Experiment No - 01

Author Name : Vinni Fengade
Roll No. : 67
Sem & Sec : 7 CSE [B]

Date Compiled : 24-Jul-2023
File Name : Prac1.txt

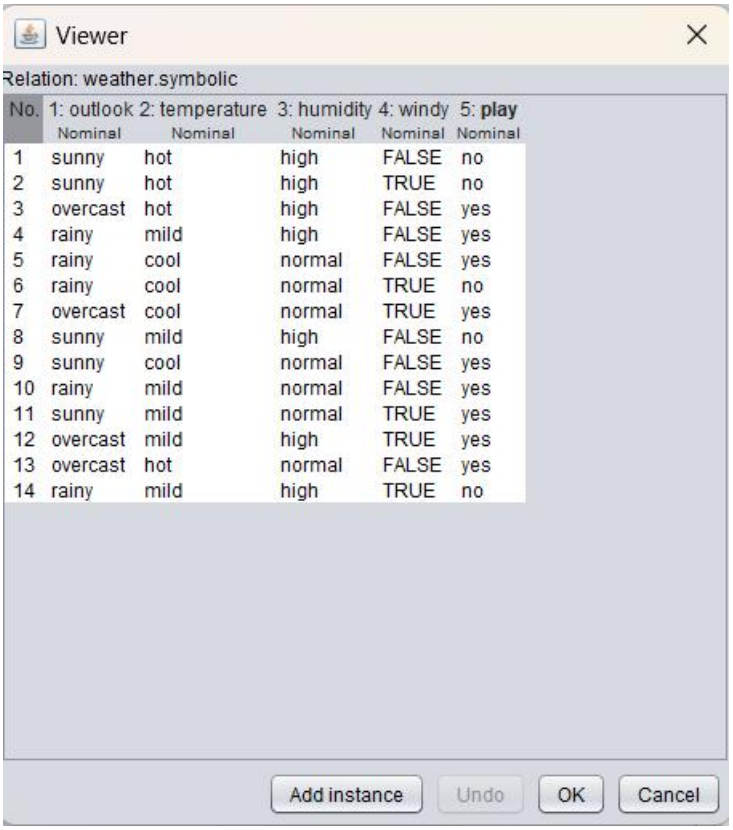
AIM: (A) Introduction to Weka tool.
(B) Performing data understanding and preprocessing on the given data set in Weka.

EXECUTION

A. Press the Explorer button on the main panel and load the weather dataset and answer the
following questions

1. How many instances are there in the dataset?

=> there are 14 instances in dataset



2.State the names of the attributes along with their types and values.

Outlook:nominal,Temperature:nominal,Humidity:Nominal,Windy:nominal, Play:nominal.

3.What is the class attribute?

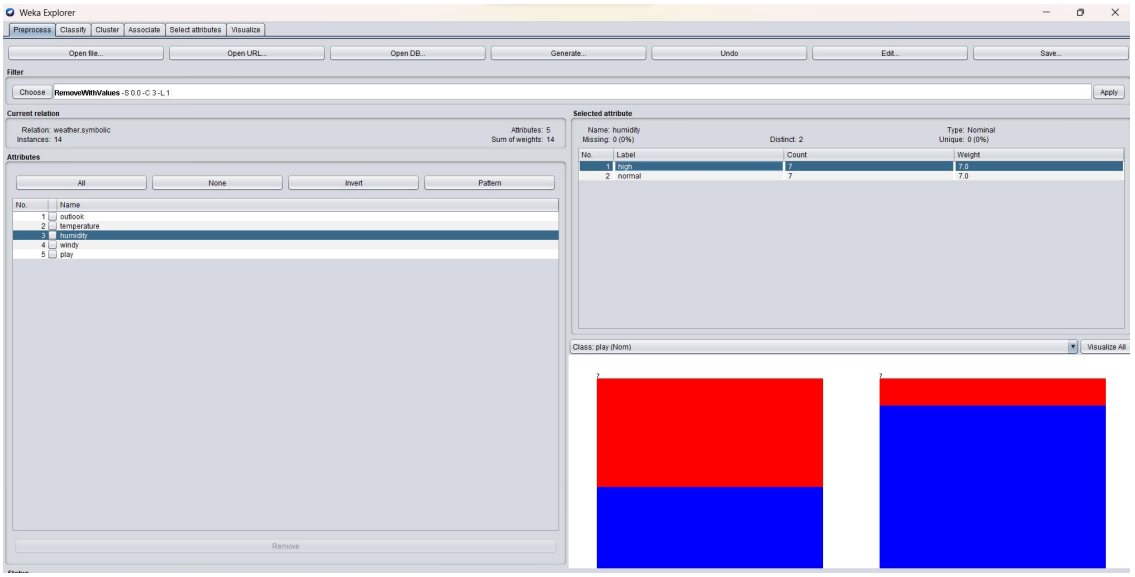
The class attribute is Play which can take 2 values Yes , No

4. How will you determine how many instances of each class are present in the data

By selecting the attribute from Atrribute list and filtering for which value of attribute will have yes and no

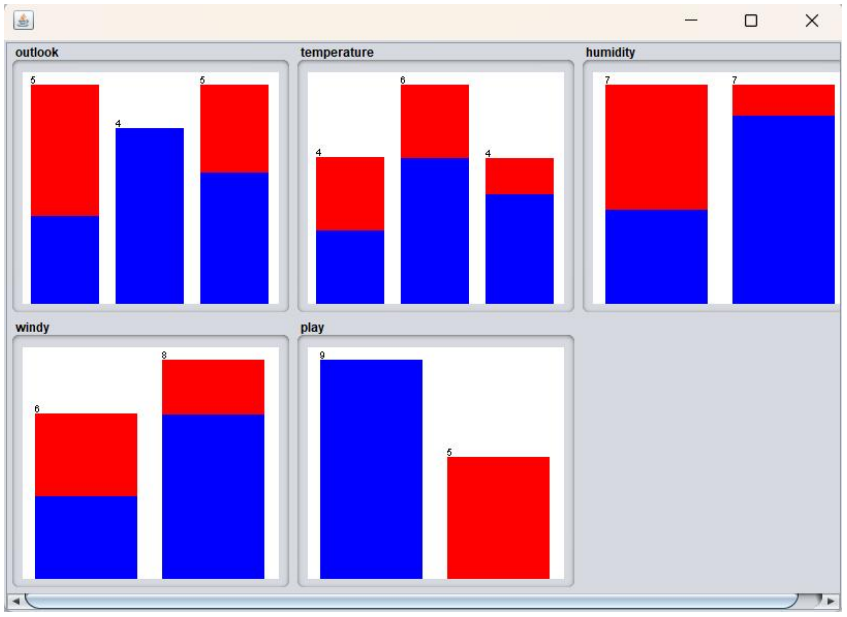
Like for example here yes-blue,no-red

So in case of humidity: for high humidity we have 7 instances in which 3 are yes and 4 are no
Normal humidity we have 7 instances in which 1 is no 6 is yes



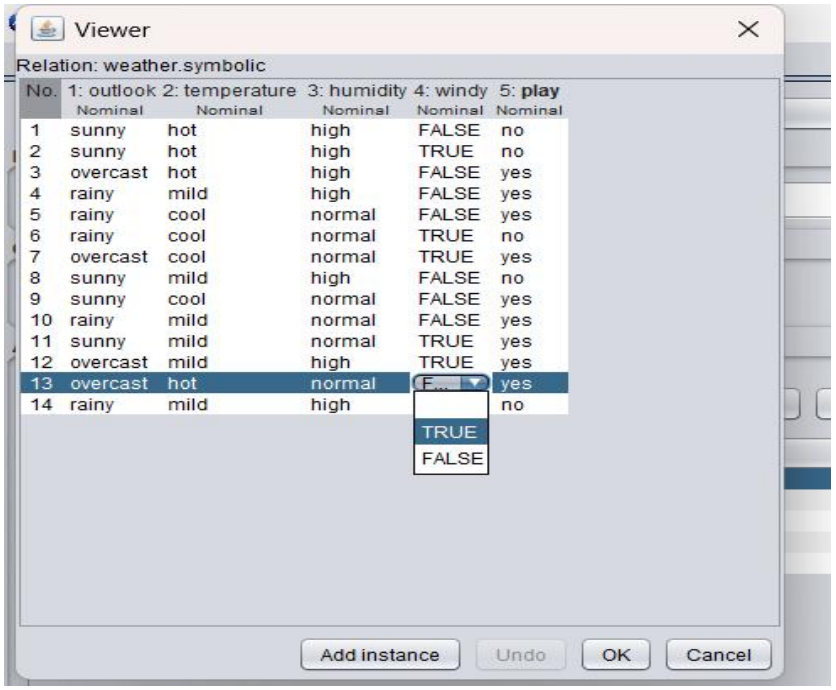
5. What happens with the Visualize All button is pressed?

It will give the all the different classification of Play attribute on basis of different classes

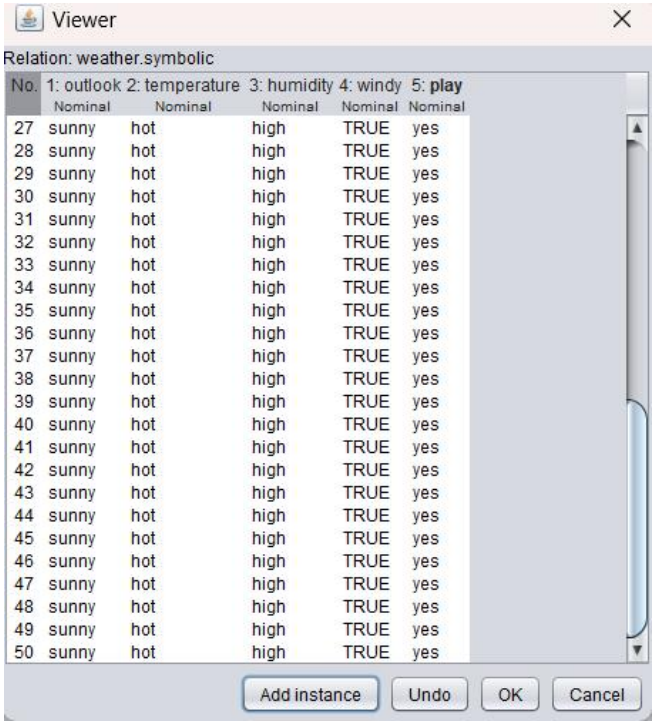


6. How will you view the instances in the dataset? How will you save the changes?

Click on edit button and then change any attribute you want and click on ok the changes will be saved:

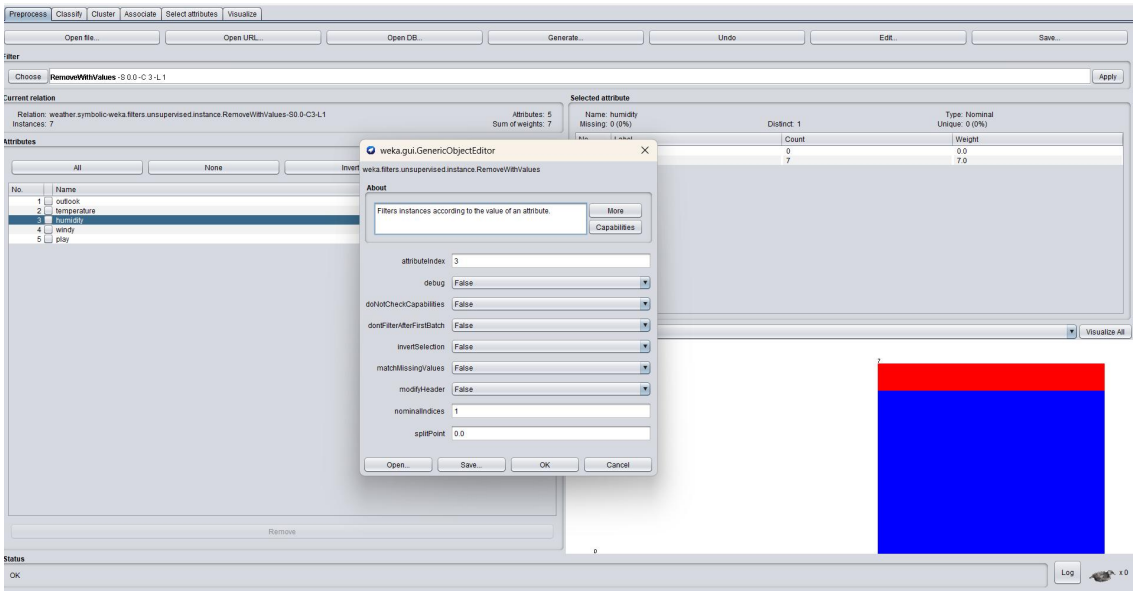


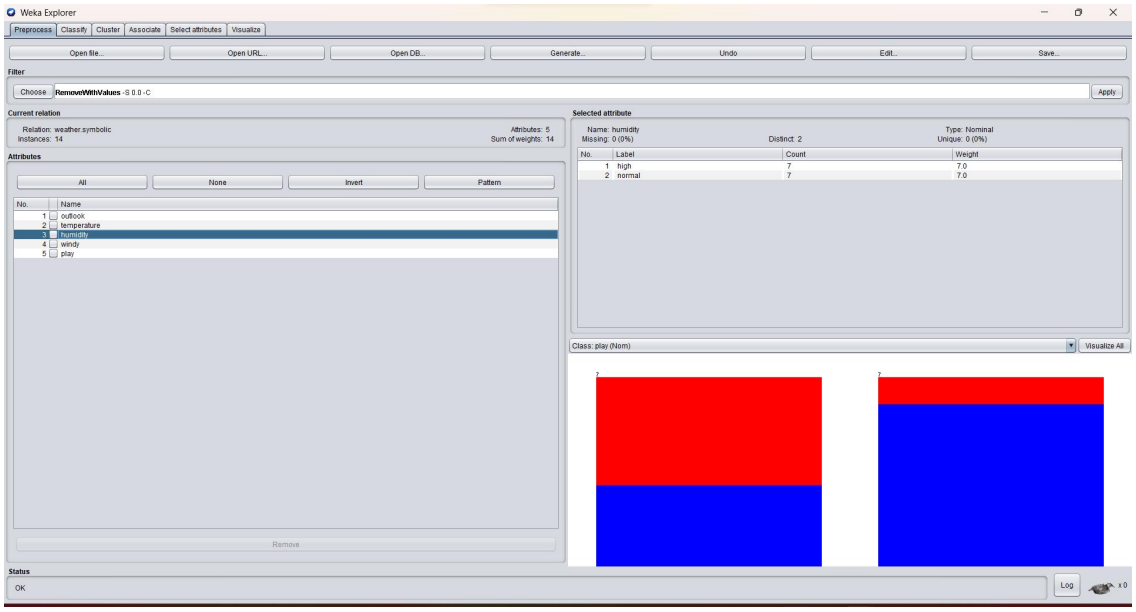
7. Now, extend the dataset to include 50 instances in total.



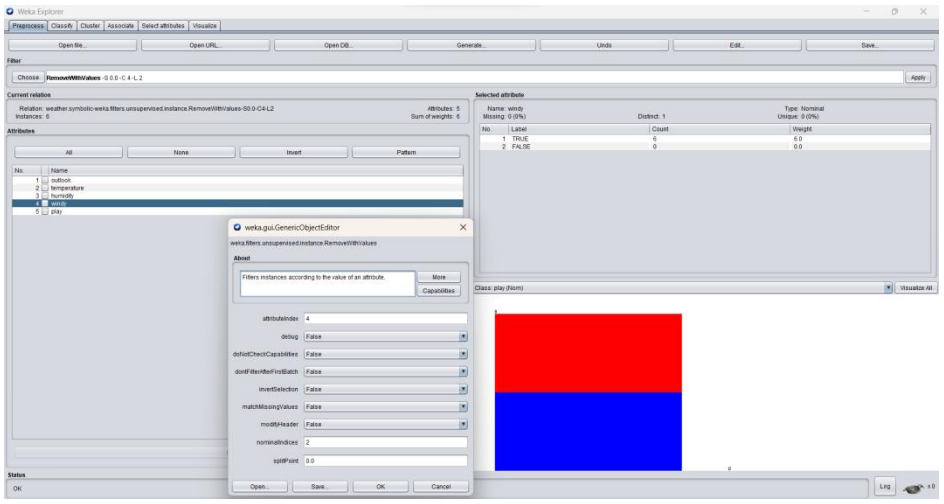
2. Do as directed to apply Filter

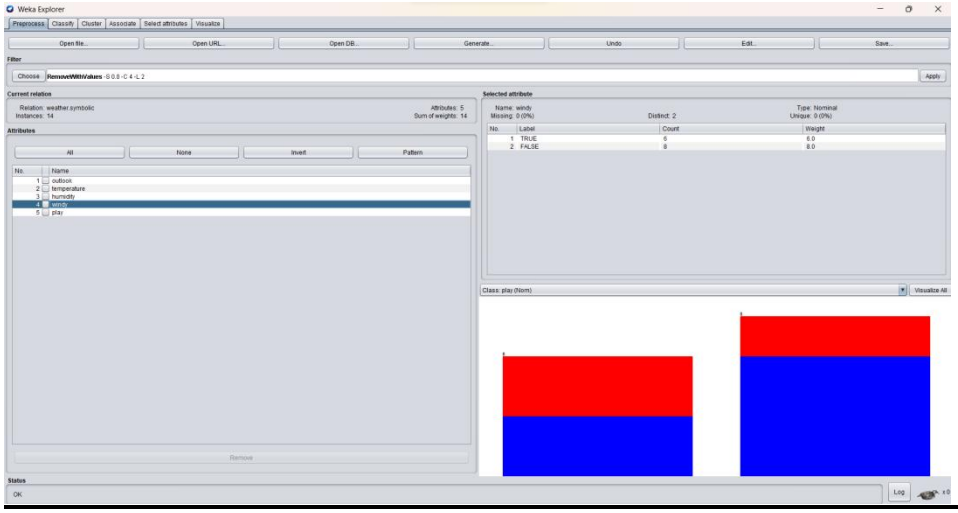
a) Use the unsupervised filter RemoveWithValues to remove all instances where the attribute ‘humidity’ has the value ‘high’? Undo the effect of the filter.



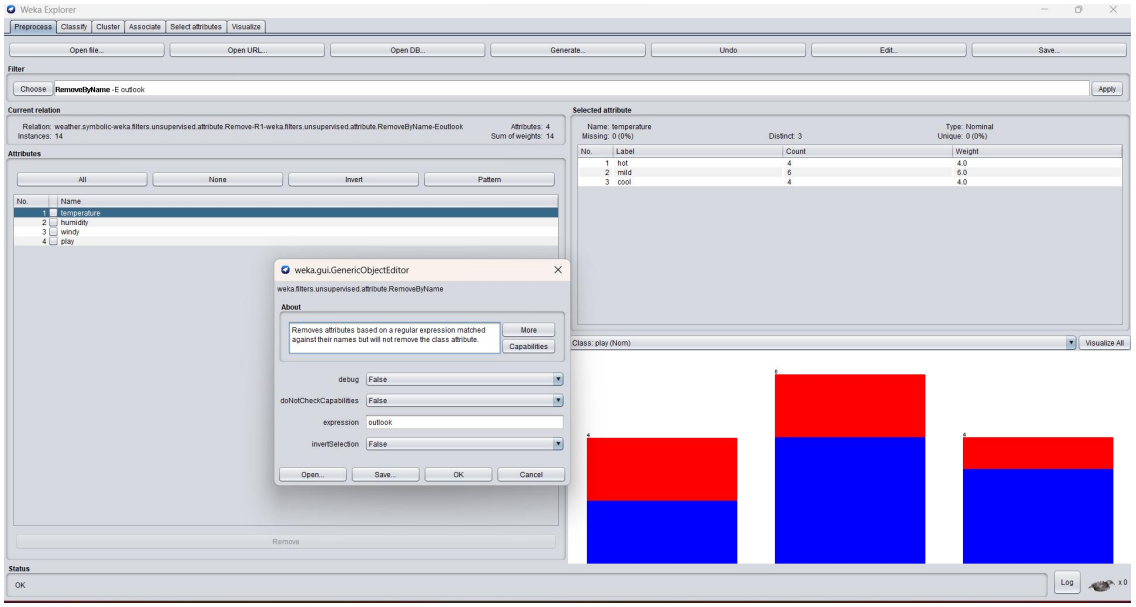


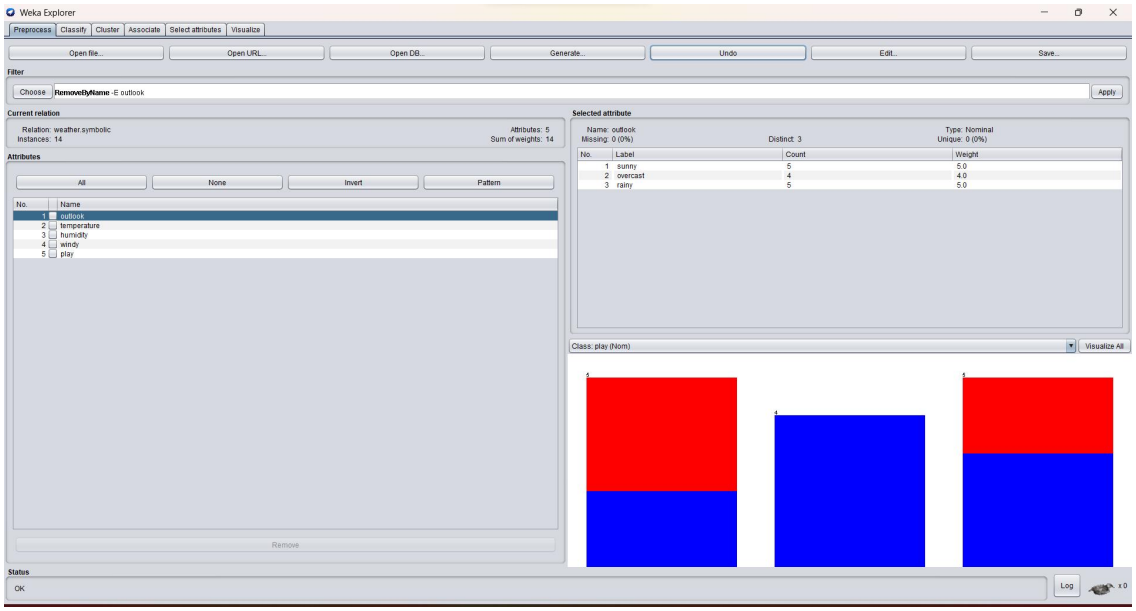
b) Remove the ‘FALSE’ instances of windy attribute and undo the effect.





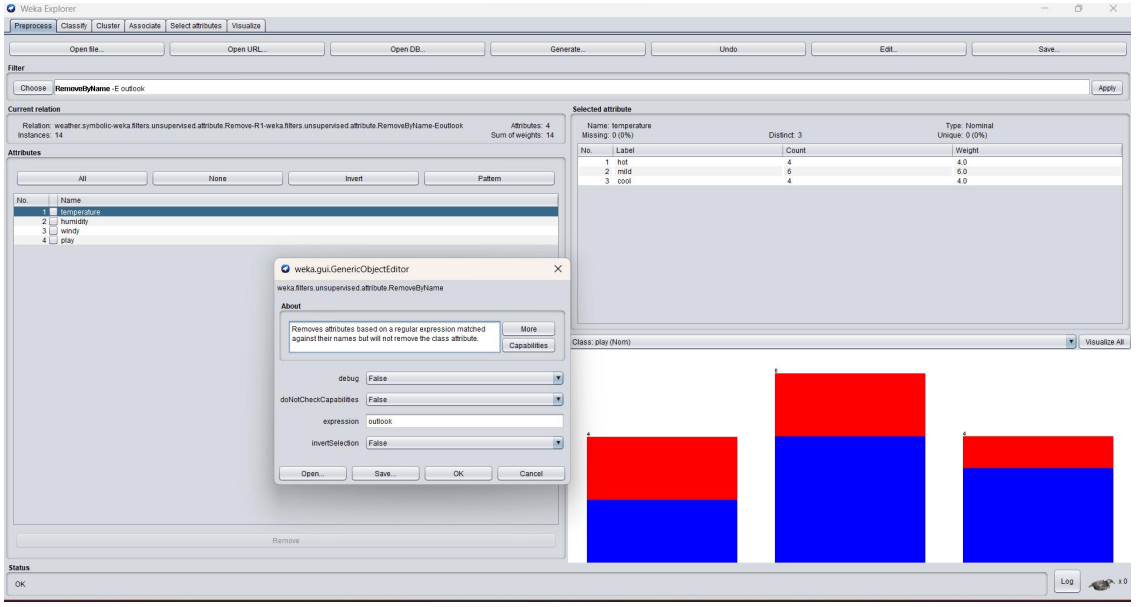
c. Remove the attribute outlook and undo the effect.





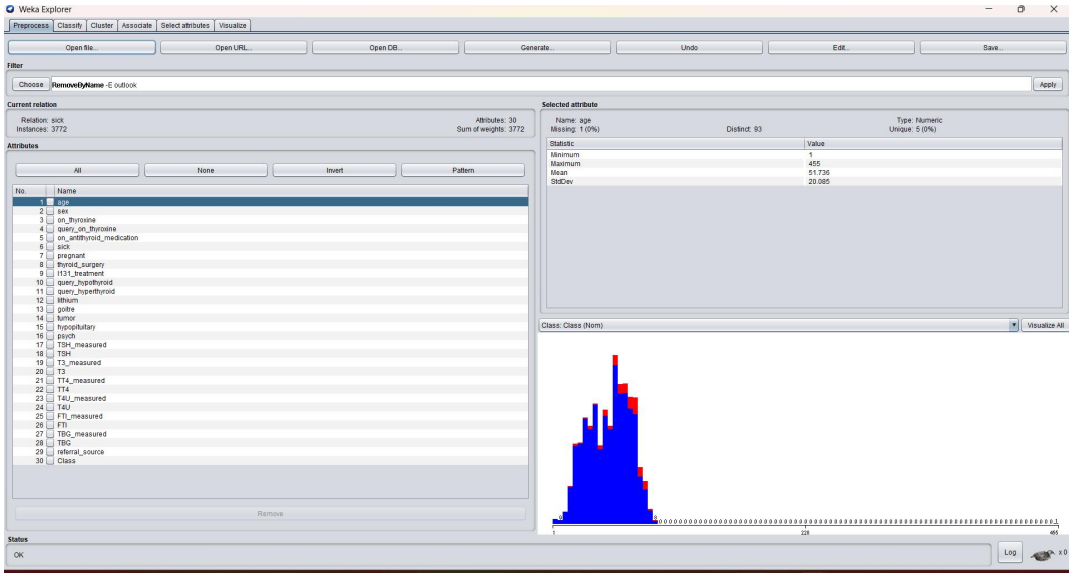
d. Experiment with different filters and report their effects.

The different filters have different effect which can be used to classify the data like withhelp of removebyvalue filter we can remove the instance according to its value and by using remove we can remove the attribute from the table for classification.

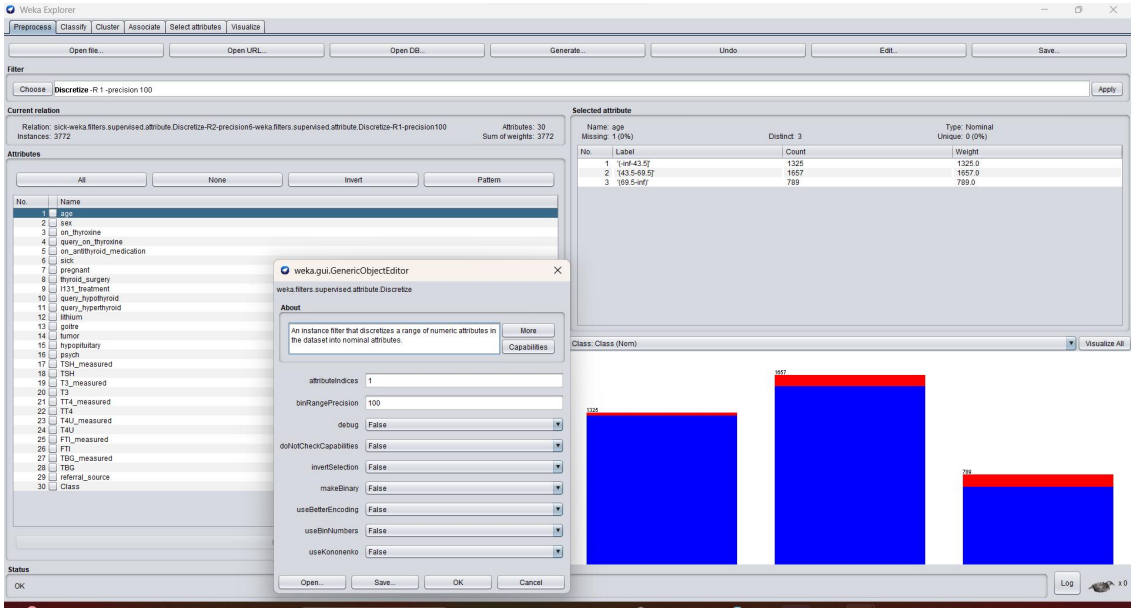


3.Application of Discretization Filters [use sick.arff dataset]

a. Load the 'sick.arff' dataset.

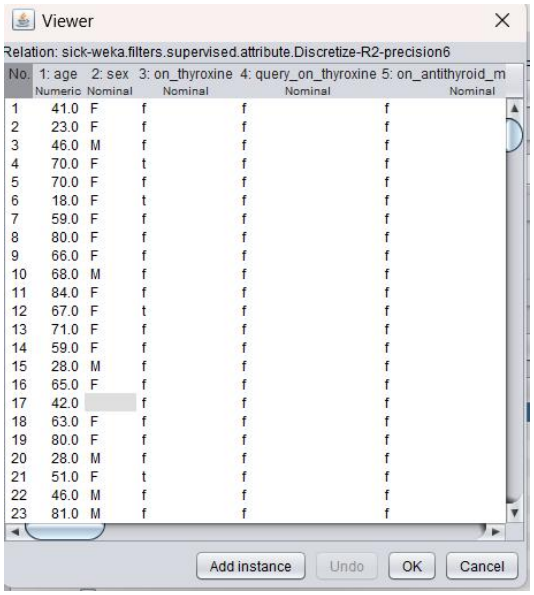


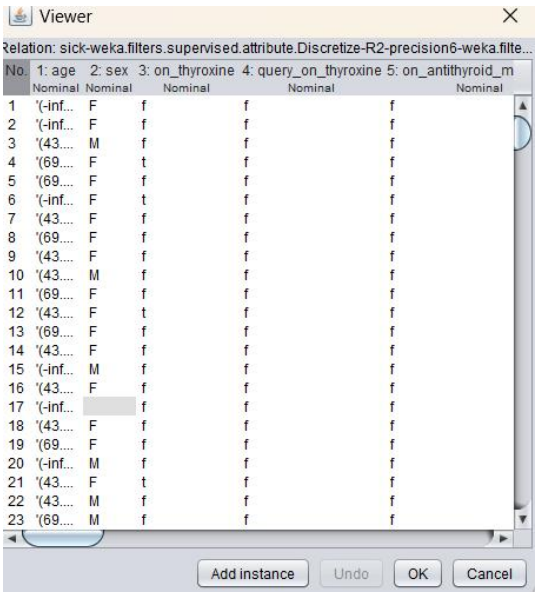
b. Apply the supervised discretization filter on different attributes.



C. What is the effect of this filter on the attributes?

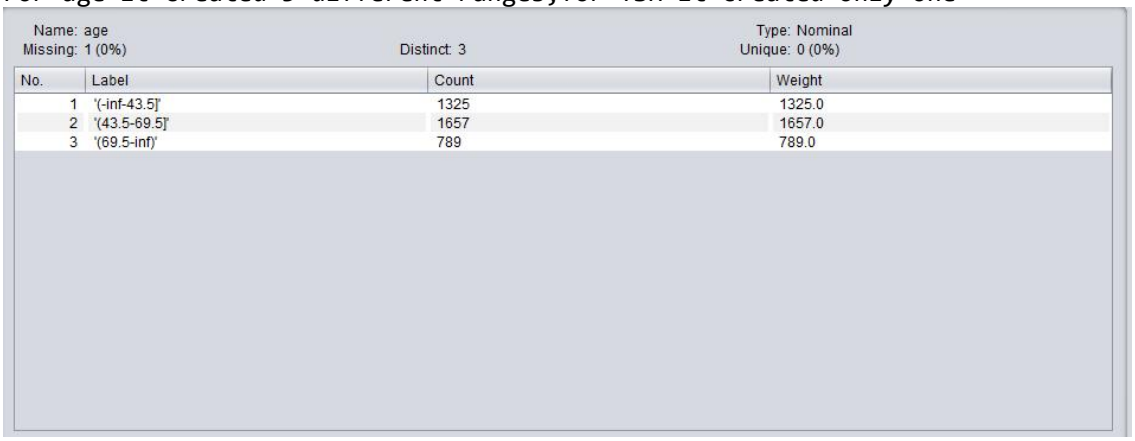
The numeric attribute is converted to nominal attribute after placing a filter on it



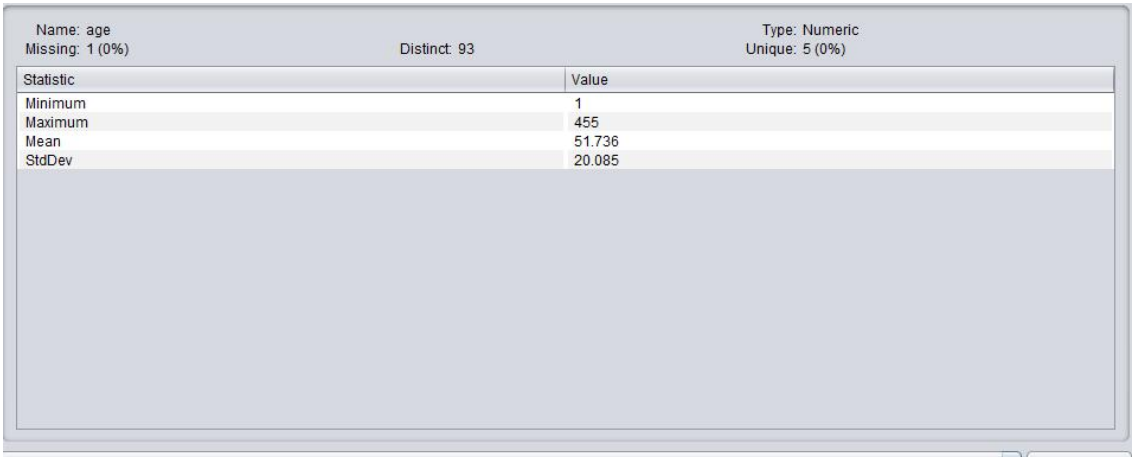


d. How many distinct ranges have been created for each attribute?

For age it created 3 different ranges,for TSH it created only one

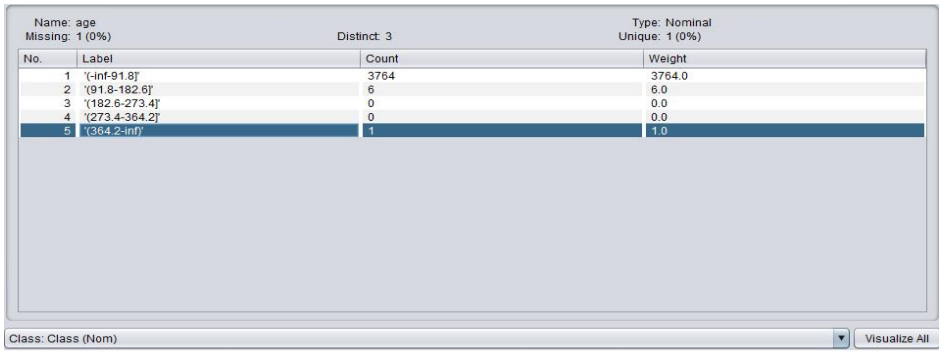


e. Undo the filter applied in the previous step.

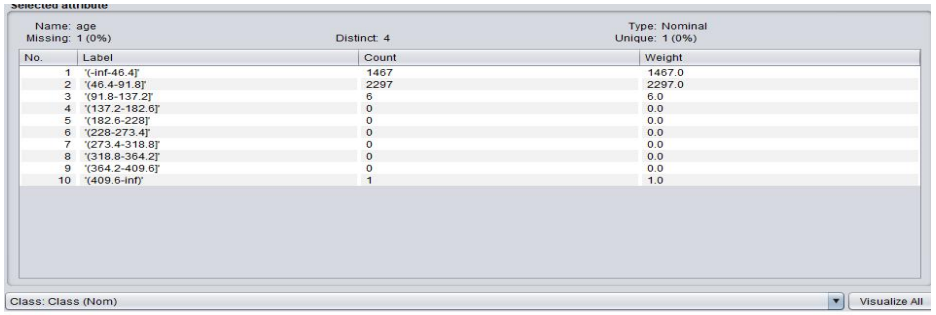


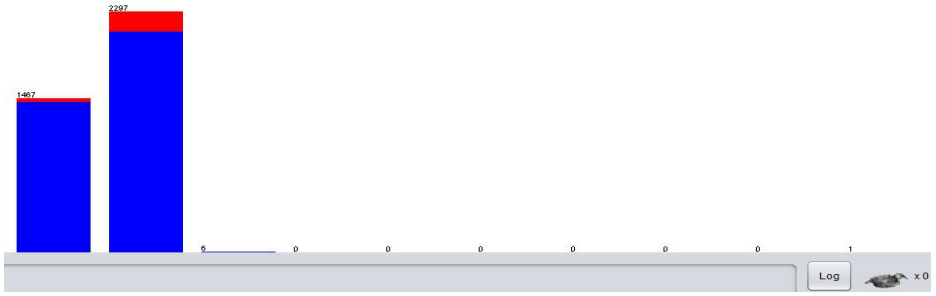
f. Apply the unsupervised discretization filter. [Use equal-width binning approach]
1. In this step, set 'bins'=5
2. In this step, set 'bins'=10
3. What is the effect of the unsupervised filter on the dataset?

Bins=5



For bin=10

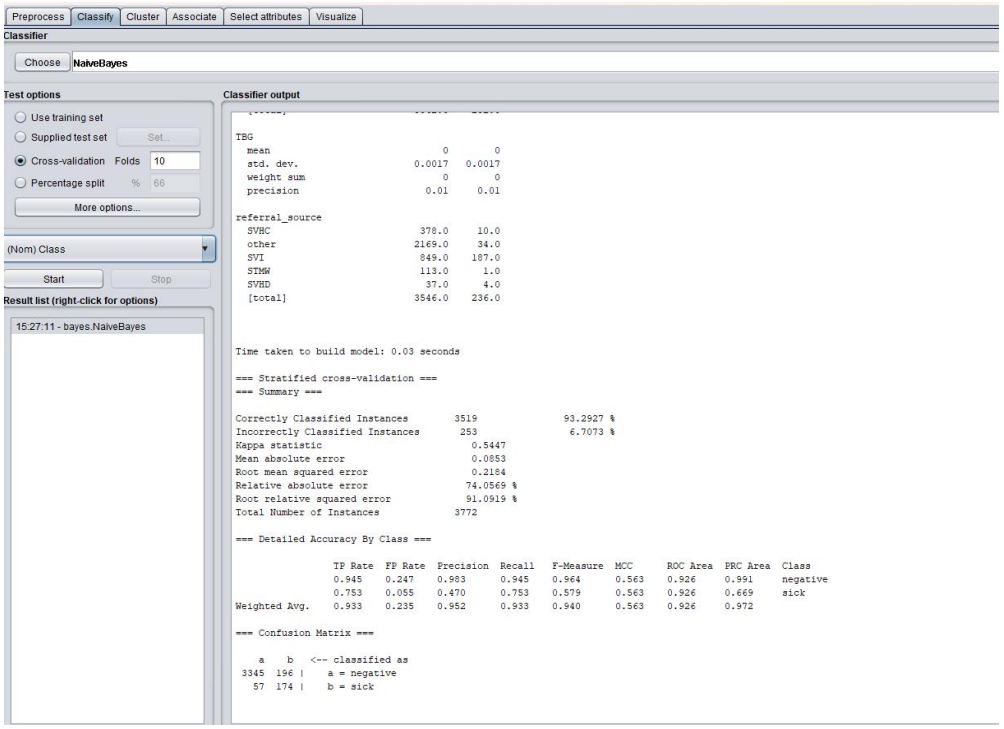




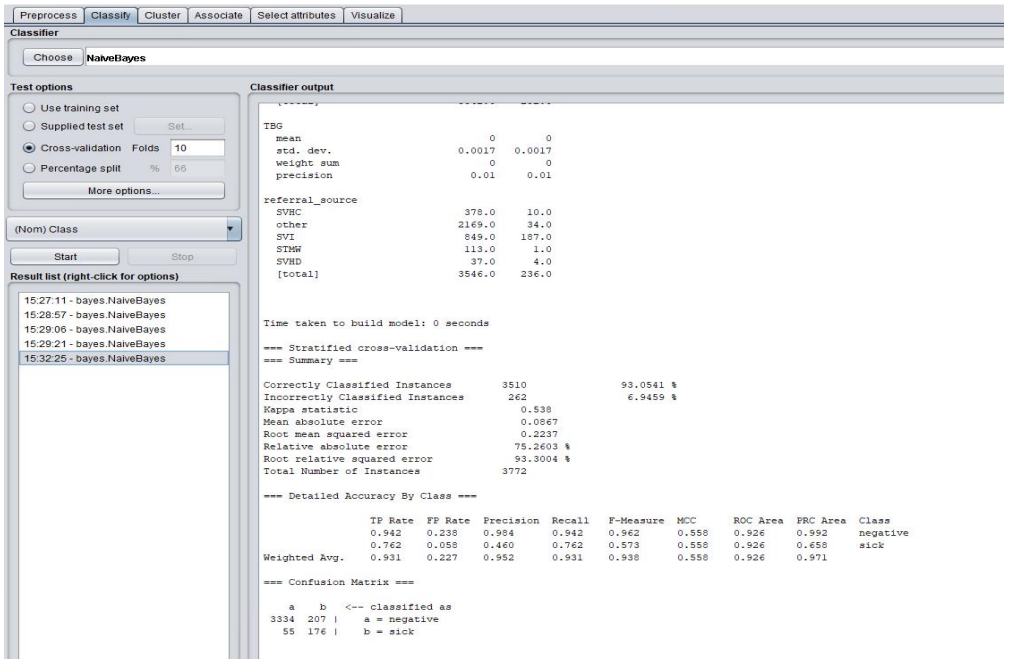
By using the unsupervised discretization filter we were able to convert the numeric data into nominal data for age etc. Also it help us to get the data in format of ranges as it will be better to form a conclusion out of the data as the ranged data is better in term for parameters like age.

g. Run the the Naive Bayes classifier after apply the following filters
1. Unsupervised discretized with 'bins'=5
2. Unsupervised discretized with 'bins'=10
3. Unsupervised discretized with 'bins''=20.

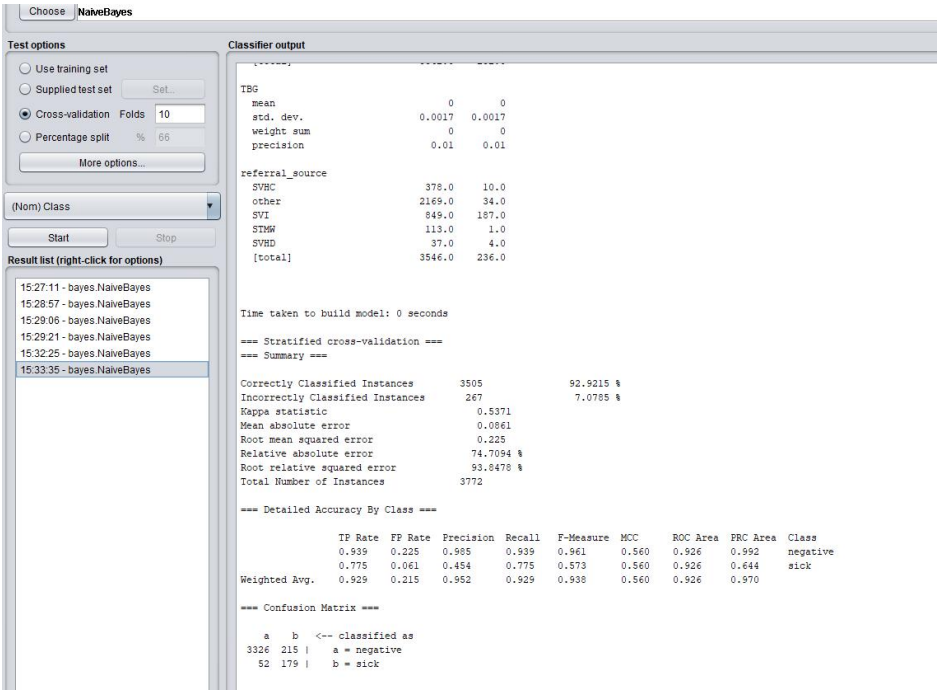
Bin =5



Bin=10



Bin =20

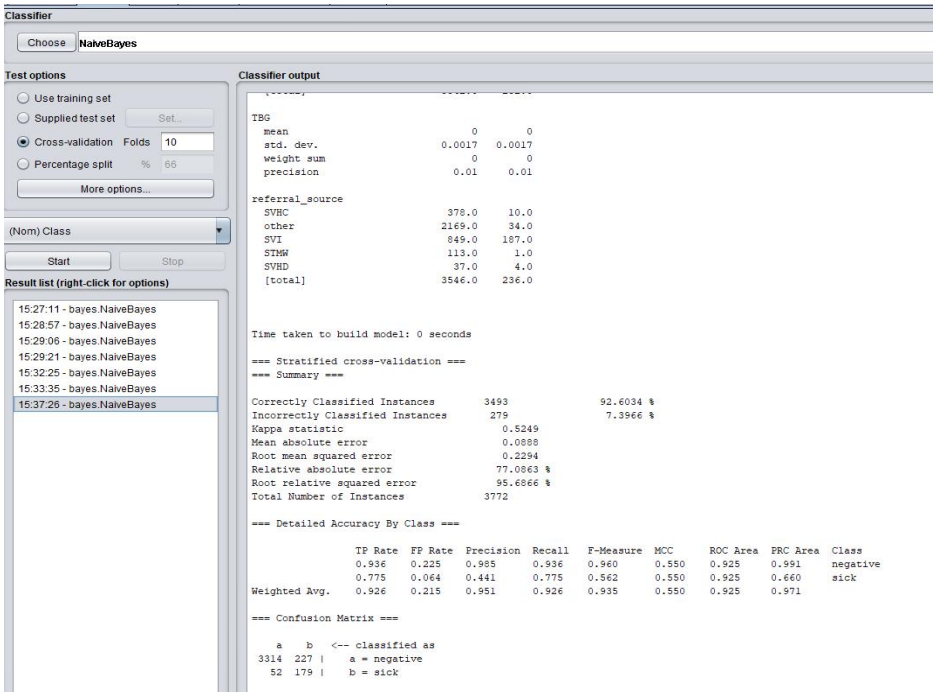


Here after changing the size of bn the data classification of data become different so the values in confusion matrix also changes as the size changed

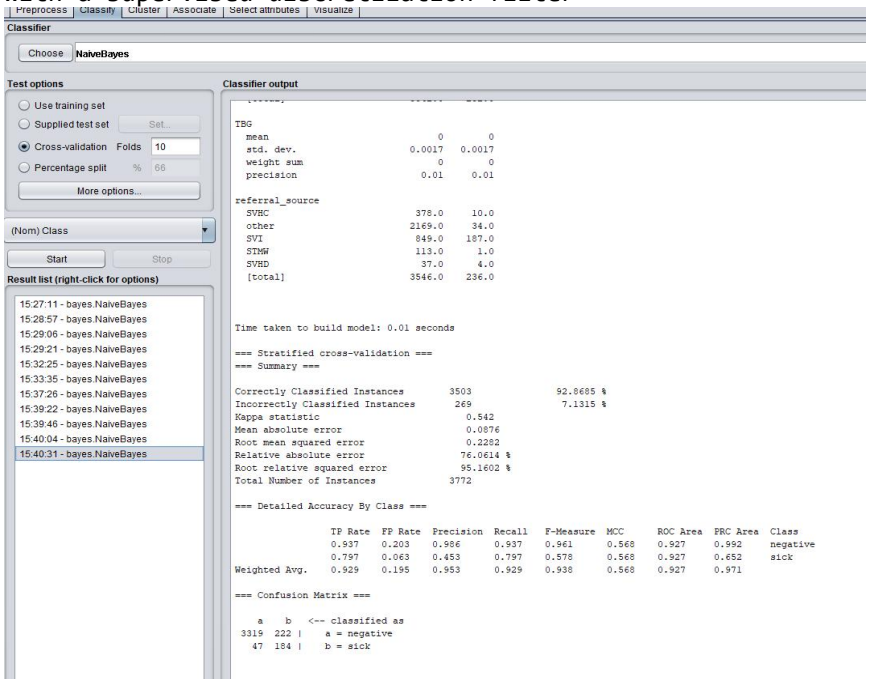
And the results were better as the no of bin were increasing to some extent

h. Compare the accuracy of the following cases
1. Naive Bayes without discretization filters
2. Naive Bayes with a supervised discretization filter
3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins attributes.

Without filter:



with a supervised discretization filter

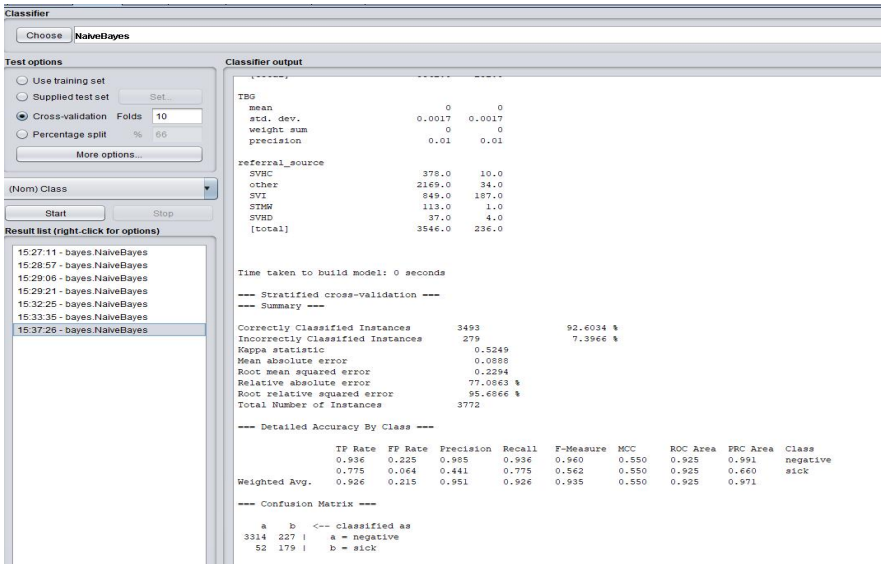


As we can see the value of class with respect to filter and non filtered data is different as is better for filttered data as it is easy to predict the reason more.

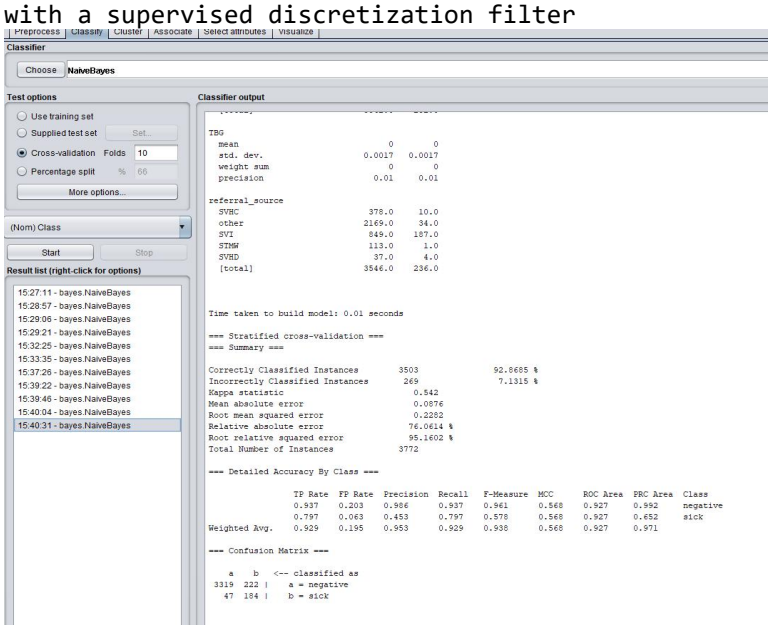
---

i. Repeat steps 6 to 8 using equal-frequency binning approach and present your conclusion.

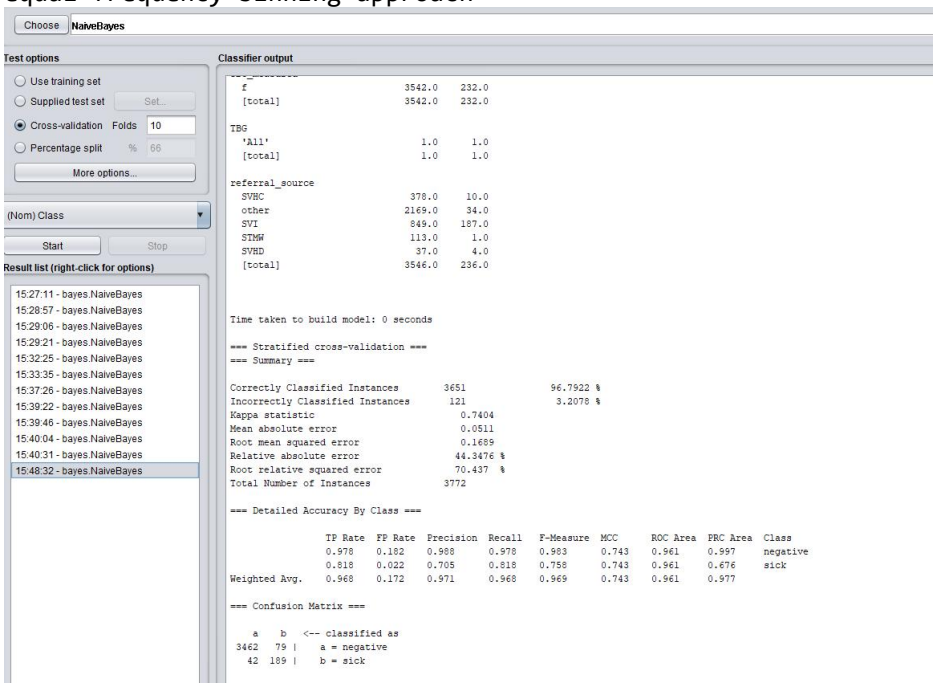
Without filter:



with a supervised discretization filter



equal-frequency binning approach



As compared to the non filtered data and supervised descritized filter the unsupervised descritized filter has better classification as only 3.6% of cases are wrongly classified which is half of supervised learning also the accuracy is incread by a large gap