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)

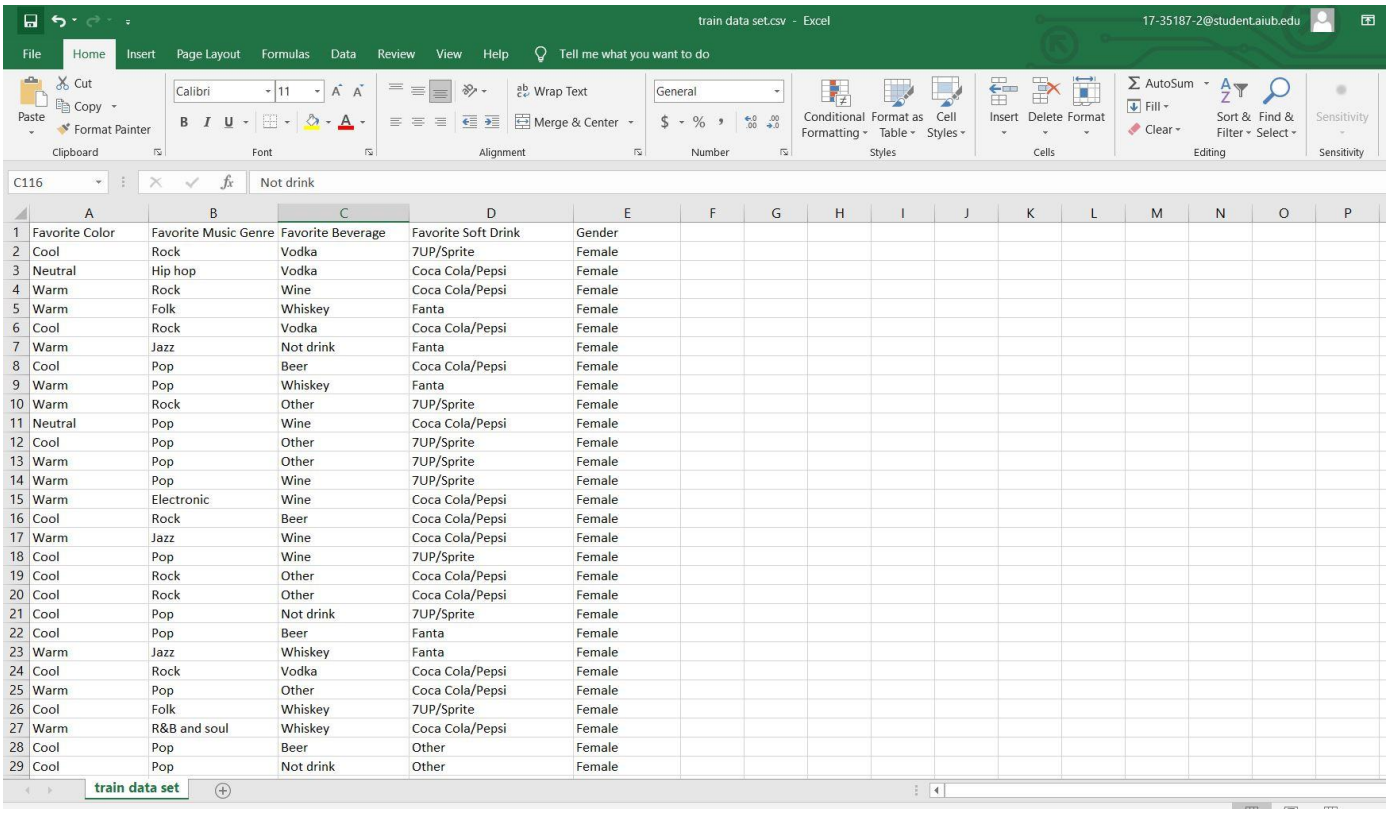
Gender represents class attribute.

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			

WEKA Process:

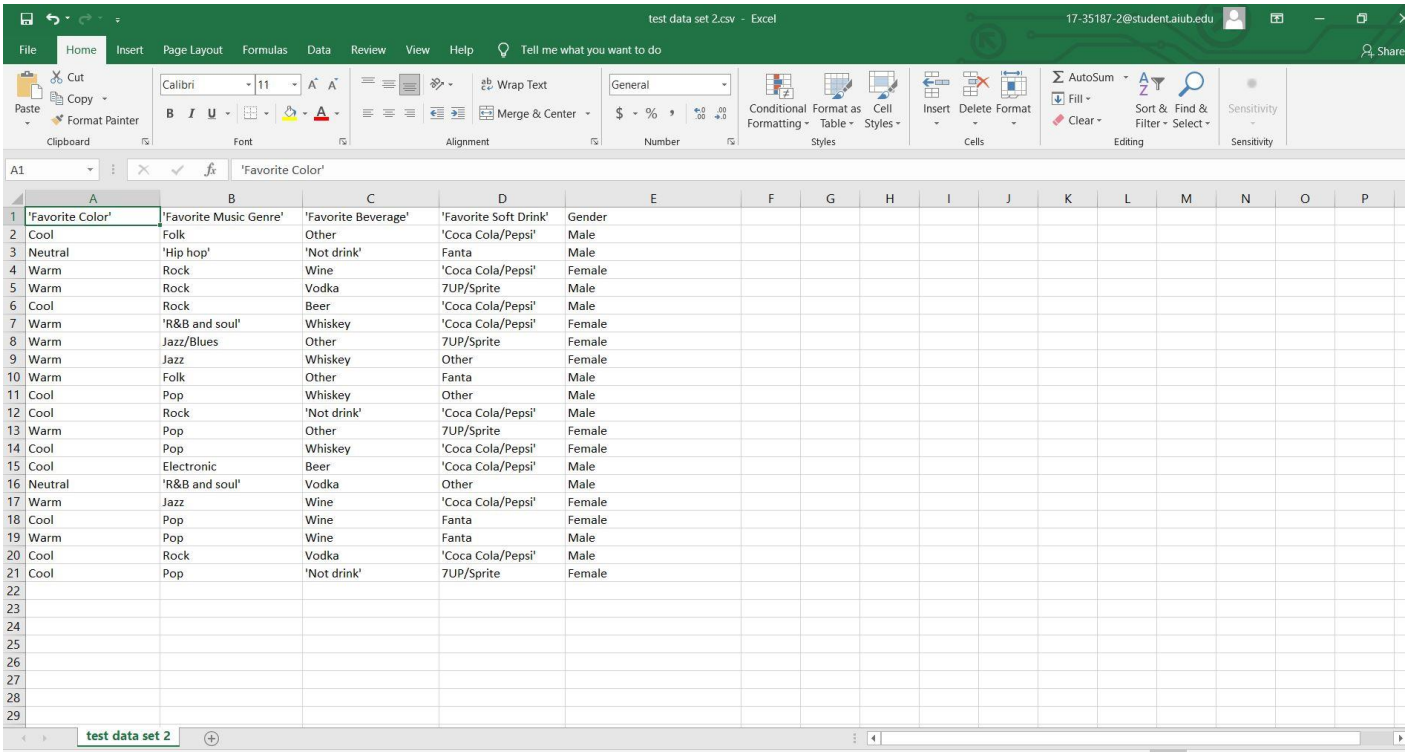
Step 1: Original Dataset or Train Dataset



	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

Step 2: Test Dataset

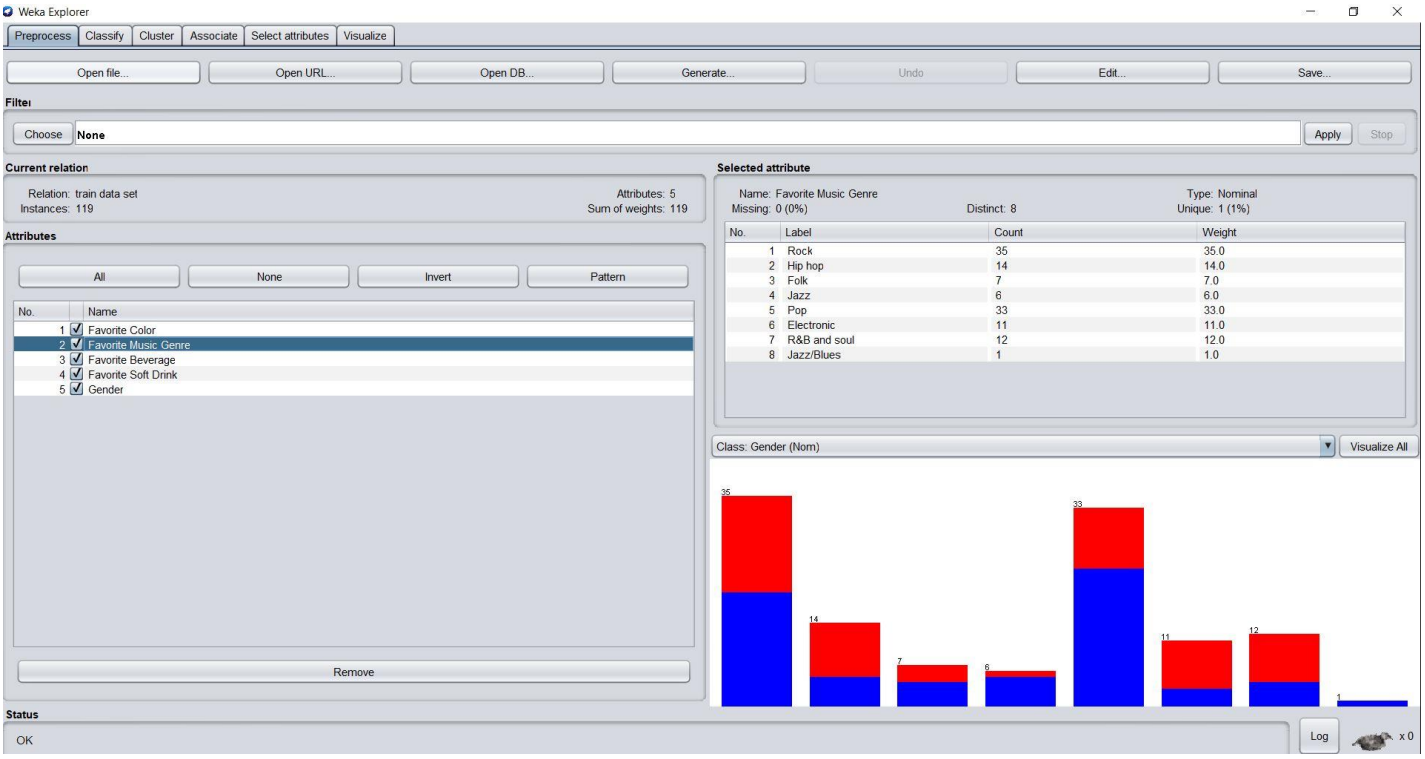
I have made this test dataset by using Weka. Weka can simply make test dataset from any train dataset.



	A	B	C	D	E
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

Step 3: Import train dataset

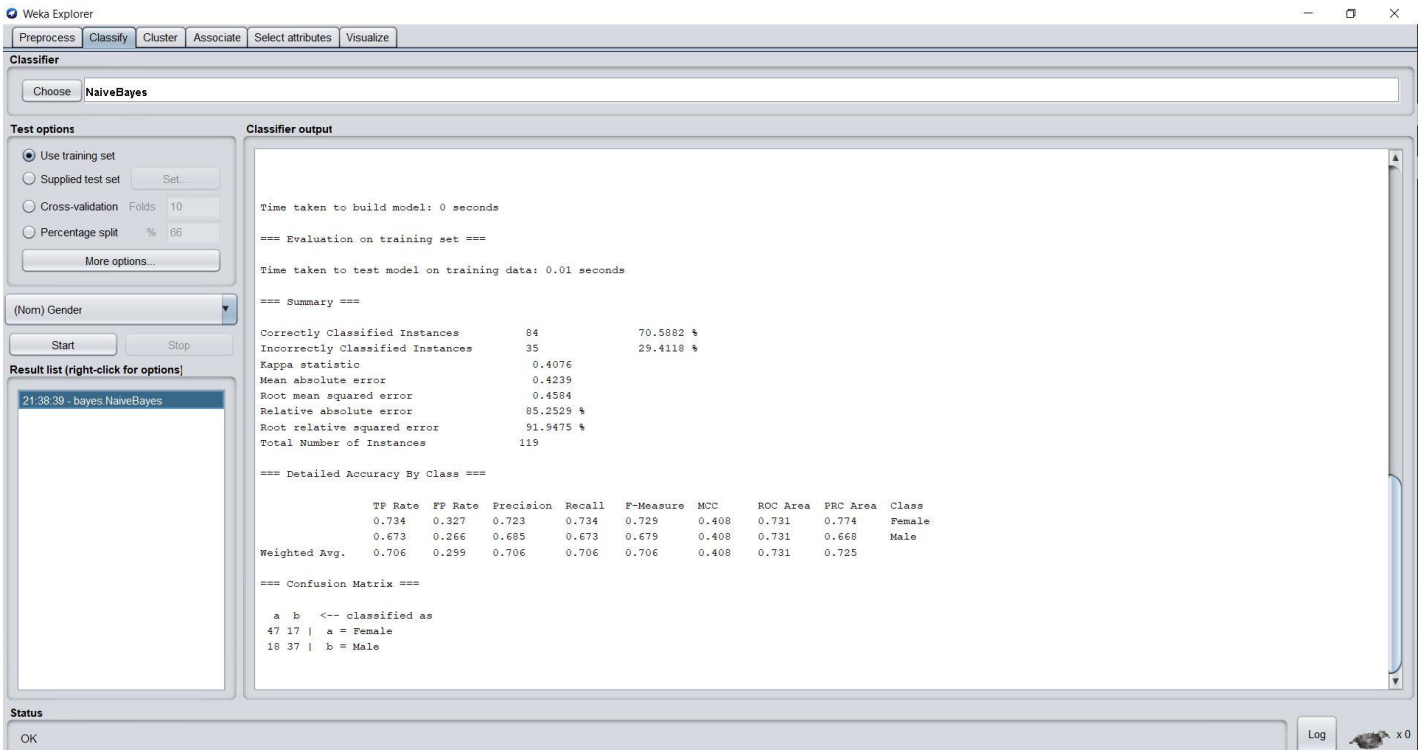
Import train dataset into WEKA. (File name: train dataset.csv)



Step 4: Train Dataset analysis

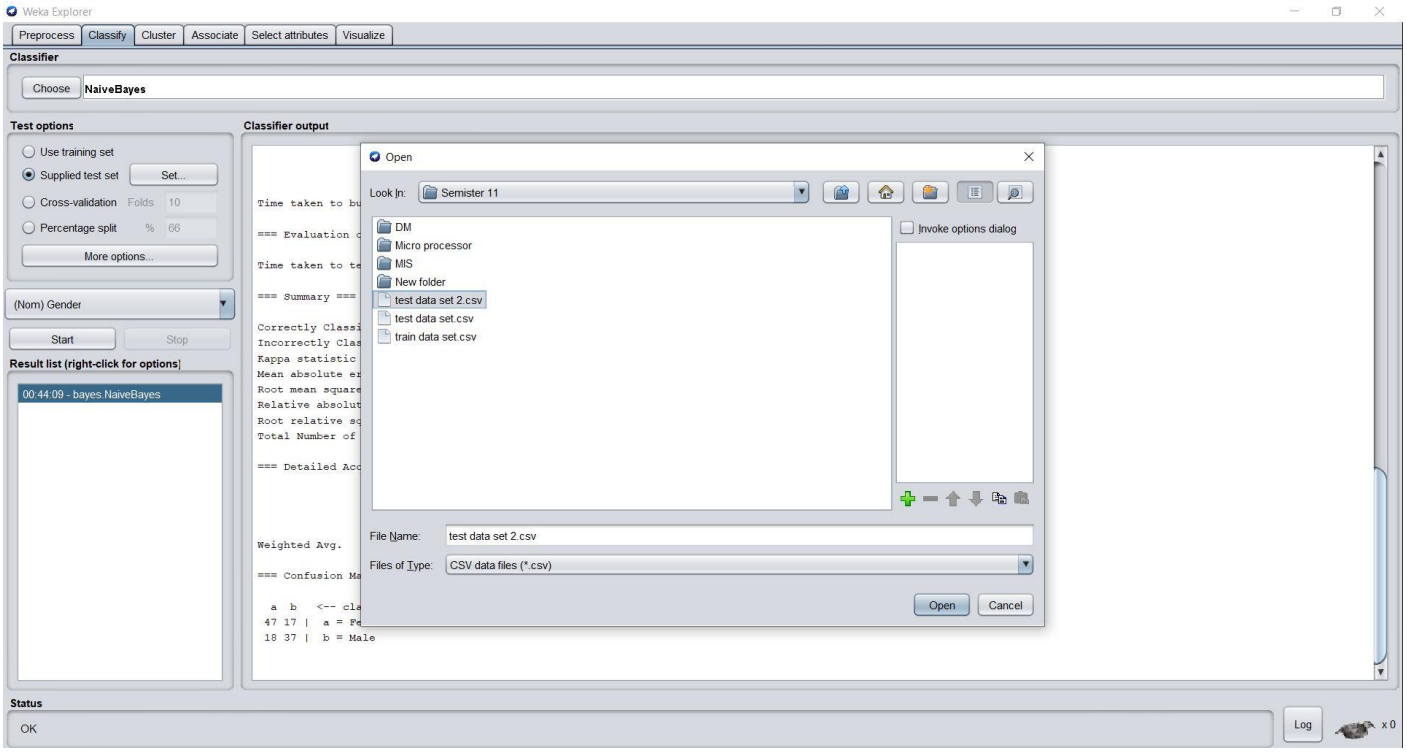
By using Naïve Bayes algorithm & training set, my train dataset accuracy is 70.5882%

Correctly Classified Instances	84	70.5882 %
Incorrectly Classified Instances	35	29.4118 %



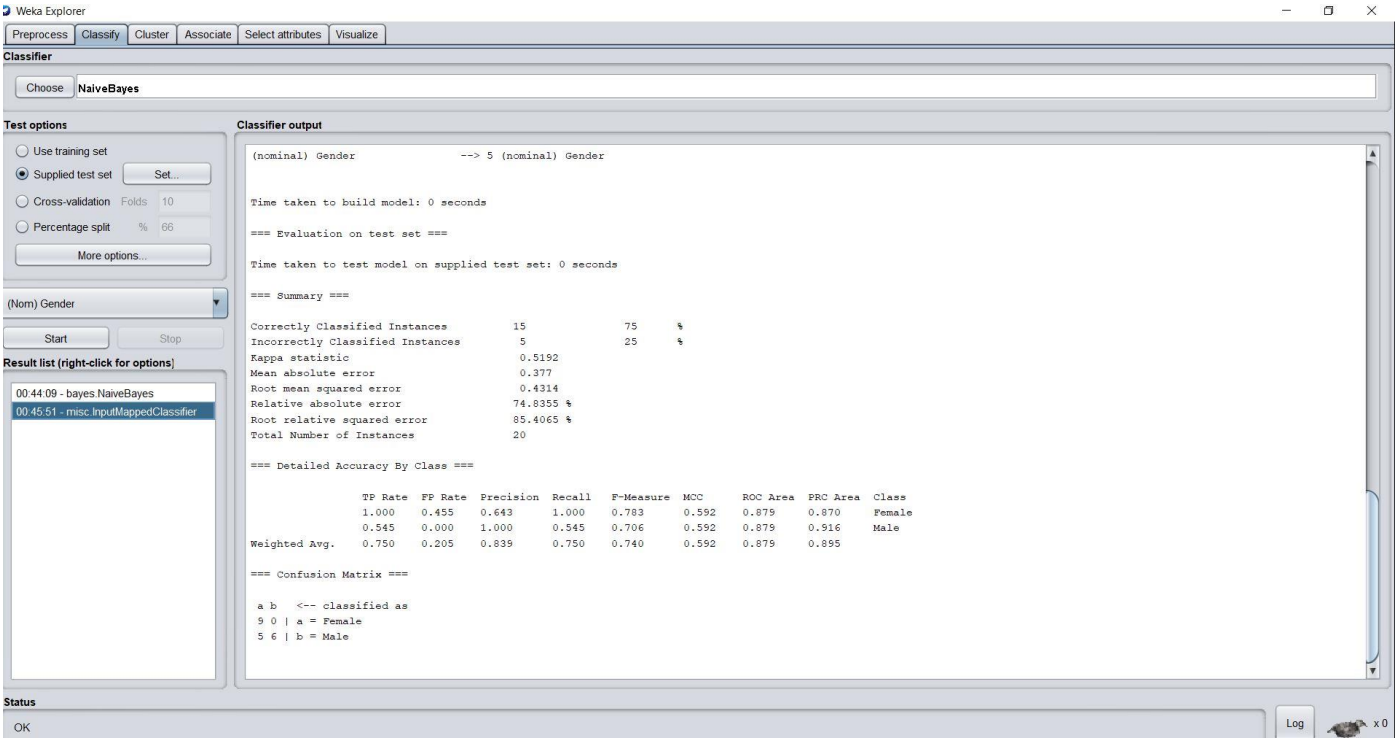
Step 5: Test Dataset Analysis from Train Dataset

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	%



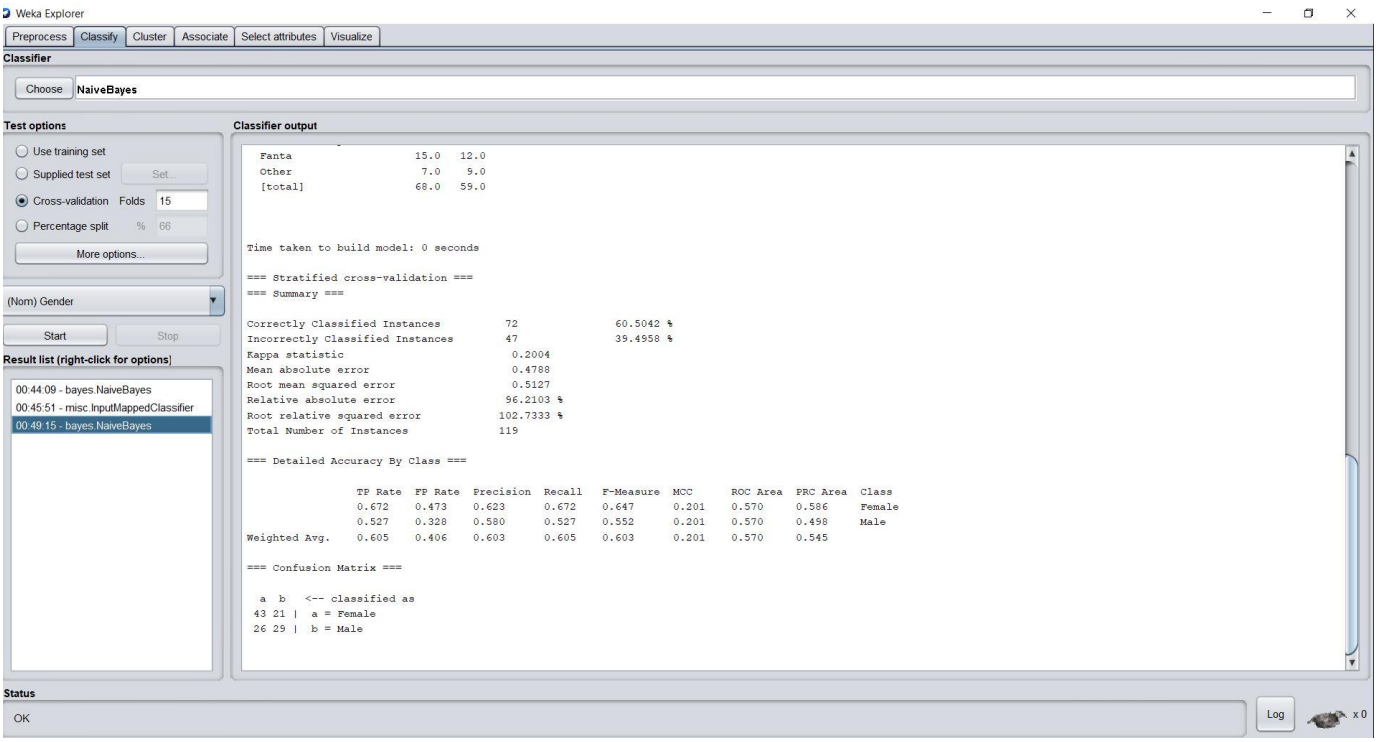
Classifier output

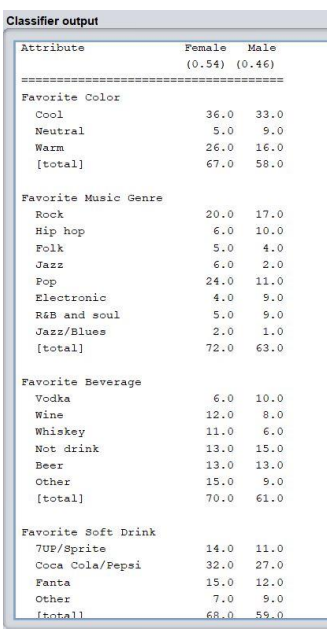
Attribute	Female (0.54)	Male (0.46)
Favorite Color		
Cool	36.0	33.0
Neutral	5.0	9.0
Warm	26.0	16.0
[total]	67.0	58.0
Favorite Music Genre		
Rock	20.0	17.0
Hip hop	6.0	10.0
Folk	5.0	4.0
Jazz	6.0	2.0
Pop	24.0	11.0
Electronic	4.0	9.0
R&B and soul	5.0	9.0
Jazz/Blues	2.0	1.0
[total]	72.0	63.0
Favorite Beverage		
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 %

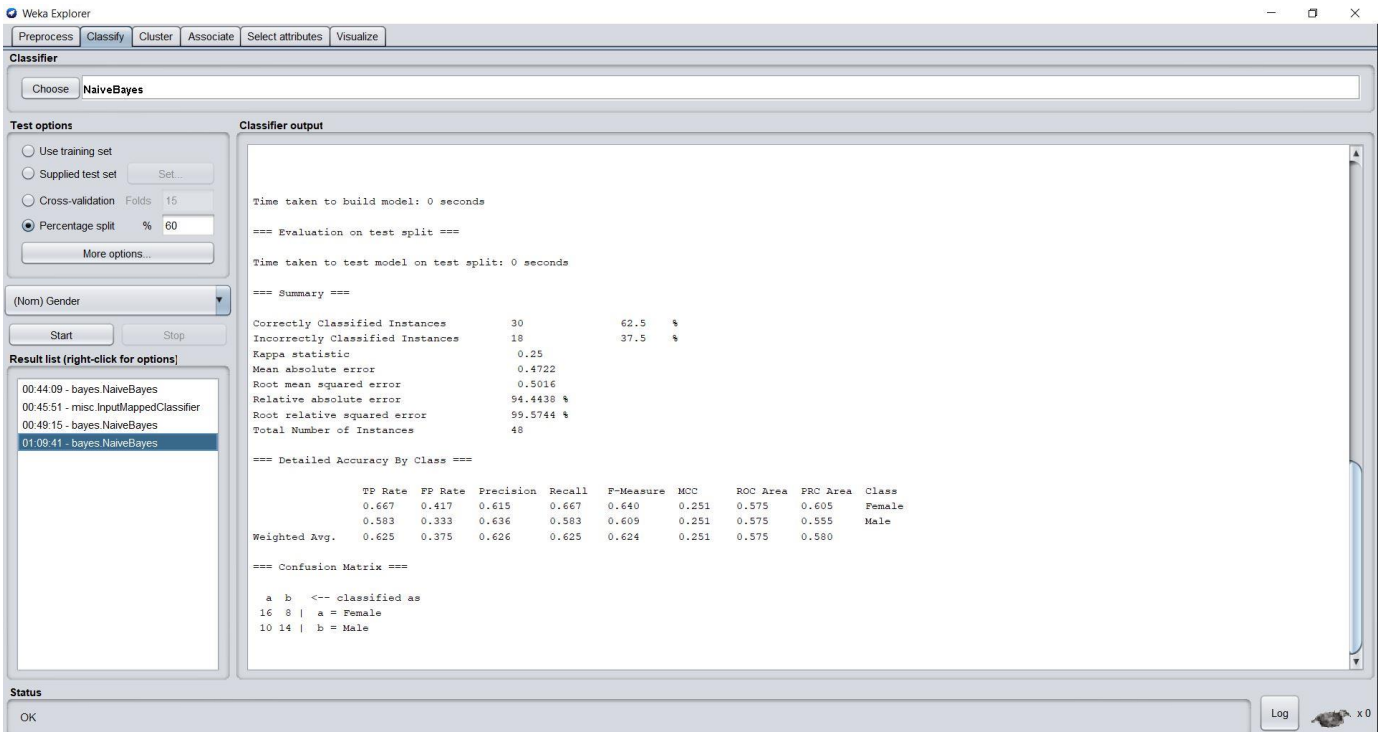


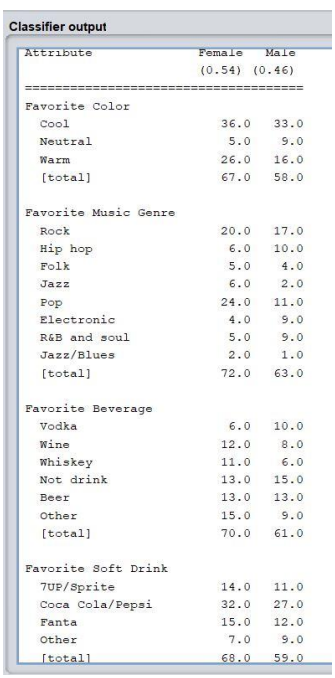


Step 7: Percentage Split Analysis

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

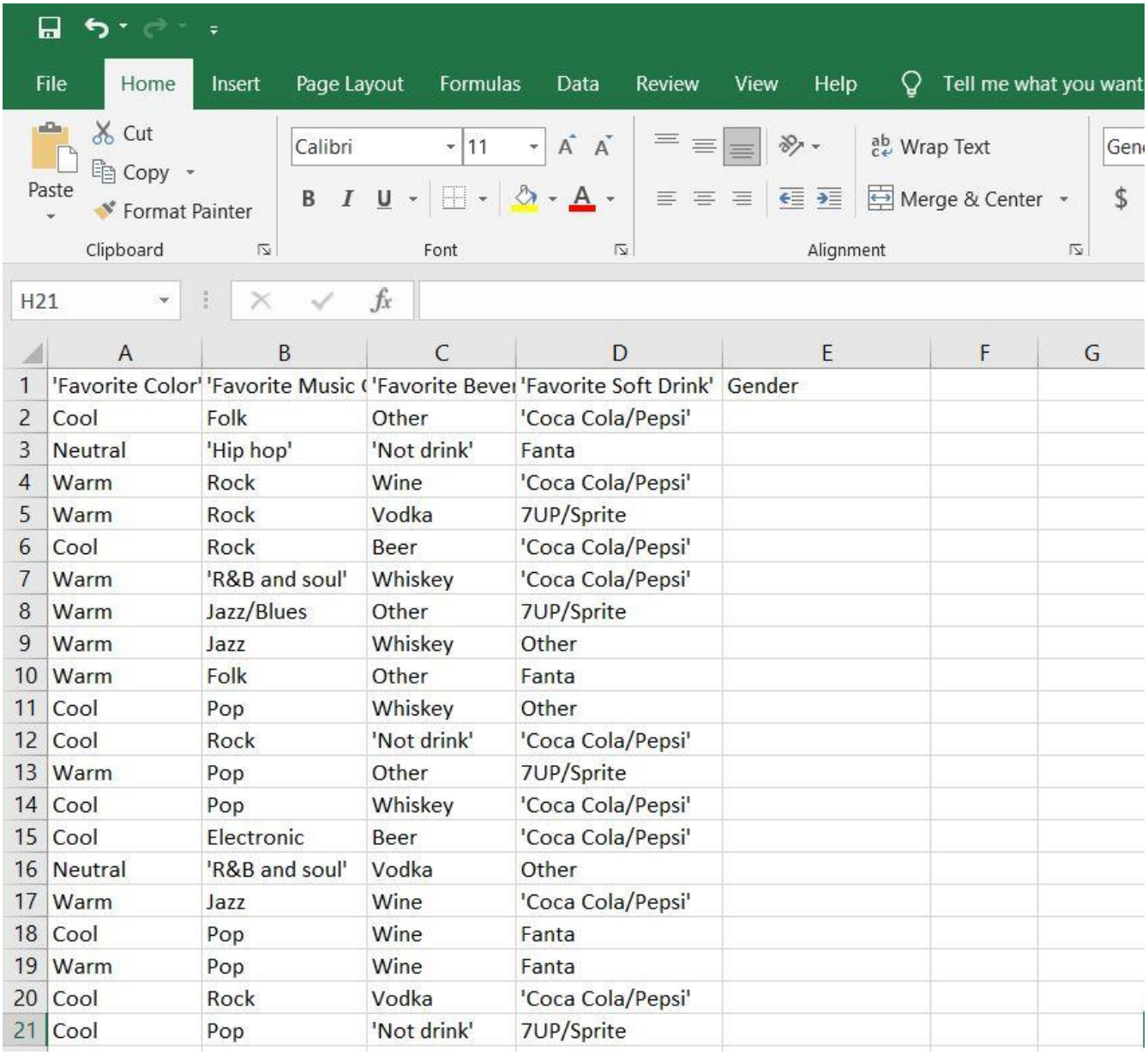
Correctly Classified Instances	30	62.5	%
Incorrectly Classified Instances	18	37.5	%





Step 8: Predict the result

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



	A	B	C	D	E	F	G
1	'Favorite Color'	'Favorite Music (	'Favorite Bever	'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
1	'Favorite Color'	'Favorite Music (	'Favorite Bevei	'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.