**

```
The accuracy of the model is:  0.3423220973782772

The classification report is:
              precision    recall  f1-score   support

       Other       1.00      1.00      1.00        12
       brand       0.00      0.00      0.00      1196
      female       0.34      0.98      0.50      1336
        male       0.37      0.05      0.08      1230
     unknown       0.00      0.00      0.00       231

    accuracy                           0.34      4005
   macro avg       0.34      0.40      0.32      4005
weighted avg       0.23      0.34      0.20      4005
```



The accuracy is **34.2%**

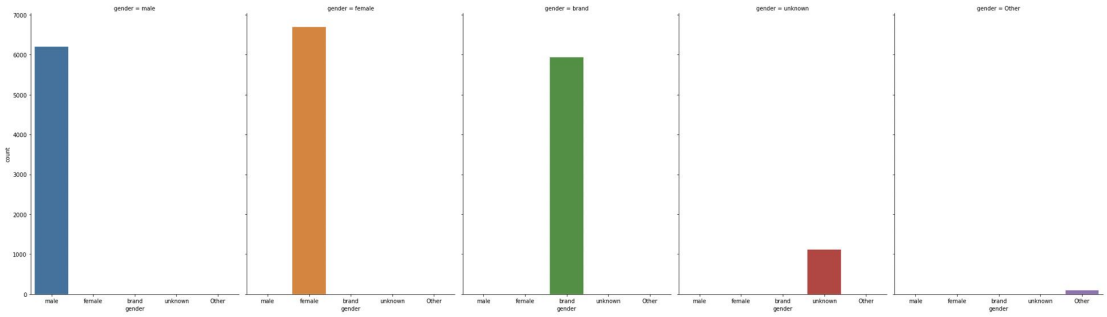
We did not use the Complement Naive Bayes model because the classes are unbalanced and the output variables are more or less uniformly distributed. As a result, we can say that the Multinomial Naive Bayes model has the highest accuracy,while the Bernoulli Naive Bayes model has the lowest accuracy.

Explaination :-

It can be observed that the approaches employed to determine the appropriate split have a significant impact on how the decision tree evolves. It is also clear that the entropy model's accuracy is superior in this scenario.

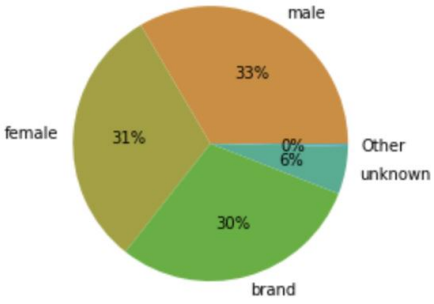
The Multinomial Naïve Bayes model achieves the maximum accuracy, whereas the Bernoulli Naïve Bayes model achieves the lowest accuracy.

Data set Visualization :-



It can be observed that the number of brand female and male classes is almost equal, however the other class values are few.

This pie chart helps to display the percentage values of all class values.



References :-

1. ntroduction to Data mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar
2. https://mljar.com/blog/visualize-decision-tree/
3. https://python-course.eu/machine-learning/naive-bayes-classifier-with-scikit.php
4. Hands-on machine Learning using Sci-kit learn.
5. https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning-894688cb1c0a