DATA MINING ASSIGNMENT-2

DECISION TREE REPORT

1. Describe the Decision Tree methods, and Naive Bayes classifier in details.

DECISION TREE:

Classification Models are built using Decision Tree Mining, a type of data mining approach. Just like the name suggests, it creates classification models in the manner of a tree-like structure. Supervised class learning encompasses this form of mining. We use the dataset to train the data model (split the dataset), and then we use testing data to test the model (to check for accuracy of the data model). A two-step process is followed, to build a classification model.

1. In the first step i.e., learning: A classification model based on training data is built.
2. In the second step i.e., Classification, the accuracy of the model is checked, and then the model is used to classify new data.

The percentage of test instances that are successfully identified determines the classifier's accuracy.

DECISION TREE CONSTRUCTION:

A decision tree is a classification technique that generates a tree and a set of rules to describe the model, with each internal node denoting an attribute test, each branch denoting the test's conclusion, and each leaf node holding a class label. The root node is the topmost node in a tree.
When data is provided, the most crucial task is to choose a root node, which is done using the Gain value =P-M formula, where P is the impurity measure before splitting and M is the impurity measure after splitting.

To calculate the impurity measure of each attribute we have a few methods like Gini Index, Entropy, and misclassification Error

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

where Pi denotes the probability of an element being classified for a distinct class.

$$\text{Entropy} = \sum_{i=1}^{n} -p(c_i) log_2(p(c_i))$$

NAÏVE BAYES CLASSIFIER:

A probabilistic machine learning model called a Naive Bayes classifier is utilized to perform classification tasks. The Bayes theorem lies at the heart of the classifier.

BAYES THEOREM:

Let X be a data tuple.Let H be some hypothesis such that the data tuple X belongs to a specified class C. We want to determine P(H|X) the probability that hypothesis H holds that the tuple X belongs to class C, given the attribute description of X. For Example P(H|X) means the probability that customer X will buy a computer given that we know the customer's age and income

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}$$

2. Describe the datasets and code.

Gender Classifier dataset has 20051 rows and 26 columns. The attributes are _unit_id, _golden, _unit_state, _trusted_judgments, _last_judgment_at, gender, gender:confidence, profile_yn, profile_yn:confidence, created, description, fav_number, gender_gold, link_color, name, profile_yn_gold, profileimage, retweet_count, sidebar_color, text, tweet_coord, tweet_count, tweet_created, tweet_id, tweet_location, user_timezone. Here, gender is the target variable.

It predicts using the training model if gender is female, male, or other. The dataset is split into 60% training, 20% validation, and 30% testing. The training data is used to train a model and this model has to predict the gender of testing data.

Pre-Processing and Code Explanation:

We read the CSV file and convert it into a data frame.

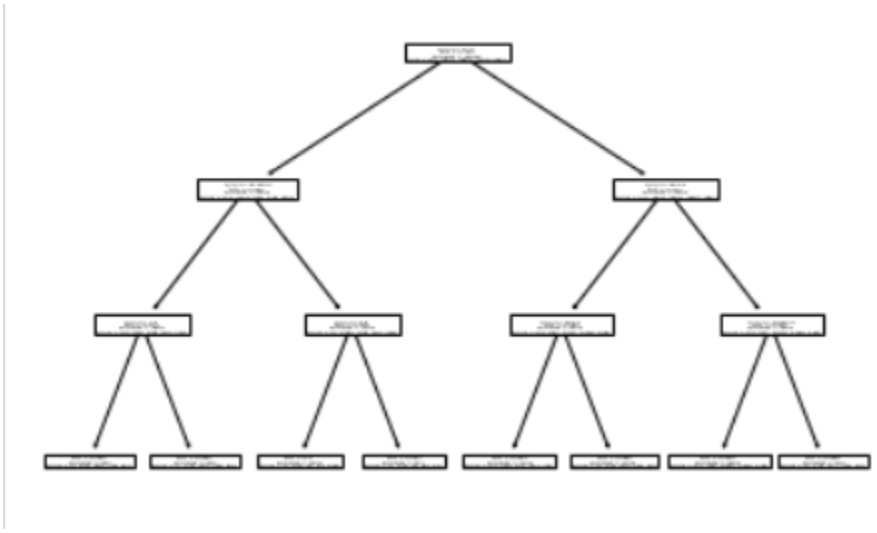
```
df_data = pd.read_csv('GenderClassifier.csv', encoding='latin1')
```

Initially, we checked for any blank spaces / NAN values in the given data set, then there are many other techniques to resolve this, but based on the labels we used a respective method like median to fill the values. We used a label encoder to solve the character data for each and every data for character format.

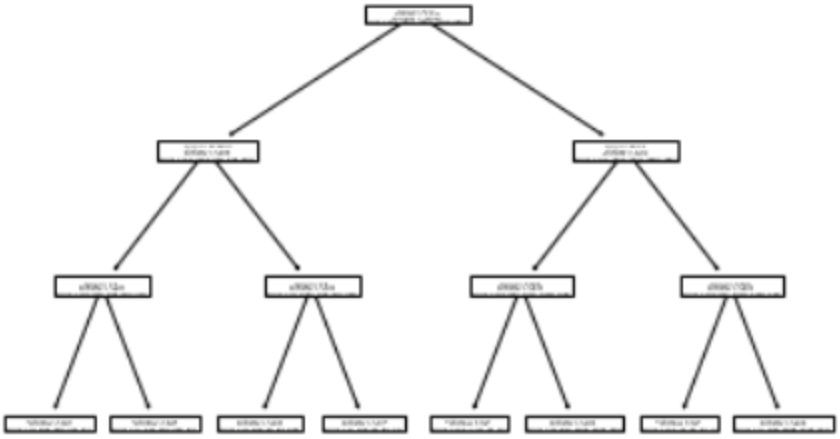
```
new_df = df_data.replace('0', np.nan, inplace= True)

df_data.isnull().sum()

df_data.isnull().any().sum()
```

```
K = df_data
labelencoder_k = LabelEncoder()
labelencoder_k.fit_transform(K['gender'])
K['gender'] = labelencoder_k.fit_transform(K['gender'])
K['_golden'] = labelencoder_k.fit_transform(K['_golden'])
K['_unit_state'] = labelencoder_k.fit_transform(K['_unit_state'])
K['profile_yn'] = labelencoder_k.fit_transform(K['profile_yn'])
K['sidebar_color'] = labelencoder_k.fit_transform(K['sidebar_color'])
K['tweet_id'] = labelencoder_k.fit_transform(K['tweet_id'])
LAA = pd.DataFrame(K)
```

3. **Visualization of the decision tree for Gini and Entropy.**



The above figure is the decision tree using Gini Index and has maximum depth = 3.



The above figure is the decision tree using Entropy and has maximum depth=3.

4. Interpret your results, and do not forget to compare Gini and entropy.

Confusion Matrix:

```
              precision    recall  f1-score   support

       Other       0.09      0.05      0.06        20
       brand       0.63      0.54      0.59      1142
      female       0.49      0.52      0.51      1428
        male       0.36      0.46      0.41      1198
     unknown       0.00      0.00      0.00       222

    accuracy                           0.48      4010
   macro avg       0.32      0.31      0.31      4010
weighted avg       0.47      0.48      0.47      4010

Confusion Matrix:  [[  1   2   8   9   0]
 [  0 621 167 354   0]
 [  2 149 742 535   0]
 [  6 131 509 552   0]
 [  2  77  74  69   0]]
```

Classification Report:

```
                precision    recall  f1-score   support

         Other       0.00      0.00      0.00         0
         brand       0.29      0.25      0.27      1131
```

female	0.42	0.37	0.40	1708
male	0.28	0.36	0.32	1171
accuracy			0.33	4010
macro avg	0.25	0.25	0.24	4010
weighted avg	0.34	0.33	0.34	4010

Using Entropy:

0.4748129675810474

Report :	precision	recall	f1-score	support
Other	0.22	0.10	0.14	20
brand	0.58	0.59	0.58	1142
female	0.48	0.55	0.52	1428
male	0.37	0.37	0.37	1198
unknown	0.00	0.00	0.00	222
accuracy			0.47	4010
macro avg	0.33	0.32	0.32	4010
weighted avg	0.45	0.47	0.46	4010

Confusion Matrix:
```
[[  2   3   9   6   0]
 [  3 673 184 282   0]
 [  1 210 789 428   0]
 [  3 185 570 440   0]
 [  0  92  81  49   0]]
```

Confusion matrix

<AxesSubplot:>

Using Gini Index:

```
0.4743142144638404

Report :                precision    recall  f1-score   support

       Other       0.00      0.00      0.00        20
       brand       0.58      0.59      0.58      1142
      female       0.48      0.55      0.51      1428
        male       0.37      0.37      0.37      1198
     unknown       0.00      0.00      0.00       222

    accuracy                           0.47      4010
   macro avg       0.28      0.30      0.29      4010
weighted avg       0.45      0.47      0.46      4010

Confusion Matrix:  [[  0   5   9   6   0]
 [  0 671 189 282   0]
 [  0 210 789 429   0]
 [  0 185 571 442   0]
 [  0  92  81  49   0]]
```

Confusion matrix

<AxesSubplot:>

Both Gini Index and Entropy have different accuracy values with some accuracy.

Naïve Bayes Classifier:

Gaussian Naive Bayes

```
-------------------- Test Accuracy --------------------
Confusion Matrix:

[[   9 1093   50   30   16]
 [   3 1346   32   24   23]
 [   5 1012   81   32   12]
 [   0  208    7    4    3]
 [   0   18    1    0    1]]
Classification Report:

              precision    recall  f1-score   support

       Other       0.02      0.05      0.03        20
       brand       0.47      0.07      0.12      1142
      female       0.37      0.94      0.53      1428
        male       0.53      0.01      0.01      1198
     unknown       0.04      0.02      0.03       222

    accuracy                           0.36      4010
   macro avg       0.29      0.22      0.14      4010
```

```
weighted avg       0.43      0.36      0.23      4010

-------------------- Val Accuracy --------------------
Confusion Matrix:

[[   8 1157   46   30   22]
 [   9 1288   31   29   26]
 [   2 1007   74   39    8]
 [   0  196    6    6    4]
 [   0   21    1    0    0]]
Classification Report:

              precision    recall  f1-score   support

       Other       0.00      0.00      0.00        22
       brand       0.47      0.07      0.11      1130
      female       0.35      0.93      0.51      1383
        male       0.42      0.01      0.01      1263
     unknown       0.06      0.03      0.04       212

    accuracy                           0.34      4010
   macro avg       0.26      0.21      0.14      4010
weighted avg       0.39      0.34      0.21      4010
```

Multinomial Naive Bayes

```
-------------------- Test Accuracy --------------------
Confusion Matrix:

[[498 510 186   4   0]
 [378 871 178   1   0]
 [247 229 665   1   0]
 [ 66  85  70   1   0]
 [  4  11   4   1   0]]
Classification Report:

              precision    recall  f1-score   support

       Other       0.00      0.00      0.00        20
       brand       0.60      0.58      0.59      1142
      female       0.51      0.61      0.56      1428
        male       0.42      0.42      0.42      1198
     unknown       0.12      0.00      0.01       222

    accuracy                           0.51      4010
   macro avg       0.33      0.32      0.31      4010
weighted avg       0.49      0.51      0.49      4010

-------------------- Val Accuracy --------------------
Confusion Matrix:

[[547 522 193   1   0]
 [329 902 151   1   0]
```

```
 [249 243 637   1   0]
 [ 65  79  68   0   0]
 [  7  11   2   2   0]]
Classification Report:

              precision    recall  f1-score   support

       Other       0.00      0.00      0.00        22
       brand       0.61      0.56      0.58      1130
      female       0.51      0.65      0.57      1383
        male       0.46      0.43      0.44      1263
     unknown       0.00      0.00      0.00       212

    accuracy                           0.52      4010
   macro avg       0.32      0.33      0.32      4010
weighted avg       0.49      0.52      0.50      4010
```

Complement Naive Bayes

```
-------------------- Test Accuracy --------------------
Confusion Matrix:

[[438 503 226  25   6]
 [352 848 205  17   6]
 [225 219 690   6   2]
 [ 60  87  69   6   0]
 [  4  11   3   2   0]]
Classification Report:

              precision    recall  f1-score   support

       Other       0.00      0.00      0.00        20
       brand       0.58      0.60      0.59      1142
      female       0.51      0.59      0.55      1428
        male       0.41      0.37      0.38      1198
     unknown       0.11      0.03      0.04       222

    accuracy                           0.49      4010
   macro avg       0.32      0.32      0.31      4010
weighted avg       0.47      0.49      0.48      4010

-------------------- Val Accuracy --------------------
Confusion Matrix:

[[504 506 228  19   6]
 [297 887 178  17   4]
 [226 238 646  19   1]
 [ 64  71  72   4   1]
 [  6  10   2   3   1]]
Classification Report:

              precision    recall  f1-score   support
```

```
              precision    recall  f1-score   support

       Other       0.08      0.05      0.06        22
       brand       0.57      0.57      0.57      1130
      female       0.52      0.64      0.57      1383
        male       0.46      0.40      0.43      1263
     unknown       0.06      0.02      0.03       212

    accuracy                           0.51      4010
   macro avg       0.34      0.34      0.33      4010
weighted avg       0.49      0.51      0.50      4010
```

Bernoulli Naive Bayes

```
-------------------- Test Accuracy --------------------
Confusion Matrix:

[[ 413  672  113    0    0]
 [ 268 1053  107    0    0]
 [ 233  324  585    0    0]
 [  49  118   55    0    0]
 [   2   16    2    0    0]]
Classification Report:

              precision    recall  f1-score   support

       Other       0.00      0.00      0.00        20
       brand       0.68      0.51      0.58      1142
      female       0.48      0.74      0.58      1428
        male       0.43      0.34      0.38      1198
     unknown       0.00      0.00      0.00       222

    accuracy                           0.51      4010
   macro avg       0.32      0.32      0.31      4010
weighted avg       0.49      0.51      0.49      4010

-------------------- Val Accuracy --------------------
Confusion Matrix:

[[ 452  681  130    0    0]
 [ 240 1053   90    0    0]
 [ 238  346  546    0    0]
 [  61  100   51    0    0]
 [   6   14    2    0    0]]
Classification Report:

              precision    recall  f1-score   support

       Other       0.00      0.00      0.00        22
       brand       0.67      0.48      0.56      1130
      female       0.48      0.76      0.59      1383
        male       0.45      0.36      0.40      1263
     unknown       0.00      0.00      0.00       212

    accuracy                           0.51      4010
```

macro avg	0.32	0.32	0.31	4010
weighted avg	0.50	0.51	0.49	4010