

Mid-term Assignment
Data Warehousing & Data Mining

Name: Parvez, Fuad Al

ID: 17-35187-2

Sec: A

To

Dr. Mahbub Chowdhury Mishu

**Department of Computer Science
American International University Bangladesh (AIUB)**

I am Choosing **Naïve Bayes** classification technique.

Naïve Bayes: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The category is determined on the basis of a training set of data which contains observations whose category membership is already known. Bayes Theorem helps us to find the probability of a hypothesis given our prior knowledge.

Choosing Dataset: I have selected this dataset from Kaggle website.

C116				Not drink	
	A	B	C	D	E
1	Favorite Color	Favorite Music Genre	Favorite Beverage	Favorite Soft Drink	Gender
2	Cool	Rock	Vodka	7UP/Sprite	Female
3	Neutral	Hip hop	Vodka	Coca Cola/Pepsi	Female
4	Warm	Rock	Wine	Coca Cola/Pepsi	Female
5	Warm	Folk	Whiskey	Fanta	Female
6	Cool	Rock	Vodka	Coca Cola/Pepsi	Female
7	Warm	Jazz	Not drink	Fanta	Female
8	Cool	Pop	Beer	Coca Cola/Pepsi	Female
9	Warm	Pop	Whiskey	Fanta	Female
10	Warm	Rock	Other	7UP/Sprite	Female
11	Neutral	Pop	Wine	Coca Cola/Pepsi	Female
12	Cool	Pop	Other	7UP/Sprite	Female
13	Warm	Pop	Other	7UP/Sprite	Female
14	Warm	Pop	Wine	7UP/Sprite	Female
15	Warm	Electronic	Wine	Coca Cola/Pepsi	Female
16	Cool	Rock	Beer	Coca Cola/Pepsi	Female
17	Warm	Jazz	Wine	Coca Cola/Pepsi	Female
18	Cool	Pop	Wine	7UP/Sprite	Female
19	Cool	Rock	Other	Coca Cola/Pepsi	Female
20	Cool	Rock	Other	Coca Cola/Pepsi	Female
21	Cool	Pop	Not drink	7UP/Sprite	Female
22	Cool	Pop	Beer	Fanta	Female
23	Warm	Jazz	Whiskey	Fanta	Female
24	Cool	Rock	Vodka	Coca Cola/Pepsi	Female
25	Warm	Pop	Other	Coca Cola/Pepsi	Female
26	Cool	Folk	Whiskey	7UP/Sprite	Female
27	Warm	R&B and soul	Whiskey	Coca Cola/Pepsi	Female
28	Cool	Pop	Beer	Other	Female
29	Cool	Pop	Not drink	Other	Female
train data set					

Here,

Total number of Attribute: 5

- Favorite Color - Data type Nominal (categorical)
- Favorite Music Genre - Data type Nominal (categorical)
- Favorite Beverage - Data type Nominal (categorical)
- Favorite Soft Drink - Data type Nominal (categorical)
- Gender - Data type Nominal (categorical)

Gender represents class attribute.

Total instance: 119

Favorite Color	Favorite Music Genre	Favorite Beverage	Favorite Soft Drink	Gender
Cool	Rock	Vodka	7up	Male
Neutral	Hip-hop	Wine	Sprite	Female
Warm	Folk	Whiskey	Coca cola	
	Jazz	Not Drink	Pepsi	
	Pop	Beer	Fanta	
	Electronic	Other	Other	
	R&B and soul			
	Blues			

Step 1: Original Dataset or Train Dataset

train data set.csv - Excel

17-35187-2@student.laib.edu

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Paste Cut Copy Format Painter Clipboard Font Alignment Number Styles Cells Editing Sensitivity

Calibri 11 A A Wrap Text General Conditional Formatting Format as Table Cell Styles Insert Delete Format AutoSum Fill Sort & Find & Filter Select Clear Sensitivity

C116 Not drink

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Favorite Color	Favorite Music Genre	Favorite Beverage	Favorite Soft Drink	Gender											
2	Cool	Rock	Vodka	7UP/Sprite	Female											
3	Neutral	Hip hop	Vodka	Coca Cola/Pepsi	Female											
4	Warm	Rock	Wine	Coca Cola/Pepsi	Female											
5	Warm	Folk	Whiskey	Fanta	Female											
6	Cool	Rock	Vodka	Coca Cola/Pepsi	Female											
7	Warm	Jazz	Not drink	Fanta	Female											
8	Cool	Pop	Beer	Coca Cola/Pepsi	Female											
9	Warm	Pop	Whiskey	Fanta	Female											
10	Warm	Rock	Other	7UP/Sprite	Female											
11	Neutral	Pop	Wine	Coca Cola/Pepsi	Female											
12	Cool	Pop	Other	7UP/Sprite	Female											
13	Warm	Pop	Other	7UP/Sprite	Female											
14	Warm	Pop	Wine	7UP/Sprite	Female											
15	Warm	Electronic	Wine	Coca Cola/Pepsi	Female											
16	Cool	Rock	Beer	Coca Cola/Pepsi	Female											
17	Warm	Jazz	Wine	Coca Cola/Pepsi	Female											
18	Cool	Pop	Wine	7UP/Sprite	Female											
19	Cool	Rock	Other	Coca Cola/Pepsi	Female											
20	Cool	Rock	Other	Coca Cola/Pepsi	Female											
21	Cool	Pop	Not drink	7UP/Sprite	Female											
22	Cool	Pop	Beer	Fanta	Female											
23	Warm	Jazz	Whiskey	Fanta	Female											
24	Cool	Rock	Vodka	Coca Cola/Pepsi	Female											
25	Warm	Pop	Other	Coca Cola/Pepsi	Female											
26	Cool	Folk	Whiskey	7UP/Sprite	Female											
27	Warm	R&B and soul	Whiskey	Coca Cola/Pepsi	Female											
28	Cool	Pop	Beer	Other	Female											
29	Cool	Pop	Not drink	Other	Female											

train data set

Step 2: Test Dataset

I have made this test dataset by using Weka. Weka can simply make test dataset from any train dataset.

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Clipboard Font Alignment Number Styles Cells Editing Sensitivity

Calibri 11 A A Wrap Text General \$ % 00 00 Conditional Formatting Format as Table Cell Styles Insert Delete Format AutoSum Fill Sort & Find & Filter Select Sensitivity

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	'Favorite Color'	'Favorite Music Genre'	'Favorite Beverage'	'Favorite Soft Drink'	Gender											
2	Cool	Folk	Other	'Coca Cola/Pepsi'	Male											
3	Neutral	'Hip hop'	'Not drink'	Fanta	Male											
4	Warm	Rock	Wine	'Coca Cola/Pepsi'	Female											
5	Warm	Rock	Vodka	7UP/Sprite	Male											
6	Cool	Rock	Beer	'Coca Cola/Pepsi'	Male											
7	Warm	'R&B and soul'	Whiskey	'Coca Cola/Pepsi'	Female											
8	Warm	Jazz/Blues	Other	7UP/Sprite	Female											
9	Warm	Jazz	Whiskey	Other	Female											
10	Warm	Folk	Other	Fanta	Male											
11	Cool	Pop	Whiskey	Other	Male											
12	Cool	Rock	'Not drink'	'Coca Cola/Pepsi'	Male											
13	Warm	Pop	Other	7UP/Sprite	Female											
14	Cool	Pop	Whiskey	'Coca Cola/Pepsi'	Female											
15	Cool	Electronic	Beer	'Coca Cola/Pepsi'	Male											
16	Neutral	'R&B and soul'	Vodka	Other	Male											
17	Warm	Jazz	Wine	'Coca Cola/Pepsi'	Female											
18	Cool	Pop	Wine	Fanta	Female											
19	Warm	Pop	Wine	Fanta	Male											
20	Cool	Rock	Vodka	'Coca Cola/Pepsi'	Male											
21	Cool	Pop	'Not drink'	7UP/Sprite	Female											
22																
23																
24																
25																
26																
27																
28																
29																

test data set 2

Step 3: Import train dataset

Import train dataset into WEKA. (File name: train dataset.csv)

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply Stop

Current relation
Relation: train data set
Instances: 119
Attributes: 5
Sum of weights: 119

Attributes
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> Favorite Color
2	<input checked="" type="checkbox"/> Favorite Music Genre
3	<input checked="" type="checkbox"/> Favorite Beverage
4	<input checked="" type="checkbox"/> Favorite Soft Drink
5	<input checked="" type="checkbox"/> Gender

Remove

Selected attribute
Name: Favorite Music Genre
Missing: 0 (0%)
Distinct: 8
Type: Nominal
Unique: 1 (1%)

No.	Label	Count	Weight
1	Rock	35	35.0
2	Hip hop	14	14.0
3	Folk	7	7.0
4	Jazz	6	6.0
5	Pop	33	33.0
6	Electronic	11	11.0
7	R&B and soul	12	12.0
8	Jazz/Blues	1	1.0

Class: Gender (Nom) Visualize All

Genre	Count
Rock	35
Hip hop	14
Folk	7
Jazz	6
Pop	33
Electronic	11
R&B and soul	12
Jazz/Blues	1

Status: OK Log x 0

Step 4: Train Dataset analysis

By using Naïve Bayes algorithm & training set, my train dataset accuracy is 70.5882%

Correctly Classified Instances 84 70.5882 %

Incorrectly Classified Instances 35 29.4118 %

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section on the left shows 'Use training set' selected. The 'Classifier output' pane on the right displays the following results:

```
Time taken to build model: 0 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0.01 seconds

=== Summary ===
Correctly Classified Instances      84      70.5882 %
Incorrectly Classified Instances    35      29.4118 %
Kappa statistic                    0.4076
Mean absolute error                 0.4239
Root mean squared error             0.4504
Relative absolute error             85.2529 %
Root relative squared error         91.9475 %
Total Number of Instances          119

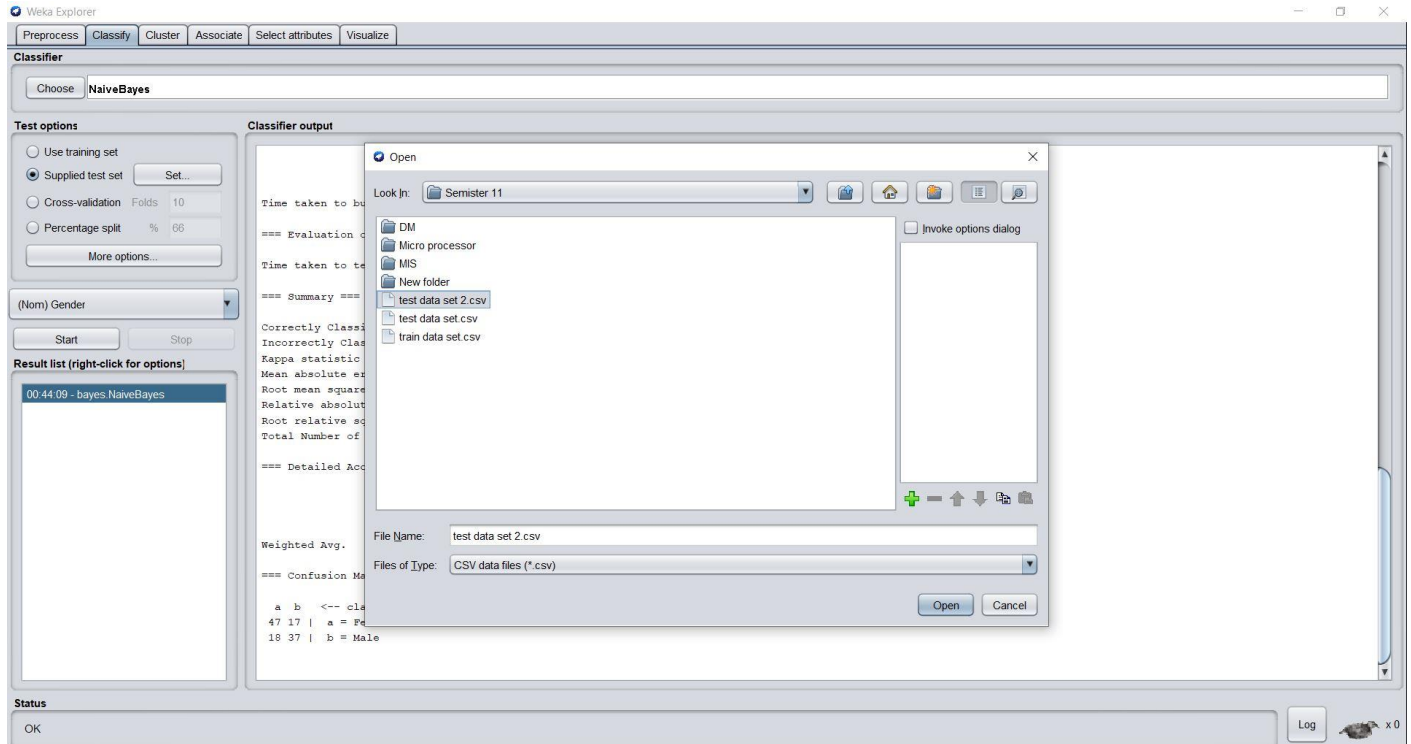
=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0.734    0.327    0.723    0.734    0.729    0.408    0.731    0.774    Female
0.673    0.266    0.685    0.673    0.679    0.408    0.731    0.668    Male
Weighted Avg.    0.706    0.299    0.706    0.706    0.706    0.408    0.731    0.725

=== Confusion Matrix ===
      a  b  <-- classified as
47 17 |  a = Female
18 37 |  b = Male
```

Classifier output		
Attribute	Female	Male
	(0.54)	(0.46)
=====		
Favorite Color		
Cool	36.0	33.0
Neutral	5.0	9.0
Warm	26.0	16.0
[total]	67.0	58.0
Favorite Music Genre		
Rock	20.0	17.0
Hip hop	6.0	10.0
Folk	5.0	4.0
Jazz	6.0	2.0
Pop	24.0	11.0
Electronic	4.0	9.0
R&B and soul	5.0	9.0
Jazz/Blues	2.0	1.0
[total]	72.0	63.0
Favorite Beverage		
Vodka	6.0	10.0
Wine	12.0	8.0
Whiskey	11.0	6.0
Not drink	13.0	15.0
Beer	13.0	13.0
Other	15.0	9.0
[total]	70.0	61.0
Favorite Soft Drink		
7UP/Sprite	14.0	11.0
Coca Cola/Pepsi	32.0	27.0
Fanta	15.0	12.0
Other	7.0	9.0
[total]	68.0	59.0

Step 5: Test Dataset Analysis from Train Dataset

Import test dataset into WEKA. (File name: test dataset 2.csv)



After run the test dataset, my test dataset accuracy is 75%

Correctly Classified Instances 15 75 %

Incorrectly Classified Instances 5 25 %

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set
☒ Supplied test set Set...
☐ Cross-validation Folds: 10
☐ Percentage split %: 66
More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

- 00:44:09 - bayes NaiveBayes
- 00:45:51 - misc InputMappedClassifier

Classifier output

```
(nominal) Gender            --> 5 (nominal) Gender

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances      15            75    %
Incorrectly Classified Instances    5            25    %
Kappa statistic                    0.5192
Mean absolute error                0.377
Root mean squared error            0.4314
Relative absolute error            74.8355 %
Root relative squared error        85.4065 %
Total Number of Instances        20

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
1.000   0.455   0.643   1.000   0.783   0.592   0.879   0.870   Female
0.545   0.000   1.000   0.545   0.706   0.592   0.879   0.916   Male
Weighted Avg.   0.750   0.205   0.839   0.750   0.740   0.592   0.879   0.895

=== Confusion Matrix ===

  a b   <-- classified as
  9 0 | a = Female
  5 6 | b = Male
```

Status

OK Log x 0

Classifier output

Attribute	Female (0.54)	Male (0.46)
=====		
Favorite Color		
Cool	36.0	33.0
Neutral	5.0	9.0
Warm	26.0	16.0
[total]	67.0	58.0
Favorite Music Genre		
Rock	20.0	17.0
Hip hop	6.0	10.0
Folk	5.0	4.0
Jazz	6.0	2.0
Pop	24.0	11.0
Electronic	4.0	9.0
R&B and soul	5.0	9.0
Jazz/Blues	2.0	1.0
[total]	72.0	63.0
Favorite Beverage		
Vodka	6.0	10.0
Wine	12.0	8.0
Whiskey	11.0	6.0
Not drink	13.0	15.0
Beer	13.0	13.0
Other	15.0	9.0
[total]	70.0	61.0
Favorite Soft Drink		
7UP/Sprite	14.0	11.0
Coca Cola/Pepsi	32.0	27.0
Fanta	15.0	12.0
Other	7.0	9.0
[total]	68.0	59.0

Step 6: Cross Validation Analysis

The accuracy of cross validation is 60.50%. I took 15 Folds to test the dataset.

Correctly Classified Instances 72 60.5042 %

Incorrectly Classified Instances 47 39.4958 %

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 15

☐ Percentage split % 66

More options...

(Norm) Gender

Start Stop

Result list (right-click for options)

- 00:44:09 - bayes.NaiveBayes
- 00:45:51 - misc.InputMappedClassifier
- 00:49:15 - bayes.NaiveBayes**

Classifier output

```
Fanta            15.0   12.0
Other            7.0    9.0
[total]           68.0   59.0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      72            60.5042 %
Incorrectly Classified Instances    47            39.4958 %
Kappa statistic                    0.2004
Mean absolute error                0.4788
Root mean squared error            0.5127
Relative absolute error            96.2103 %
Root relative squared error        102.7333 %
Total Number of Instances        119

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRG Area   Class
          0.672    0.473    0.623    0.672    0.647    0.201   0.570    0.586   Female
          0.527    0.328    0.580    0.527    0.552    0.201   0.570    0.498   Male
Weighted Avg.   0.605    0.406    0.603    0.605    0.603    0.201   0.570    0.545

=== Confusion Matrix ===
  a b   <-- classified as
43 21 | a = Female
26 29 | b = Male
```

Status

OK Log x0

Classifier output

Attribute	Female	Male
	(0.54)	(0.46)
=====		
Favorite Color		
Cool	36.0	33.0
Neutral	5.0	9.0
Warm	26.0	16.0
[total]	67.0	58.0
Favorite Music Genre		
Rock	20.0	17.0
Hip hop	6.0	10.0
Folk	5.0	4.0
Jazz	6.0	2.0
Pop	24.0	11.0
Electronic	4.0	9.0
R&B and soul	5.0	9.0
Jazz/Blues	2.0	1.0
[total]	72.0	63.0
Favorite Beverage		
Vodka	6.0	10.0
Wine	12.0	8.0
Whiskey	11.0	6.0
Not drink	13.0	15.0
Beer	13.0	13.0
Other	15.0	9.0
[total]	70.0	61.0
Favorite Soft Drink		
7UP/Sprite	14.0	11.0
Coca Cola/Pepsi	32.0	27.0
Fanta	15.0	12.0
Other	7.0	9.0
[total]	68.0	59.0

Step 7: Percentage Split Analysis

The accuracy of my dataset's percentage split is 62.5%

Correctly Classified Instances 30 62.5 %

Incorrectly Classified Instances 18 37.5 %

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 15

☒ Percentage split % 60

More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

- 00:44:09 - bayes NaiveBayes
- 00:45:51 - misc.InputMappedClassifier
- 00:49:15 - bayes NaiveBayes
- 01:09:41 - bayes NaiveBayes**

Classifier output

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances      30      62.5 %
Incorrectly Classified Instances    18      37.5 %
Kappa statistic                    0.25
Mean absolute error                0.4722
Root mean squared error            0.5016
Relative absolute error             94.4438 %
Root relative squared error        99.5744 %
Total Number of Instances          48

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MDC     ROC Area  PRC Area  Class
0.667    0.417    0.615    0.667    0.640    0.251    0.575    0.605    Female
0.583    0.333    0.636    0.583    0.609    0.251    0.575    0.555    Male
Weighted Avg.   0.625    0.375    0.626    0.625    0.624    0.251    0.575    0.580

=== Confusion Matrix ===

  a  b  <-- classified as
16  8  |  a = Female
10 14 |  b = Male
```

Status

OK

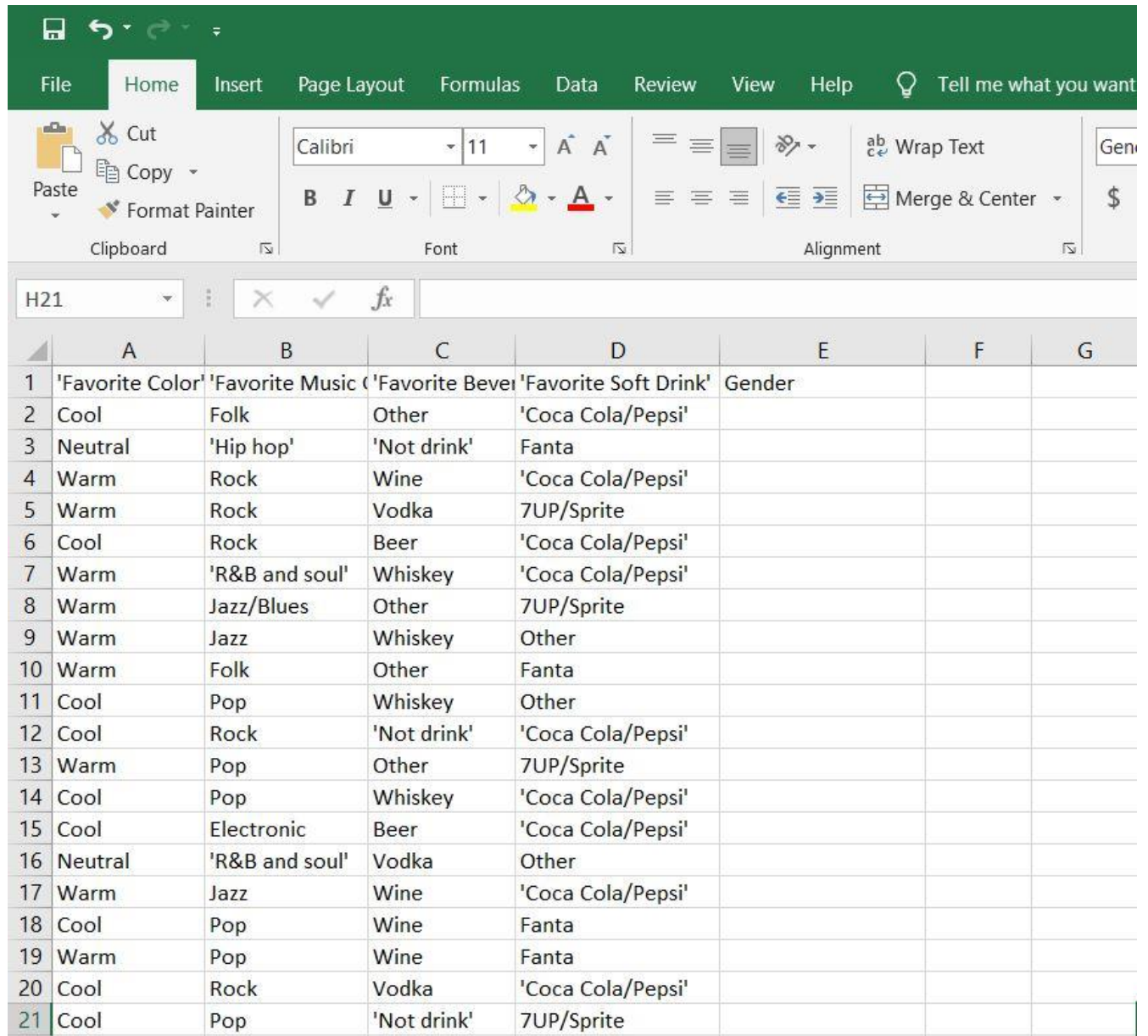
Log x 0

Classifier output

Attribute	Female	Male
	(0.54)	(0.46)
=====		
Favorite Color		
Cool	36.0	33.0
Neutral	5.0	9.0
Warm	26.0	16.0
[total]	67.0	58.0
Favorite Music Genre		
Rock	20.0	17.0
Hip hop	6.0	10.0
Folk	5.0	4.0
Jazz	6.0	2.0
Pop	24.0	11.0
Electronic	4.0	9.0
R&B and soul	5.0	9.0
Jazz/Blues	2.0	1.0
[total]	72.0	63.0
Favorite Beverage		
Vodka	6.0	10.0
Wine	12.0	8.0
Whiskey	11.0	6.0
Not drink	13.0	15.0
Beer	13.0	13.0
Other	15.0	9.0
[total]	70.0	61.0
Favorite Soft Drink		
7UP/Sprite	14.0	11.0
Coca Cola/Pepsi	32.0	27.0
Fanta	15.0	12.0
Other	7.0	9.0
[total]	68.0	59.0

Step 8: Predict the result

Here, I will take test dataset but gender attribute will remain empty as I will predict this.



The screenshot shows the Microsoft Excel interface with the 'Home' tab selected. The ribbon includes options for Clipboard, Font, and Alignment. The formula bar shows 'H21'. The dataset is organized in a table with 8 columns (A-G) and 21 rows (1-21). The first four columns represent input features: 'Favorite Color', 'Favorite Music', 'Favorite Beverage', and 'Favorite Soft Drink'. The fifth column, 'Gender', is the target variable to be predicted. The last two columns (F and G) are currently empty.

	A	B	C	D	E	F	G
1	'Favorite Color'	'Favorite Music'	'Favorite Beverage'	'Favorite Soft Drink'	Gender		
2	Cool	Folk	Other	'Coca Cola/Pepsi'			
3	Neutral	'Hip hop'	'Not drink'	Fanta			
4	Warm	Rock	Wine	'Coca Cola/Pepsi'			
5	Warm	Rock	Vodka	7UP/Sprite			
6	Cool	Rock	Beer	'Coca Cola/Pepsi'			
7	Warm	'R&B and soul'	Whiskey	'Coca Cola/Pepsi'			
8	Warm	Jazz/Blues	Other	7UP/Sprite			
9	Warm	Jazz	Whiskey	Other			
10	Warm	Folk	Other	Fanta			
11	Cool	Pop	Whiskey	Other			
12	Cool	Rock	'Not drink'	'Coca Cola/Pepsi'			
13	Warm	Pop	Other	7UP/Sprite			
14	Cool	Pop	Whiskey	'Coca Cola/Pepsi'			
15	Cool	Electronic	Beer	'Coca Cola/Pepsi'			
16	Neutral	'R&B and soul'	Vodka	Other			
17	Warm	Jazz	Wine	'Coca Cola/Pepsi'			
18	Cool	Pop	Wine	Fanta			
19	Warm	Pop	Wine	Fanta			
20	Cool	Rock	Vodka	'Coca Cola/Pepsi'			
21	Cool	Pop	'Not drink'	7UP/Sprite			

Classifier output

=== Predictions on test set ===

inst#	actual	predicted	error	prediction
1	1:?	1:Female	0.642	
2	1:?	2:Male	0.806	
3	1:?	1:Female	0.693	
4	1:?	2:Male	0.507	
5	1:?	1:Female	0.503	
6	1:?	1:Female	0.566	
7	1:?	1:Female	0.821	
8	1:?	1:Female	0.822	
9	1:?	1:Female	0.738	
10	1:?	1:Female	0.693	
11	1:?	2:Male	0.533	
12	1:?	1:Female	0.833	
13	1:?	1:Female	0.775	
14	1:?	2:Male	0.724	
15	1:?	2:Male	0.913	
16	1:?	1:Female	0.852	
17	1:?	1:Female	0.748	
18	1:?	1:Female	0.815	
19	1:?	2:Male	0.622	
20	1:?	1:Female	0.636	

=== Evaluation on test set ===

prediction data set.csv

	A	B	C	D	E	F	G	H	I
1	'Favorite Color'	'Favorite Music'	'Favorite Beve'	'Favorite Soft Drink'	Gender				
2	Cool	Folk	Other	'Coca Cola/Pepsi'	Female				
3	Neutral	'Hip hop'	'Not drink'	Fanta	Male				
4	Warm	Rock	Wine	'Coca Cola/Pepsi'	Female				
5	Warm	Rock	Vodka	7UP/Sprite	Male				
6	Cool	Rock	Beer	'Coca Cola/Pepsi'	Female				
7	Warm	'R&B and soul'	Whiskey	'Coca Cola/Pepsi'	Female				
8	Warm	Jazz/Blues	Other	7UP/Sprite	Female				
9	Warm	Jazz	Whiskey	Other	Female				
10	Warm	Folk	Other	Fanta	Female				
11	Cool	Pop	Whiskey	Other	Female				
12	Cool	Rock	'Not drink'	'Coca Cola/Pepsi'	Male				
13	Warm	Pop	Other	7UP/Sprite	Female				
14	Cool	Pop	Whiskey	'Coca Cola/Pepsi'	Female				
15	Cool	Electronic	Beer	'Coca Cola/Pepsi'	Male				
16	Neutral	'R&B and soul'	Vodka	Other	Male				
17	Warm	Jazz	Wine	'Coca Cola/Pepsi'	Female				
18	Cool	Pop	Wine	Fanta	Female				
19	Warm	Pop	Wine	Fanta	Female				
20	Cool	Rock	Vodka	'Coca Cola/Pepsi'	Male				
21	Cool	Pop	'Not drink'	7UP/Sprite	Female				

Reason for choosing Naïve Bayes:

- Better algorithm for working with text classification.
- Relatively simple approach.
- Independent assumption.
- Dataset analysis accuracy is good enough.

Reason for choosing the Dataset:

The dataset I have been choose from Kaggle is good dataset. As the dataset contains 5 Attribute with 119 instances, so it is kind of average big data. Big dataset analysis showed me much better idea.

Discussion:

- As I have used WEKA software for the first time, faced some difficulties by using it.
- Some types of data were not working in WEKA, so chosen a relevant simple dataset was challenging.
- As WEKA could accept .csv format file, I have used this type of data file in excel.
- WEKA runtime is so fast.
- As I have learned 3 classifier technique, it was more challenging which one to use. Finally, I have selected the Naïve Bayes.
- I have checked train dataset accuracy, test set accuracy, cross validation, percentage split & prediction of a target attribute.