Naive Bayes & Decision Tree Report

Student Name and ID: -

1. MEGHANA RAMIDI - 1002036880

2. SWATHI SHANAM - 1002023662

Citations

- https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
- https://seaborn.pydata.org/generated/seaborn.scatterplot.html
- https://data-flair.training/blogs/train-test-set-in-python-ml/
- https://towardsdatascience.com/introduction-to-decision-tree-classifiers-from-scikit-learn-32cd5d23f4d
- https://towardsdatascience.com/gini-index-vs-information-entropy-7a7e4fed3fcb
- https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8
- https://www.youtube.com/watch?v=O2L2Uv9pdDA

1) Describe the Decision Tree methods, and Naive Bayes classifier in details in your own words. Don’t copy paste it fromthe internet. Write it on your own.

Decision Tree:

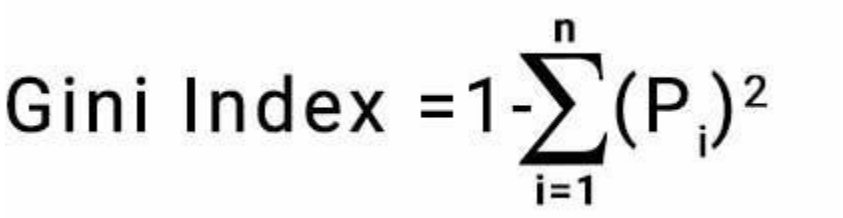
A decision tree is used for collecting the nodes which are intended to create a decision with a mix of class values or numerical target values. Every node or each node of a decision tree is provided with rules of separation which are uniquely used for specific attributes. And the decision tree is providing each node stage into different categories and divides each value in order to minimize chances of the error in the specific order which is specifically selected for the specific use-case.

As the start of the decision tree, it will start with a root node following the children node which has the structure like main information to specific sub-division of main information based on the levels of the Decision tree. A decision tree is always used for gaining the information and for that there are two decision tree methods we can use Gini index and entropy.

Gini Index:

Gini index is used as a statistical probability to calculate specific attribute features which are classified incorrectly when data were chosen randomly. the Gini index varies between the values 0 and 1, where 0 indicates the purity of the division, i.e. All items belonging to a particular category, or a single category are present. And 1 shows the random distribution of elements in different categories. A value of 0.5 Gini Index indicates an equal distribution of the material in certain categories.

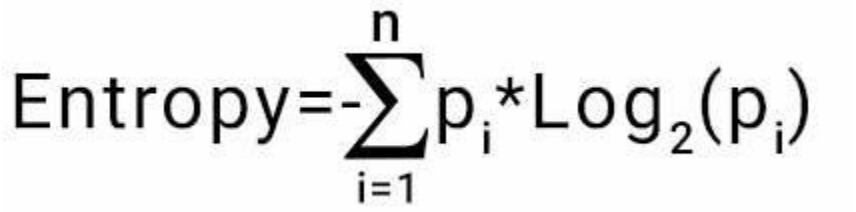
Gini index can be expressed as,



Entropy:

Entropy is a measure of the randomness in the information gain. The higher the value of entropy, it is hard to get any conclusions from that information. It also provides the impurity and heterogeneity of the target variable.

Entropy can be expressed as,



Naive Bayes Classifier:

The Naive Bayes separator is a model used for the split function. This division is based on the theorem of the Bayes. The Bayes theorem is used to find the possibilities of event A happening given that event B took place. We need to find independent opportunities for everything which feature flexibility in the database. Also, opportunities for the target class in the given dataset. Therefore, when calculating to predict the final event that will be happened or not, we can use pre- calculated data.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

2) Describe the datasets like what do you understand from the dataset? and if you have done any pre-processing, and your code, please write down your observations.

Gender Classification dataset gives us an idea about the tweets that are being posted by different genders like male, female. It also gives us an idea about the number of tweets and retweets for a particular account which can be uniquely identified by _unit_id.

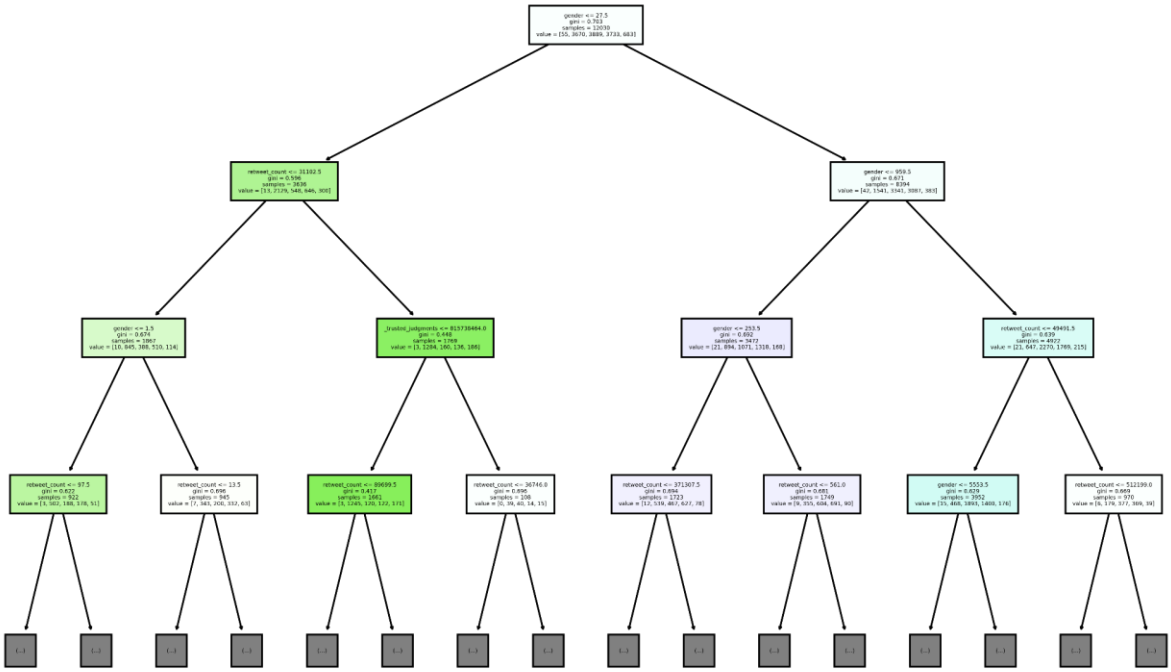
It also has descriptive columns like profile description and text fields.

We have chosen Gender as the Target variable and split the dataset into 60%,20%,20% respectively for Training, Testing and Validation.

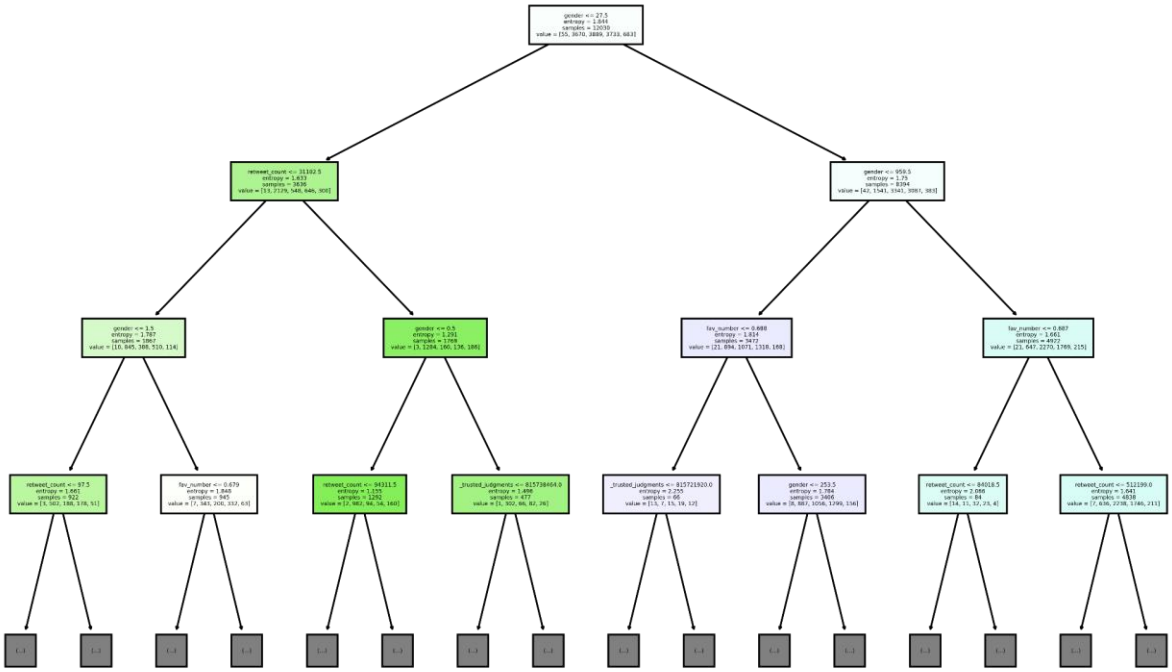
Many columns also have null values and string values, we choose to use the numeric values values to build the model and predict the decision and implement Naive Bayes Classification.

4) Visualization of the decision tree for gini and entropy.

Gini Decision Tree



Entropy Decision Tree



5) Interpret your results, compare gini and entropy

The range of Entropy lies between 0 to 1 and the range of Gini Impurity lies between 0 to 0.5. Based on that Gini Impurity is better as compared to entropy for selecting the best features. In the given observation for the cardio dataset.

```
#5)Gini
#creating model by providing criterion "gini".

gini_classifi_model = DecisionTreeClassifier(criterion = "gini")
#Training the gini model with the data as train_x_data_heart and train_y_data_heart.
trained_gini_classfi = gini_classifi_model.fit(train_x_data_gender, train_y_data_gender)
#predicting the values using tranied gini model by testing test_X_data_heart.
predict_gini_model = trained_gini_classfi.predict(test_x_data_gender)
#Using Accuracy Score printing the model accuracy for measuring the quality of a split.
print ("Gini's Accuracy are : ", accuracy_score(test_y_data_gender, predict_gini_model))
```

```
Gini's Accuracy are :  0.39201995012468827
```

```
#5)entropy
#creating model by providing criterion "entropy".

entropy_classifi_model = DecisionTreeClassifier(criterion = "entropy")

#Training the entropy model with the data as train_x_data_heart and train_y_data_heart.
trained_entropy_clf = entropy_classifi_model.fit(train_x_data_gender,train_y_data_gender)

#predicting the values using tranied entropy model by testing test_X_data_heart.
predict_entropy_model = trained_entropy_clf.predict(test_x_data_gender)

#Using Accuracy Score printing the model accuracy for measuring the quality of a split.
print ("Entropy's Accuracy are : ", accuracy_score(test_y_data_gender,predict_entropy_model))
```

```
Entropy's Accuracy are :  0.39276807980049877
```

Based on the results of accuracy, both have the same accuracy but, in some cases, Gini is best because of faster execution and entropy uses complex logarithms, so performance will be slow, and it will take more timeto execute the dataset.

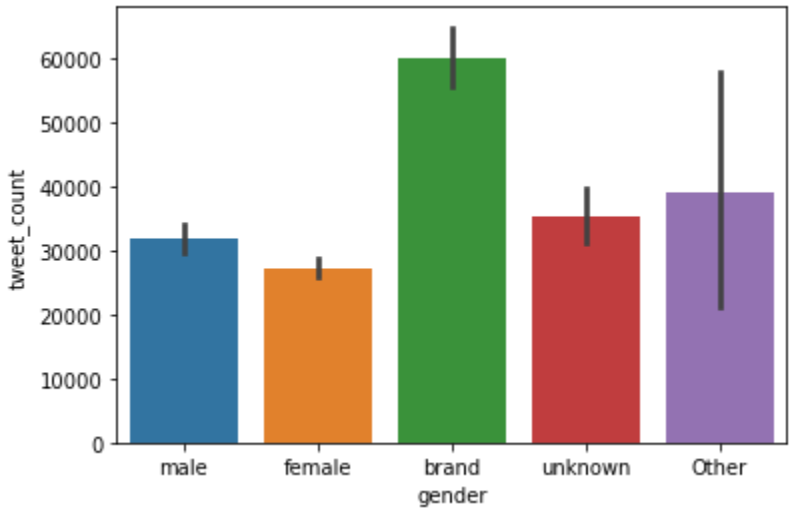
6) Visualize the dataset, for the target variable

Plotting bar plot between gender and tweet_count to see who tweets more often.

print("Bar graph for determinig no.of.tweets compared between different values of gender: ")

sns.barplot(x = "gender", y = "tweet_count", data = df_genderclass)

plt.show()



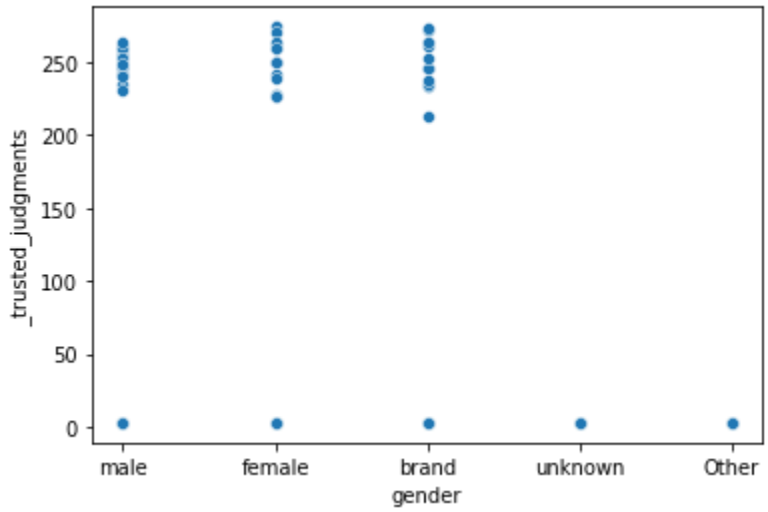
This shows the relation between amount of tweets posted by all the gender types.

Plotting scatter plot between gender and tweet_count to see who tweets more often.

print("Bar graph for determinig trusted judgments compared between different values of gender: ")

sns.scatterplot(data = df_genderclass, x = "gender", y="_trusted_judgments")

plt.show()



This graph shows the relation between the amount of trusted judgments on behalf of male,female and all the gender forms.