

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:**

### Attributes Information:

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.

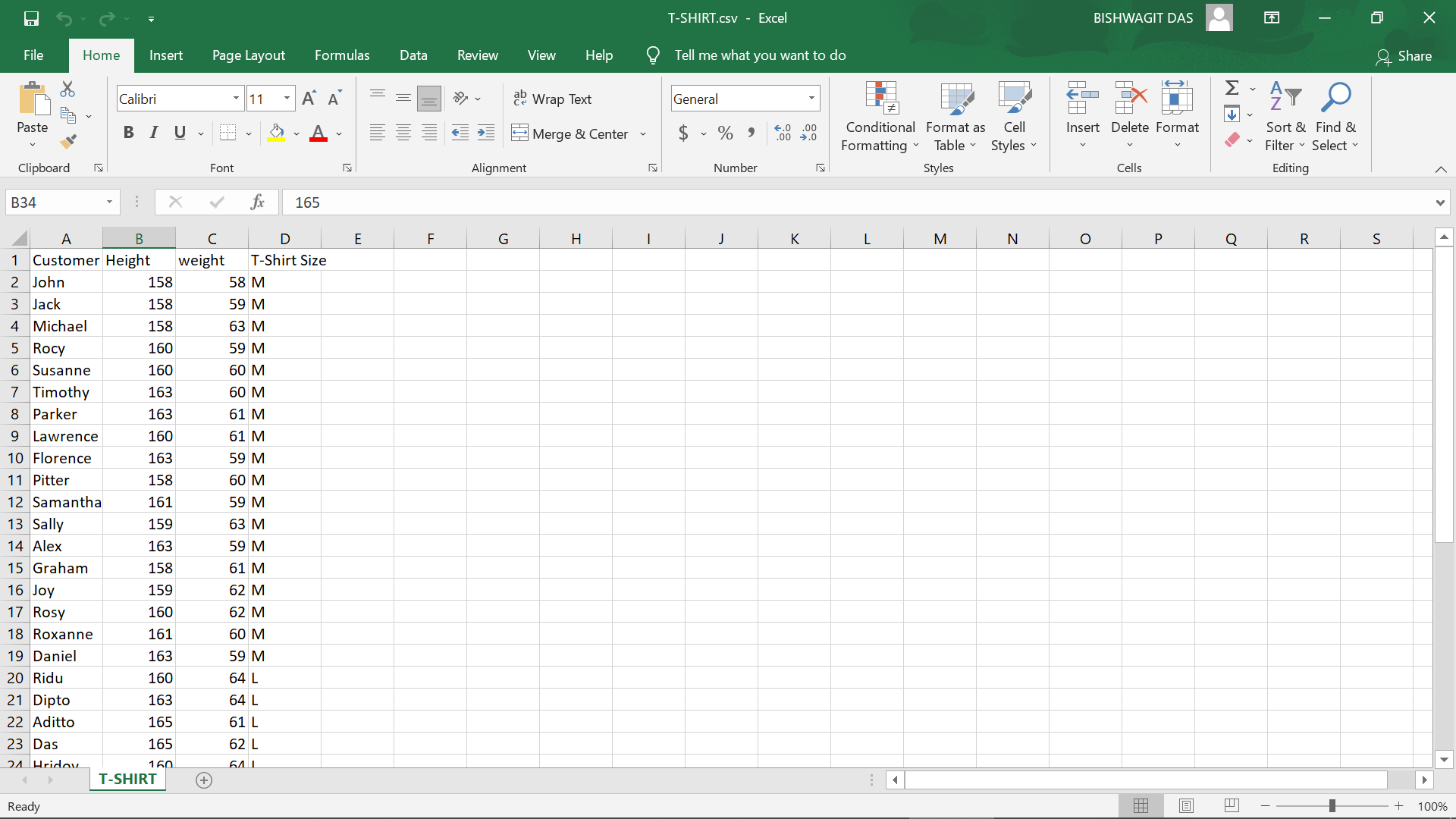


Fig 01: Train Dataset

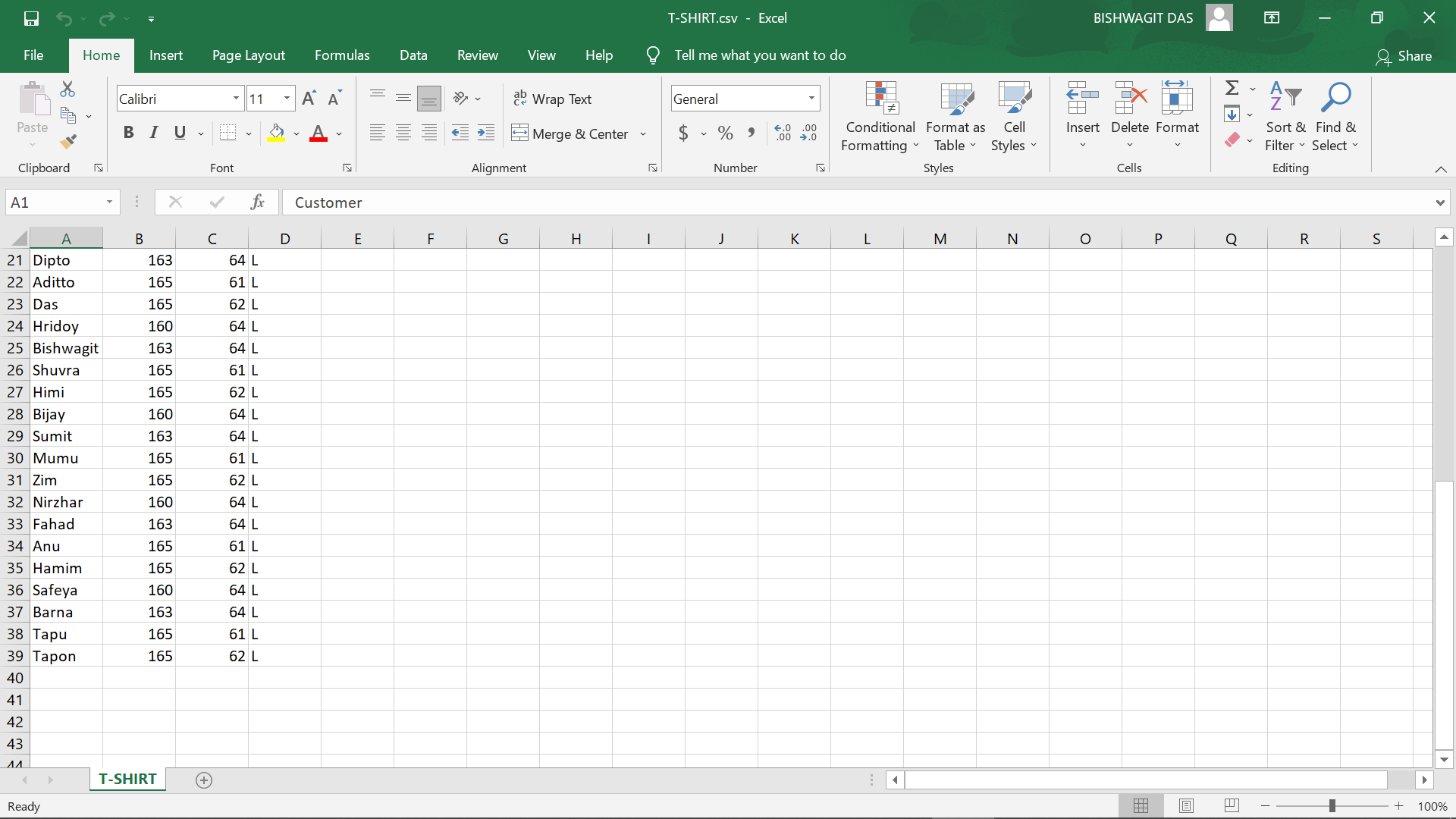


Fig 02: Train Dataset

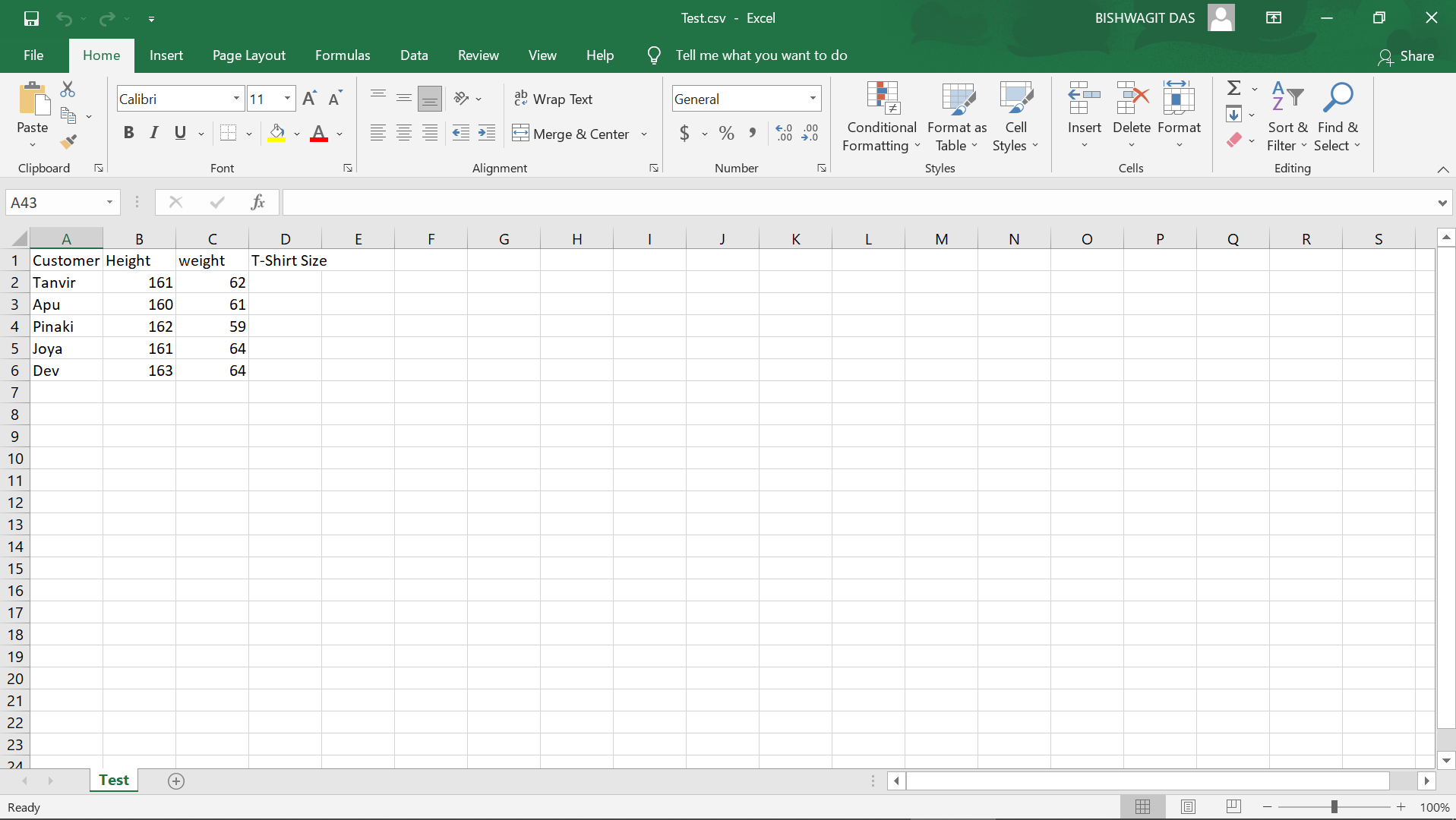
****

Fig 03: Test Dataset

**Procedure:**

Step 1:

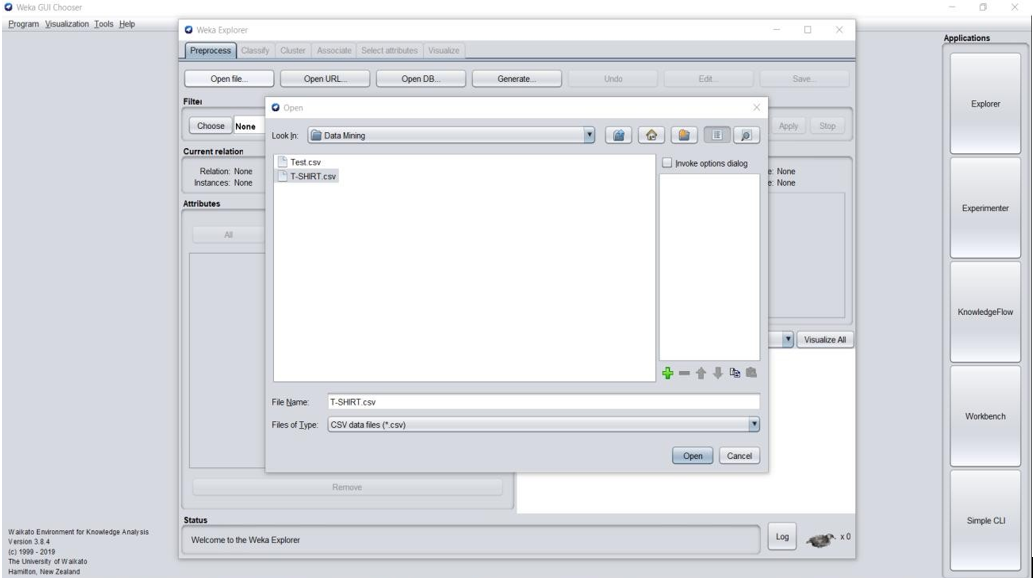
****

Fig 04: Select Train Dataset

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

Step 2:

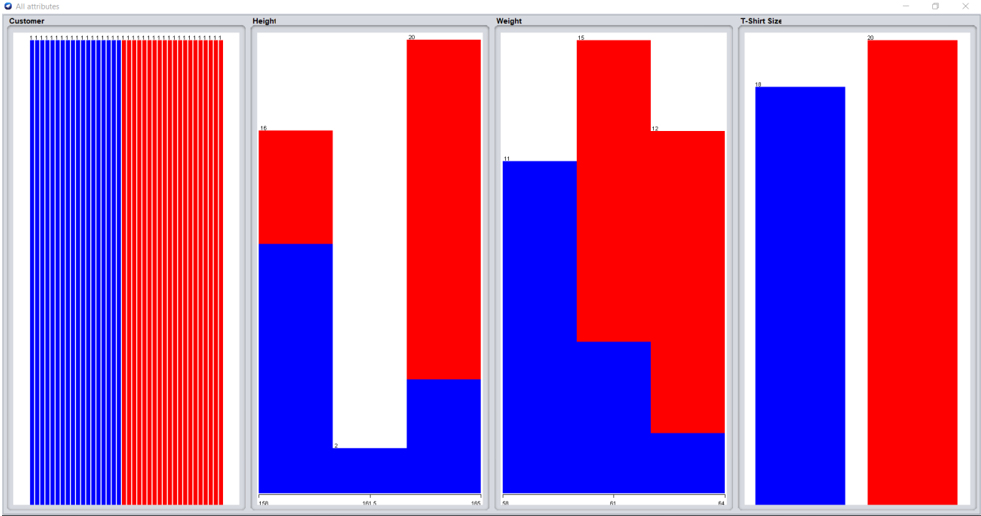
****

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:

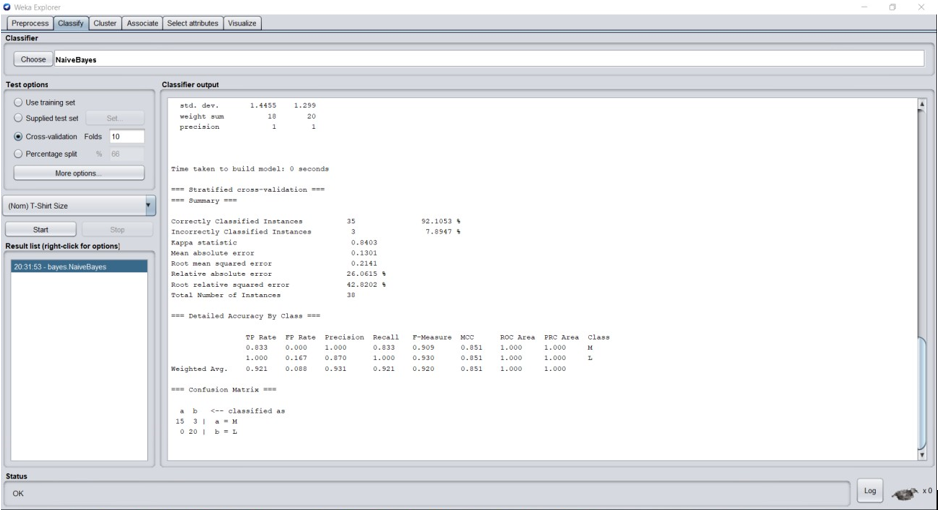


Fig 06: Train Dataset with Cross-validation.

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:

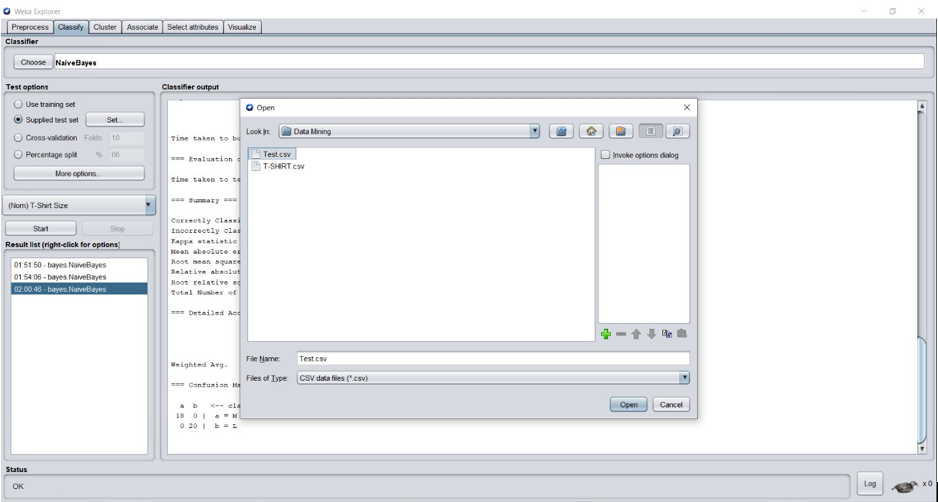


Fig 07: Test Dataset.

Select test dataset to compare train dataset.

Step 5:

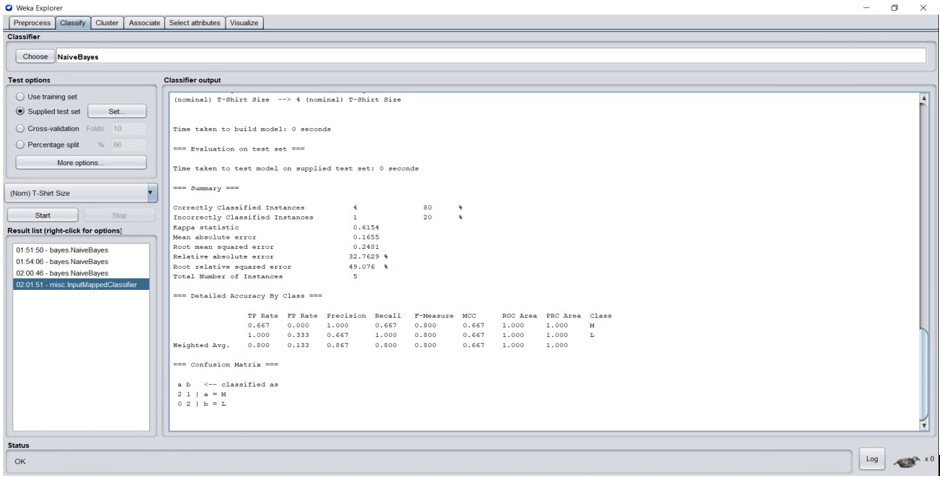


Fig 08: Test Dataset with Cross-validation.

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:

### 

### 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.