AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

ASSIGNMENT

DATA WAREHOUSING AND DATA MINING

SUBMITTED BY

NAME: DAS, BISHWAGIT

ID:18-36730-1

SECTION: A

SUBMITTED TO

DR. MD. MAHBUB CHOWDHURY MISHU

ASSISTANT PROFESSOR AND HEAD

DEPT. OF COMPUTER SCIENCE FST, AIUB

SUBMISSION DATE

30/06/2021

Introduction:

The first algorithm to consider when tackling a text classification problem is Naive Bayes. The Naive Bayes algorithm is a straightforward machine learning algorithm. It aids in classifying a new observation class from a set of classes. The type is determined using a training set of data that includes words that have already been organized. The Bayes Theorem aids us in determining the likelihood of a hypothesis based on our prior information. With the Naive Bayes method, creating models and making predictions is quick.

Dataset:

<u>Attributes Information:</u>

Number of Attributes: 4 (including the class attribute)

Customer (categorical, 38 categories)

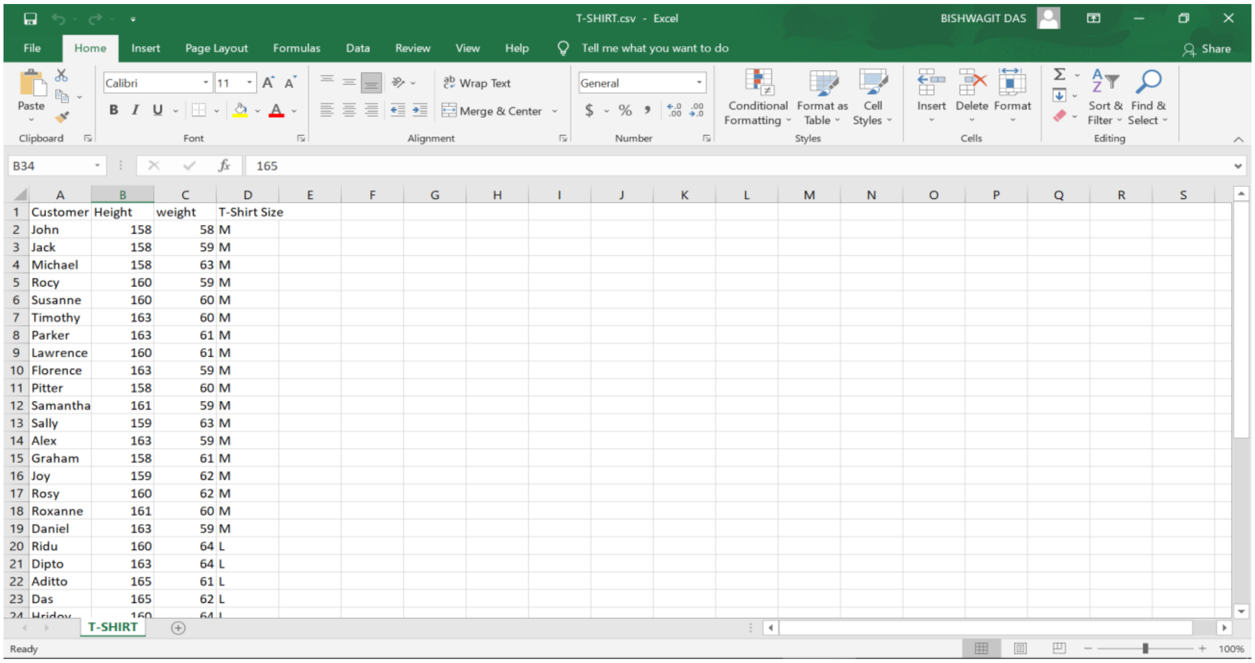
Height (numerical)

Weight (numerical)

T-Shirt Size (categorical) 1=M, 2=L

Here, the class attribute is T-Shirt Size.

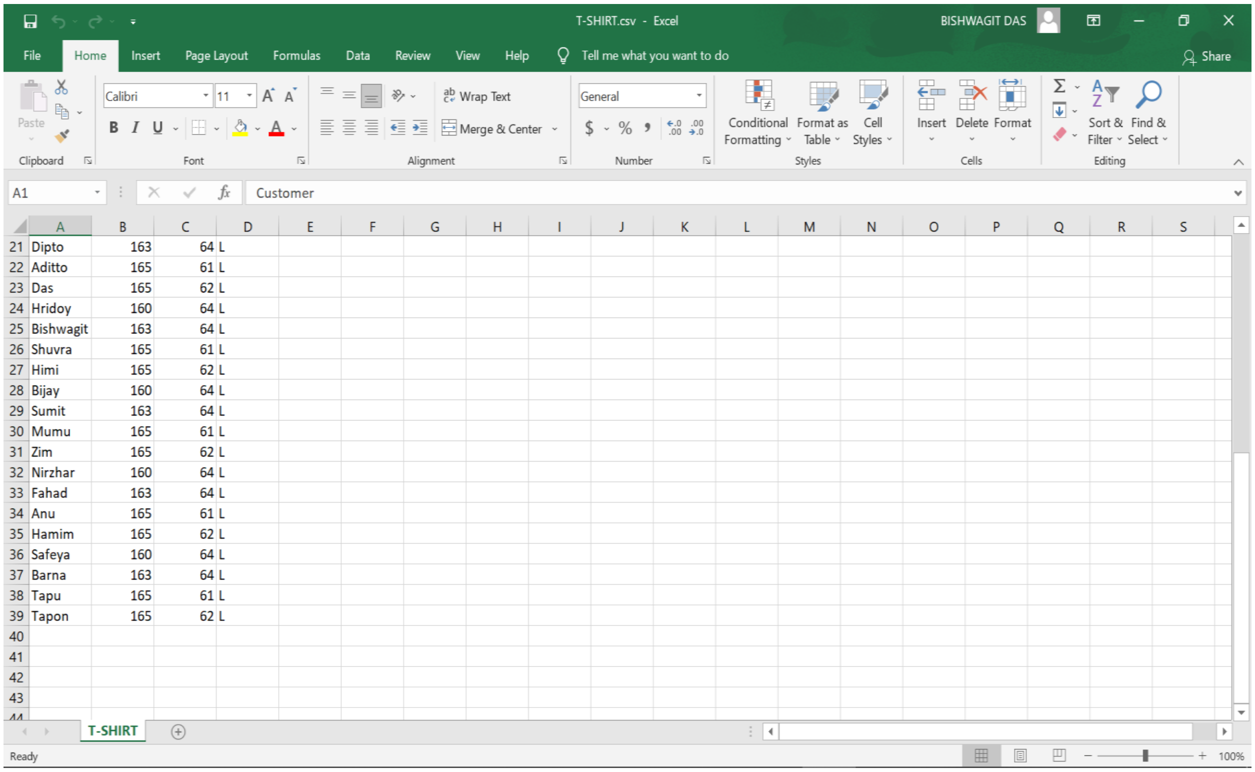
I have prepared the data set according to the given data and attribute information. This report is conducted using the WEKA tool. This tool is a combination of machine learning and data mining techniques. This dataset in .csv format.



B34 165

	A	B	C	D
1	Customer	Height	weight	T-Shirt Size
2	John	158	58	M
3	Jack	158	59	M
4	Michael	158	63	M
5	Rocy	160	59	M
6	Susanne	160	60	M
7	Timothy	163	60	M
8	Parker	163	61	M
9	Lawrence	160	61	M
10	Florence	163	59	M
11	Pitter	158	60	M
12	Samantha	161	59	M
13	Sally	159	63	M
14	Alex	163	59	M
15	Graham	158	61	M
16	Joy	159	62	M
17	Rosy	160	62	M
18	Roxanne	161	60	M
19	Daniel	163	59	M
20	Ridu	160	64	L
21	Dipto	163	64	L
22	Aditto	165	61	L
23	Das	165	62	L
24	Hriday	160	64	L

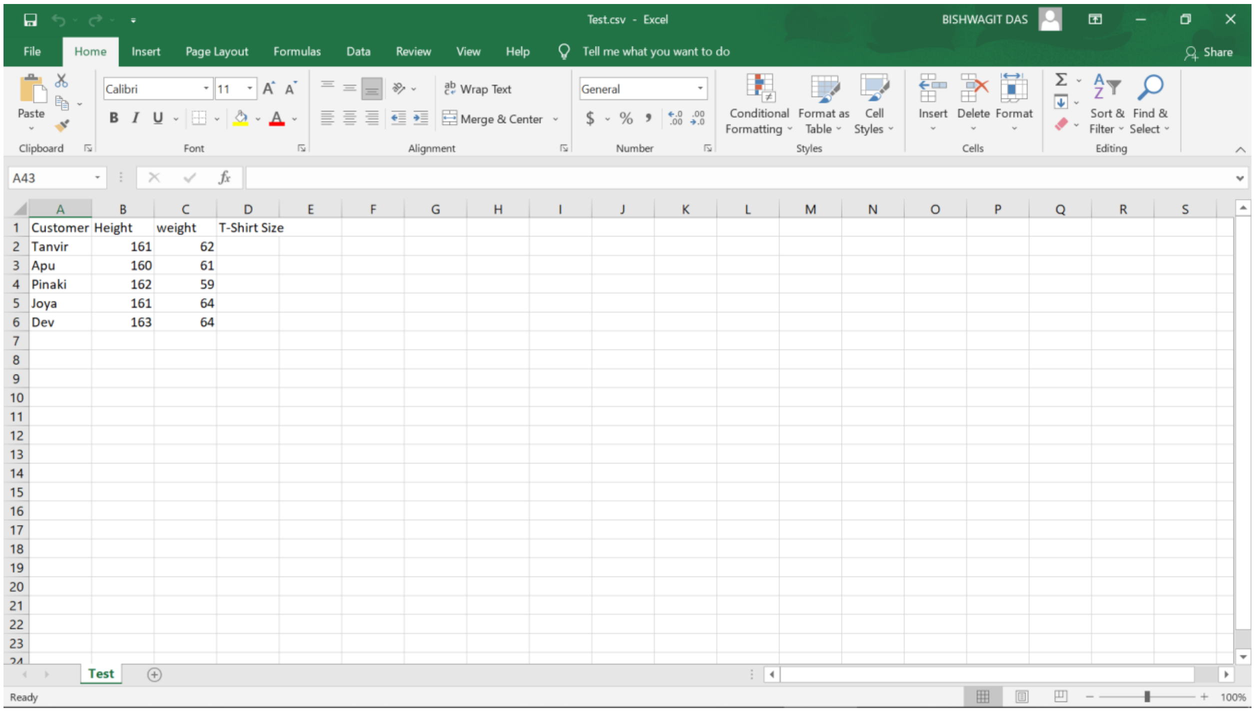
Fig 01: Train Dataset



A1 Customer

	A	B	C	D
21	Dipto	163	64	L
22	Aditto	165	61	L
23	Das	165	62	L
24	Hridoy	160	64	L
25	Bishwagit	163	64	L
26	Shuvra	165	61	L
27	Himi	165	62	L
28	Bijay	160	64	L
29	Sumit	163	64	L
30	Mumu	165	61	L
31	Zim	165	62	L
32	Nirzhar	160	64	L
33	Fahad	163	64	L
34	Anu	165	61	L
35	Hamim	165	62	L
36	Safeya	160	64	L
37	Barna	163	64	L
38	Tapu	165	61	L
39	Tapon	165	62	L

Fig 02: Train Dataset



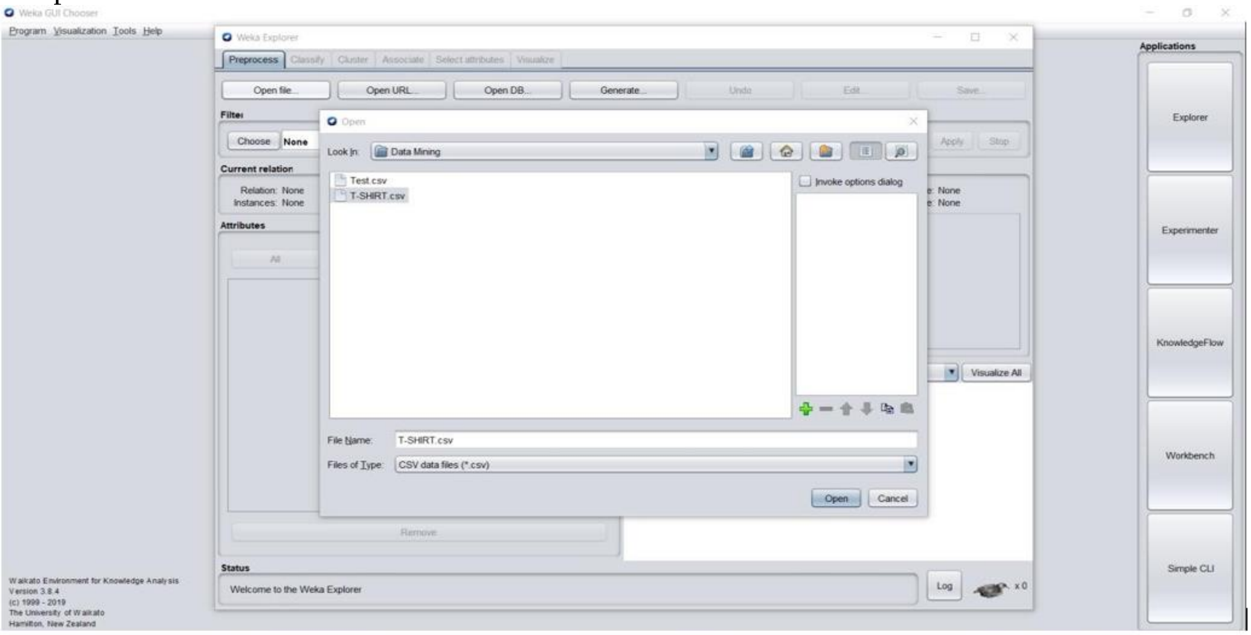
A43

	A	B	C	D
1	Customer	Height	weight	T-Shirt Size
2	Tanvir	161	62	
3	Apu	160	61	
4	Pinaki	162	59	
5	Joya	161	64	
6	Dev	163	64	

Fig 03: Test Dataset

Procedure:

Step 1:



File Name: T-SHIRT.csv

Files of Type: CSV data files (*.csv)

Welcome to the Weka Explorer

Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

Fig 04: Select Train Dataset

First of all, open Weka tools then select the training dataset.

Step 2:

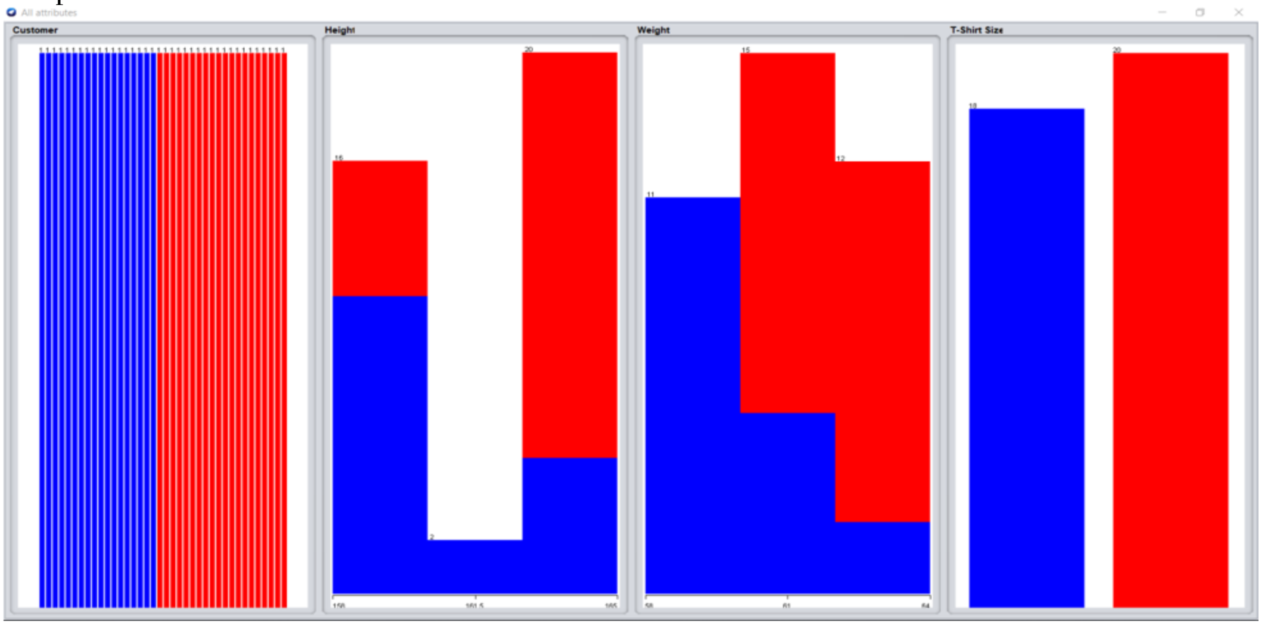
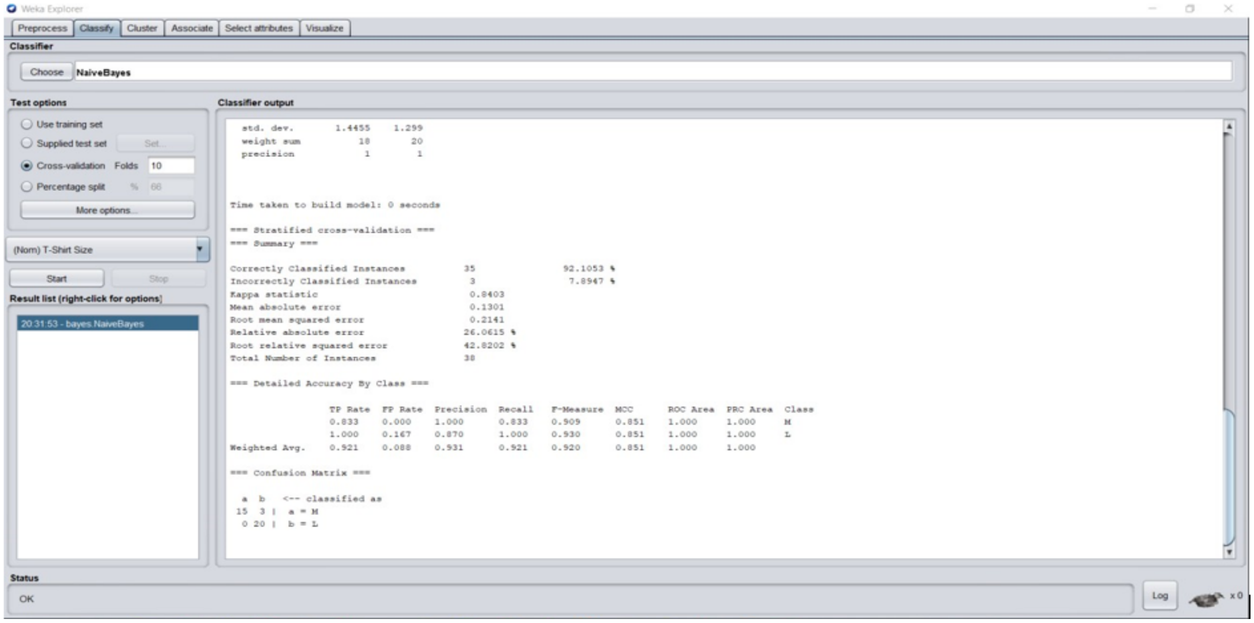


Fig 05: Visualization of all Attributes.

Train dataset select Naïve Bayes classify then show a visualization of all attributes. All attributes are visualized in this graph. There are no missing values, and all characteristics have unique values.

Step 3:



Classifier: NaiveBayes

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) T-Shirt Size

Result list (right-click for options):
20:31:53 - bayes.NaiveBayes

Classifier output:

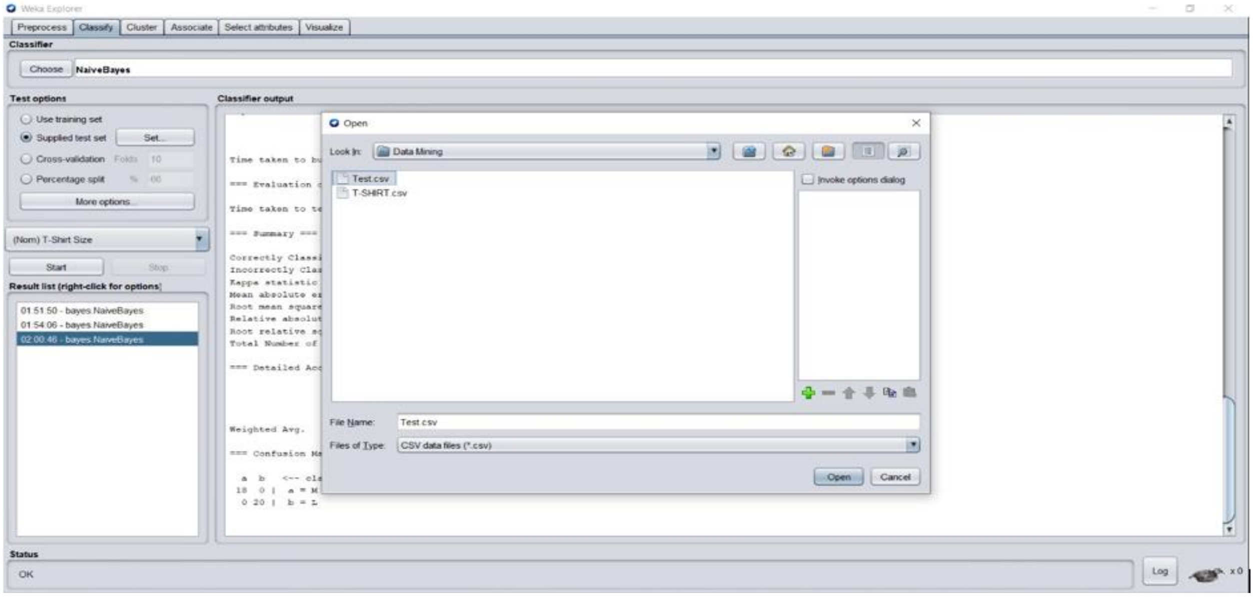
```
  std. dev.       1.4455    1.299
  weight sum          18       20
  precision            1        1



Time taken to build model: 0 seconds

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

Correctly Classified Instances          35               92.1053 %
Incorrectly Classified Instances         3                7.8947 %
Kappa statistic                          0.8403
Mean absolute error                      0.1301
Root mean squared error                  0.2141
Relative absolute error                 26.0615 %
Root relative squared error             42.8202 %
Total Number of Instances               38

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.833    0.000    1.000      0.833    0.909      0.851    1.000     1.000     M
                 1.000    0.167    0.870      1.000    0.930      0.851    1.000     1.000     L
Weighted Avg.    0.921    0.088    0.931      0.921    0.920      0.851    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 15  3 |  a = M
  0 20 |  b = L
```

Status: OK

Fig 06: Train Dataset with Cross-validation.

Cross-validation result for the training dataset. Total occurrences are 38. However, 35 are correctly identified, and 3 are wrongly classified. Accuracy here is 92.1053 percent. The True Positive Rate (TPR) is 0.921, and the False Positive Rate (FPR) is 0.088 on a weighted average basis.

Step 4:



Classifier: NaiveBayes

Test options:
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) T-Shirt Size

Result list (right-click for options):
01:51:50 - bayes.NaiveBayes
01:54:06 - bayes.NaiveBayes
02:00:46 - bayes.NaiveBayes

Classifier output:

```
Time taken to bu

=== Evaluation o

Time taken to te

=== Summary ===

Correctly Classi
Incorrectly Clas
Kappa statistic
Mean absolute er
Root mean square
Relative absolut
Root relative sq
Total Number of

=== Detailed Acc



Weighted Avg.

=== Confusion Ma

  a  b   <-- cla
 18  0 |  a = M
  0 20 |  b = L
```

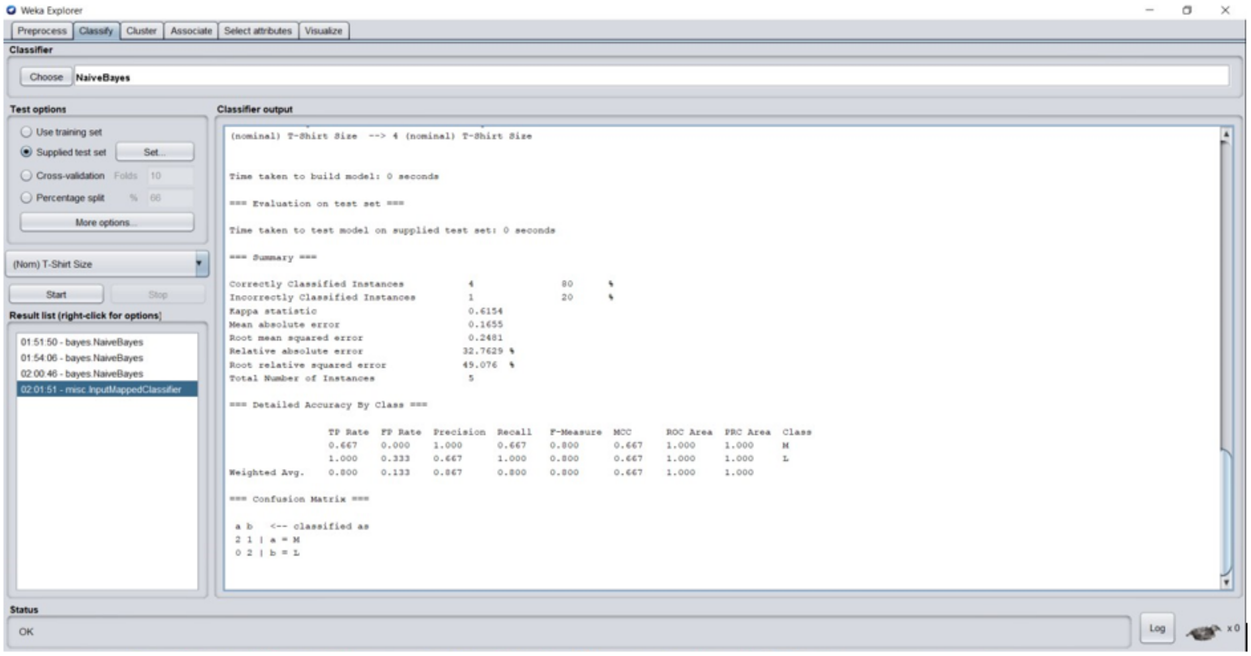
Open
Look In: Data Mining
Test.csv
T-SHIRT.csv
□ Invoke options dialog
File Name: Test.csv
Files of Type: CSV data files (*.csv)

Status: OK

Fig 07: Test Dataset.

Select test dataset to compare train dataset.

Step 5:

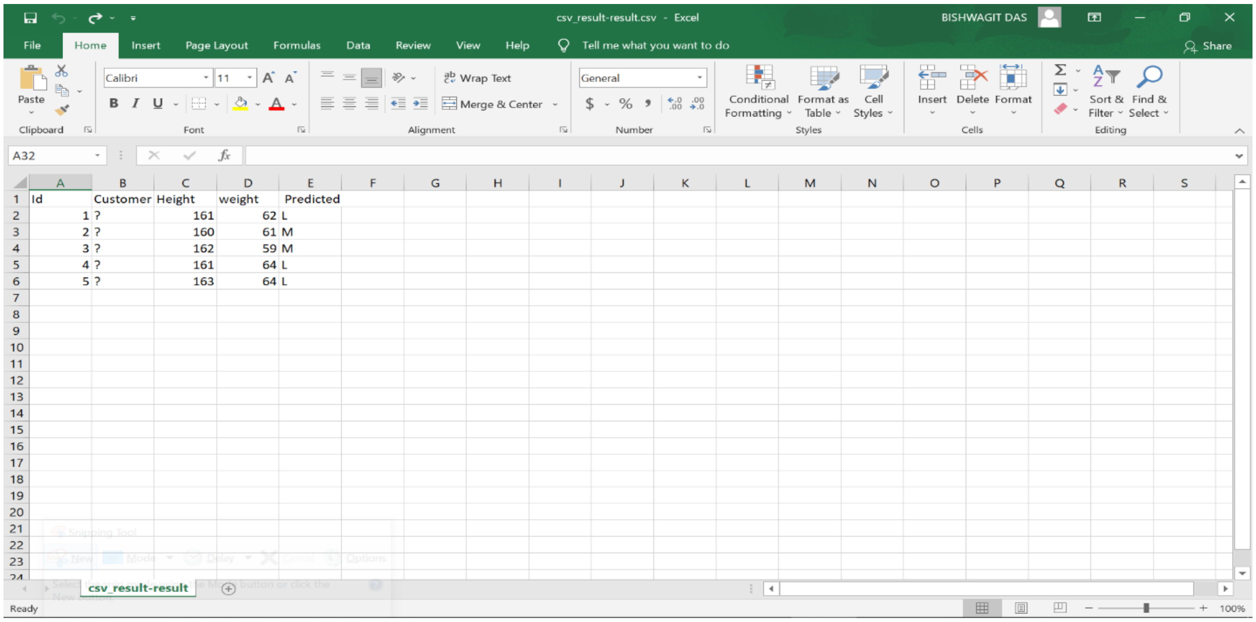


```
(nominal) T-Shirt Size  --> 4 (nominal) T-Shirt Size


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           4               80      %
Incorrectly Classified Instances         1               20      %
Kappa statistic                          0.6154
Mean absolute error                      0.1655
Root mean squared error                  0.2481
Relative absolute error                 32.7629 %
Root relative squared error             49.076  %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.000    1.000      0.667    0.800      0.667    1.000     1.000     M
                 1.000    0.333    0.667      1.000    0.800      0.667    1.000     1.000     L
Weighted Avg.    0.800    0.133    0.867      0.800    0.800      0.667    1.000     1.000

=== Confusion Matrix ===

 a b   <-- classified as
 2 1 | a = M
 0 2 | b = L
```

Fig 08: Test Dataset with Cross-validation.

The outcome of the test dataset. And total occurrences are 5, but 4 are correctly classified, and one is wrongly classified. Accuracy in this area is 80%. The actual Positive Rate (TPR) is 0.800, and the False Positive Rate (FPR) is 0.133 on a weighted average basis. The final output should be saved as a comma-separated values (CSV) file.

Result:



Id	Customer	Height	weight	Predicted
1	?	161	62	L
2	?	160	61	M
3	?	162	59	M
4	?	161	64	L
5	?	163	64	L

Fig 09: Predicted result

This figure shows how the classified match T-shirt size for each customer based on their height and weight.

Conclusion:

Because of its simplicity, elegance, also robustness, the Naive Bayes model is exceptionally appealing. It is one of the earliest formal classification methods, and even in its most basic form, it is often highly effective. It's commonly used in text classification and spam filtering, for example. This classifier predicts T-Shirt size with a near-perfect accuracy rate.