Step 1: Download and installation Weka

i) Download WEKA from here: https://sourceforge.net/projects/weka/
ii) Click on the downloaded file.
iii) Install as it is by clicking next.
iv) Tutorial video link: https://www.youtube.com/watch?v=vERzuq5SawI

Here I've already downloaded and installed WEKA.

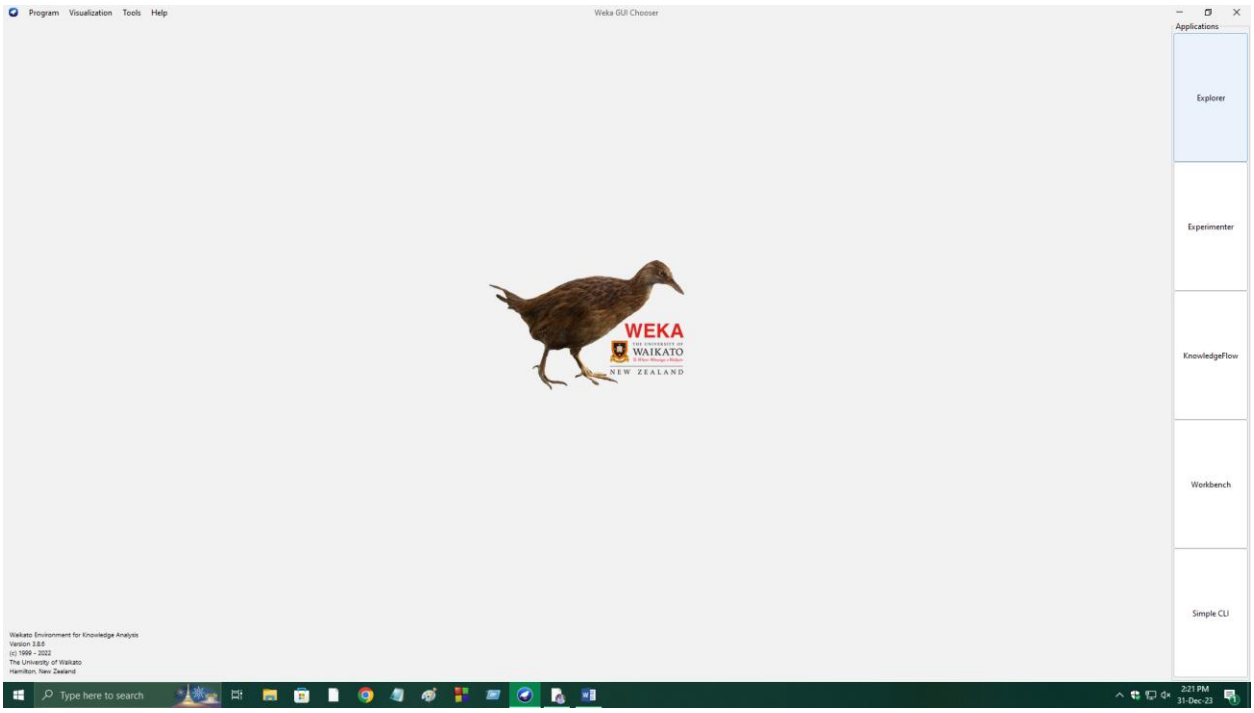
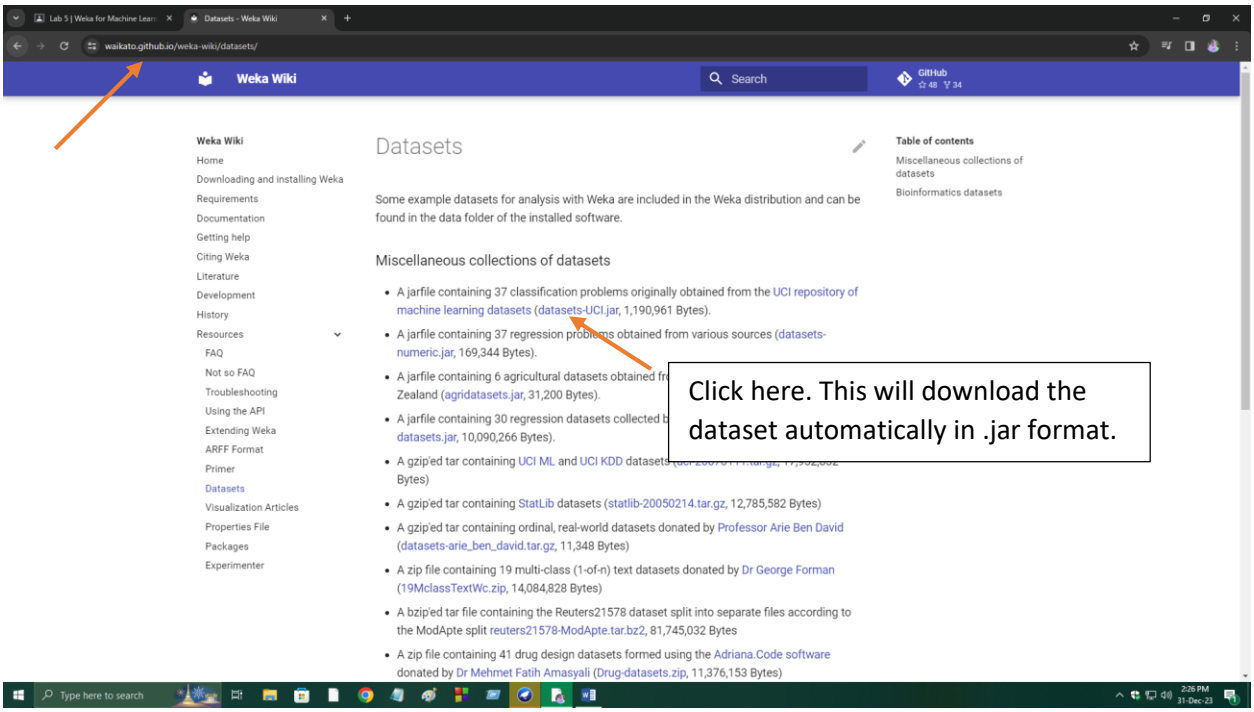
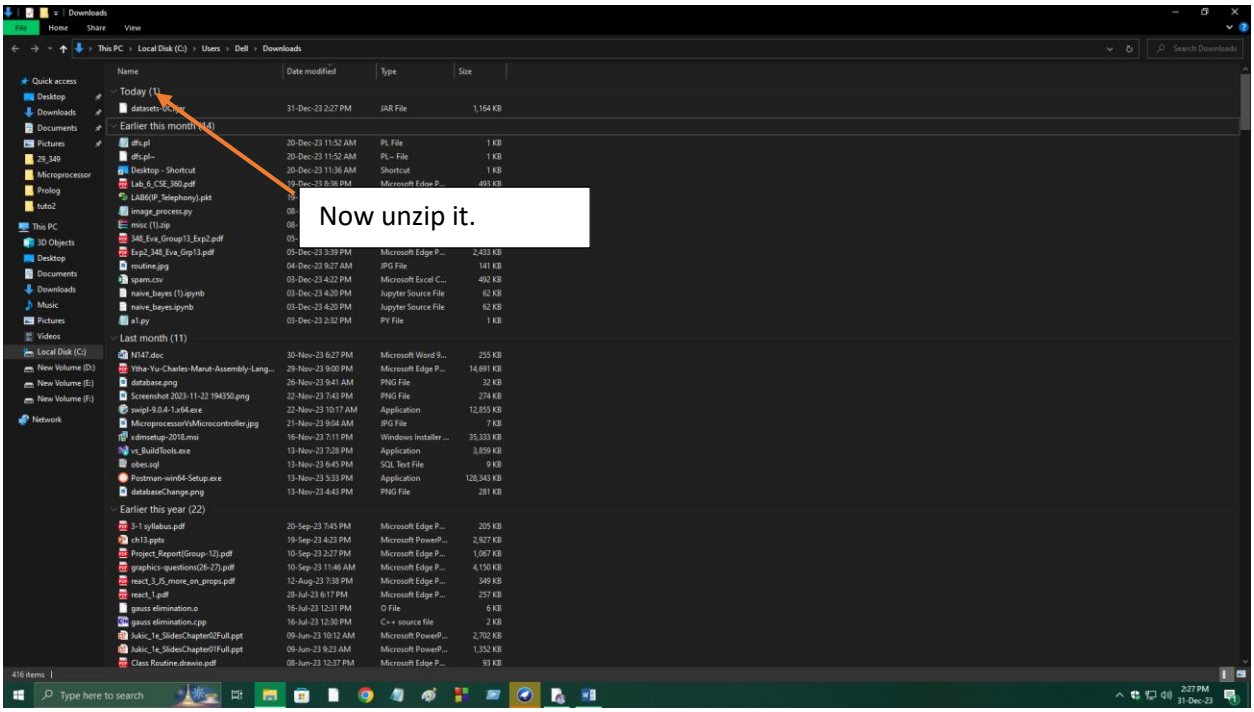


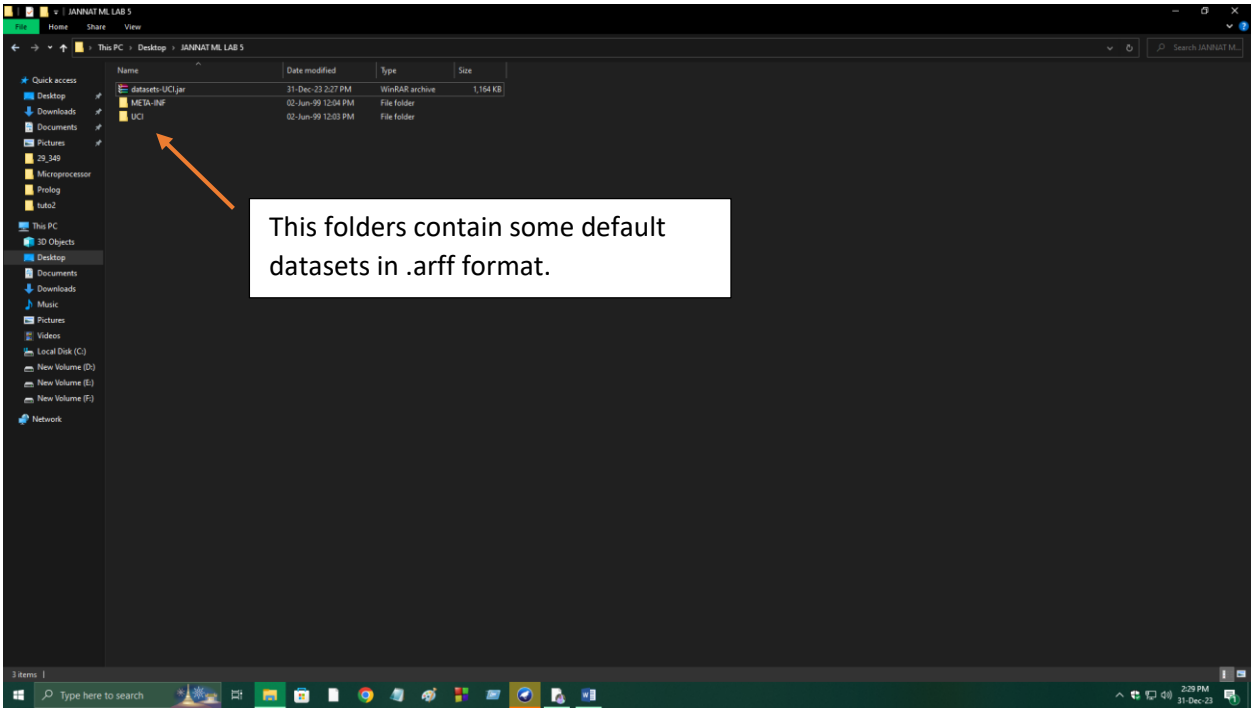
Fig 1: After launching WEKA

Step 2: Getting started with Weka

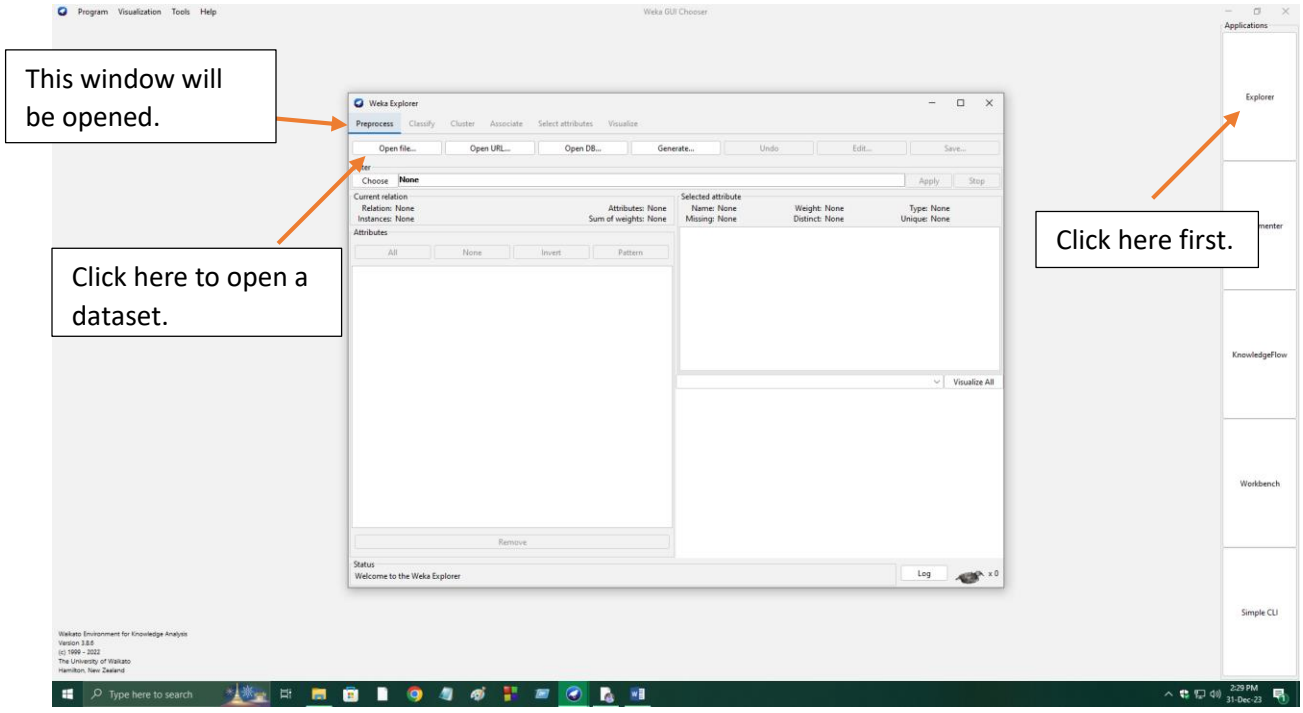
i) At first, we will download a dataset from Weka Wiki. Link: https://waikato.github.io/weka-wiki/datasets/



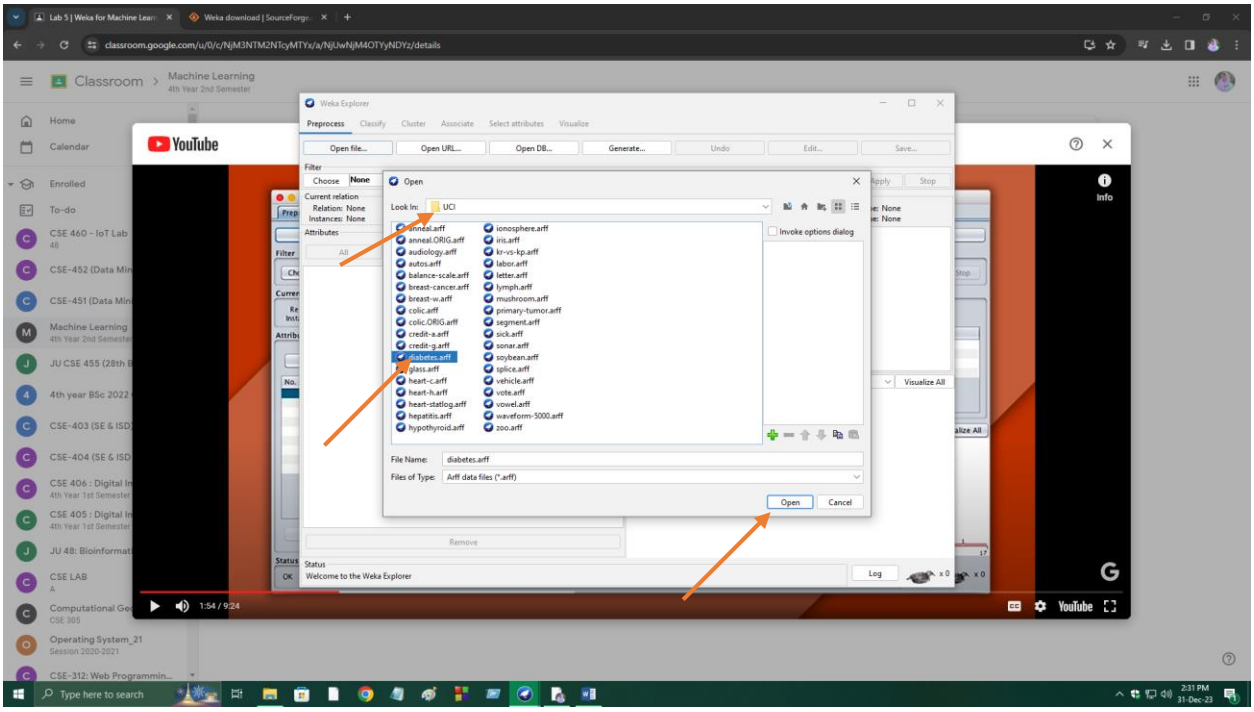


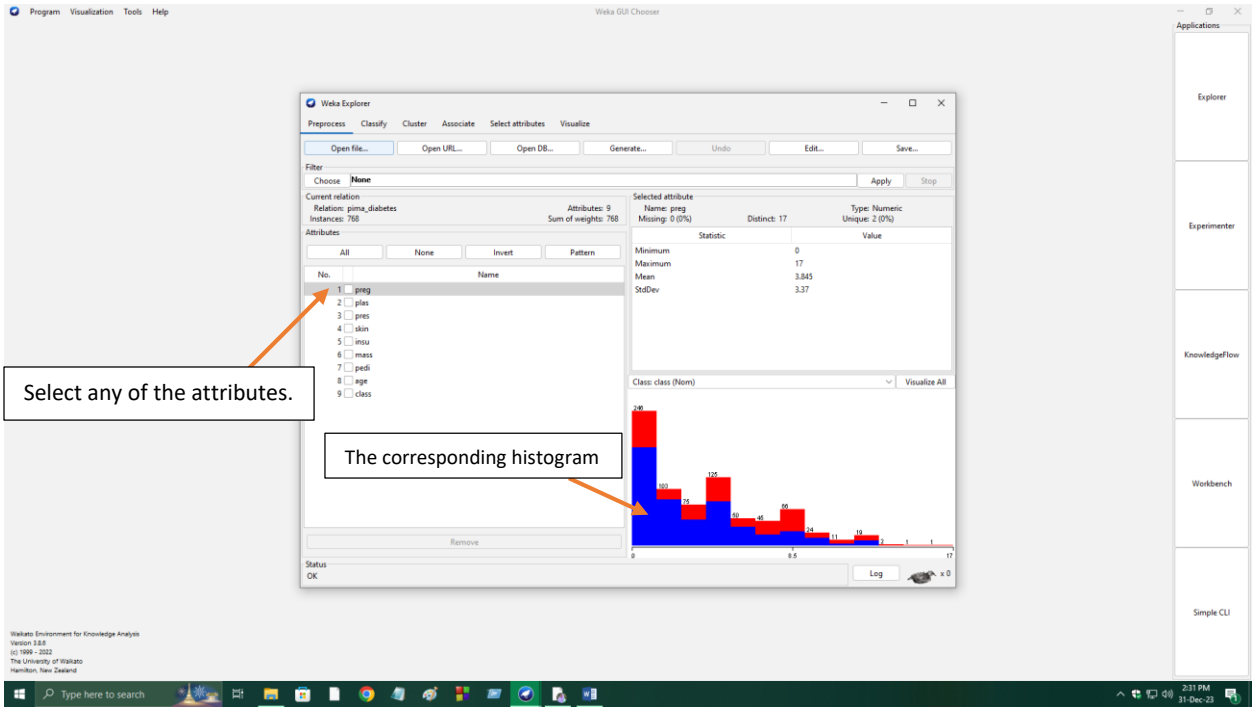


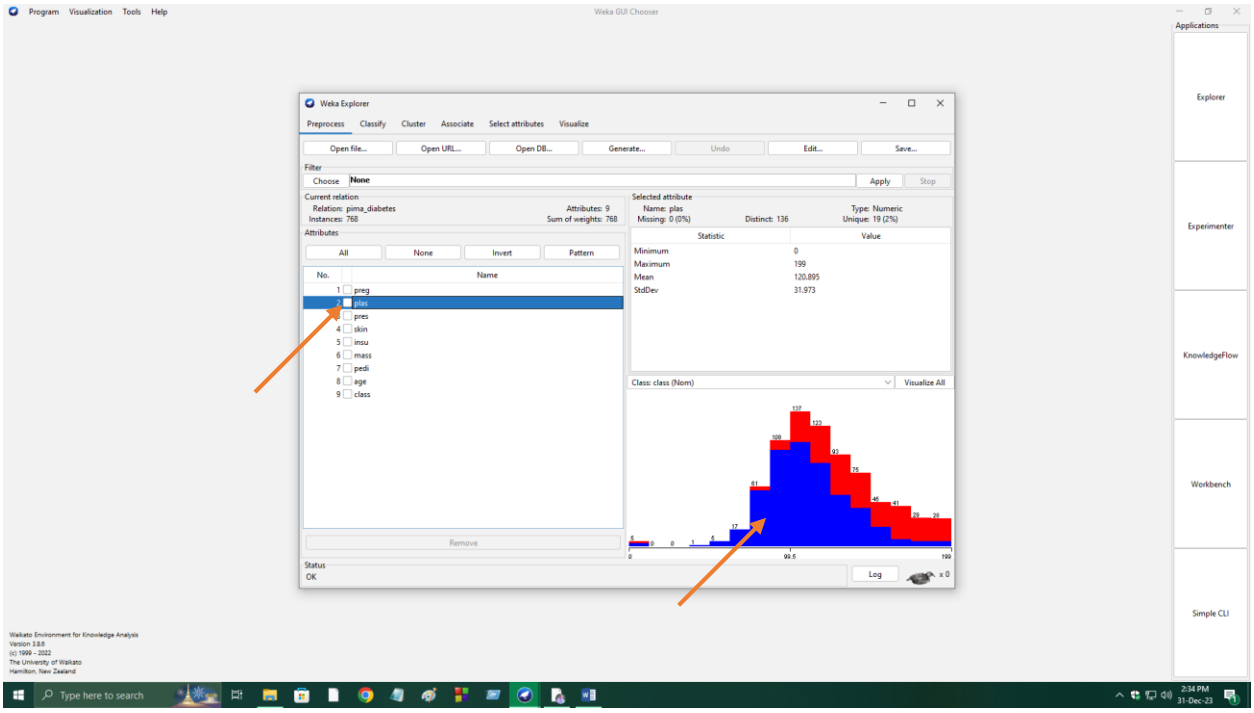
ii) Now launch Weka and click on explorer.



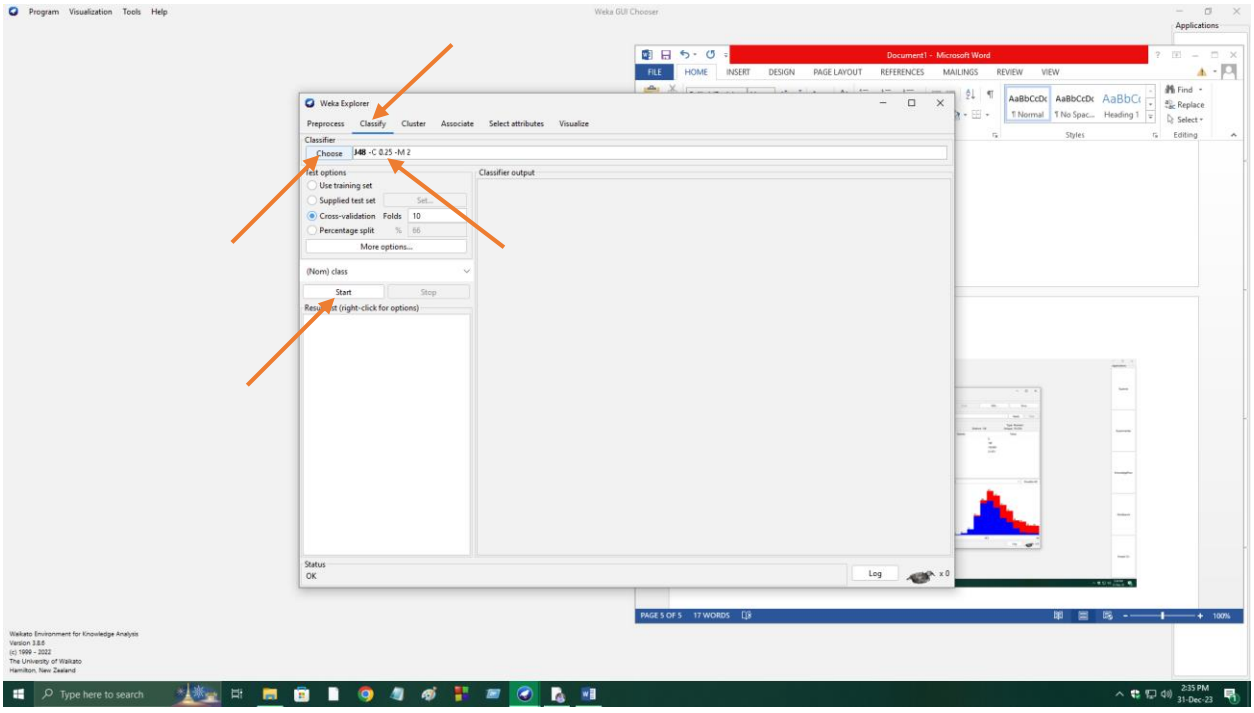
iii) Select diabetes.arff file to open. (You can open any file)

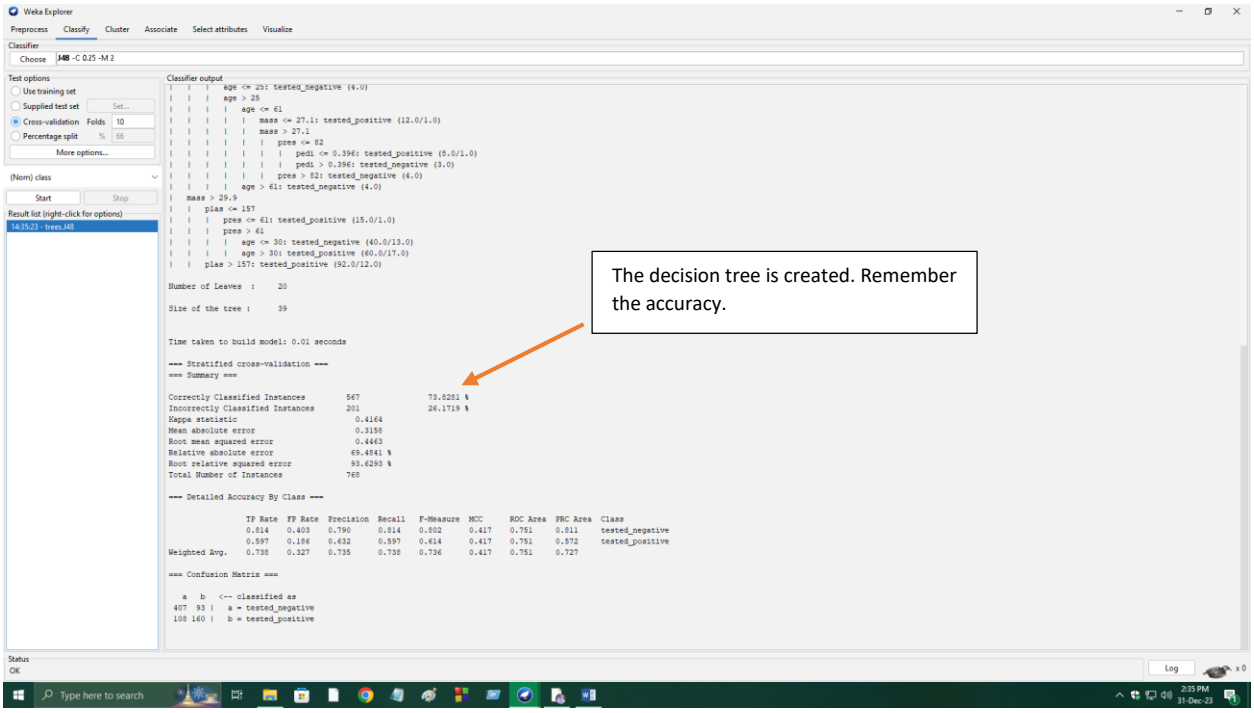




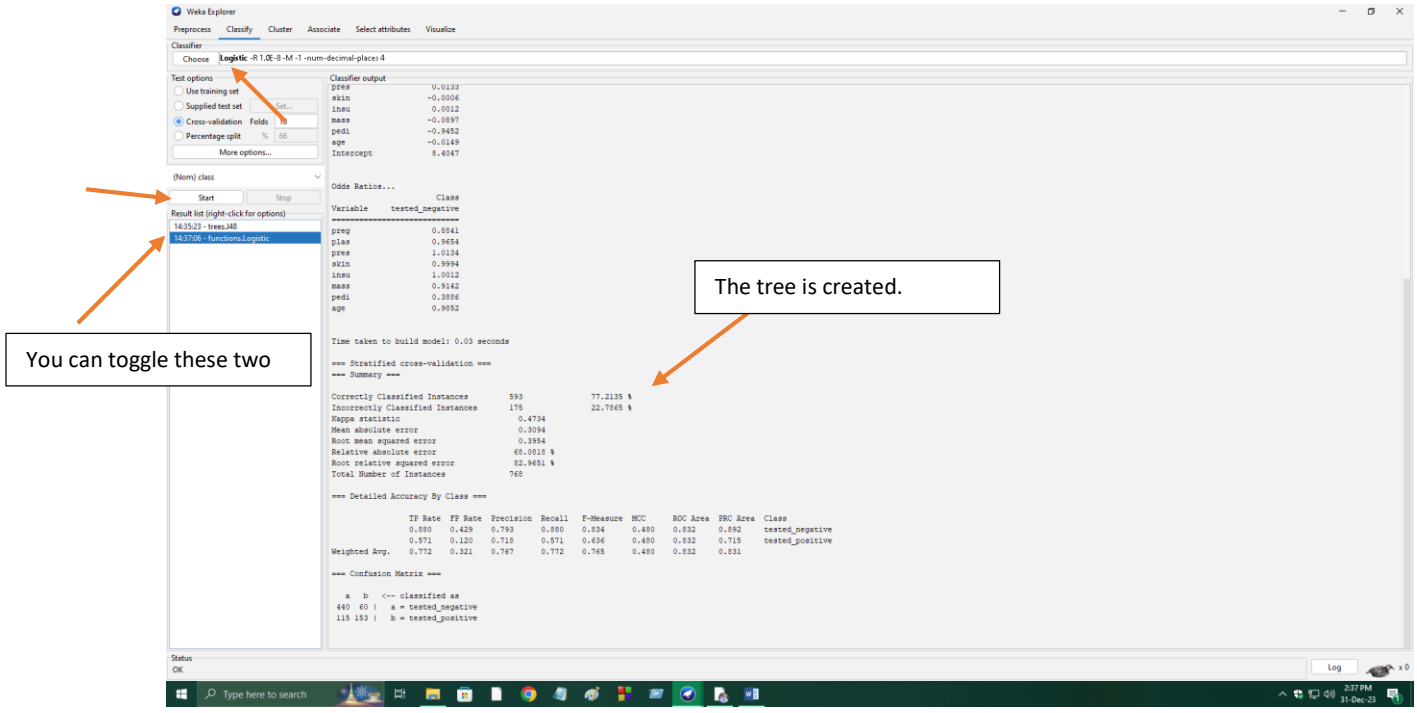


iv) Now go to classify. Then choose tree algorithm J48. Click start.

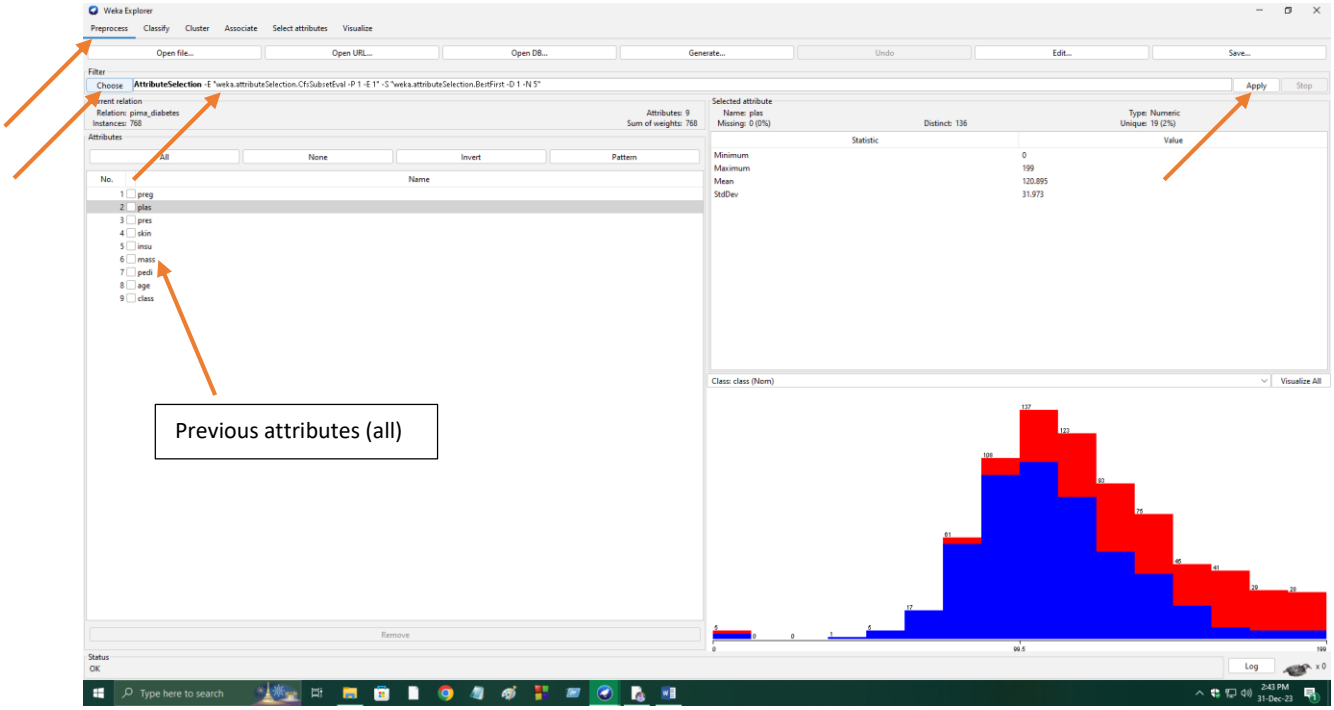


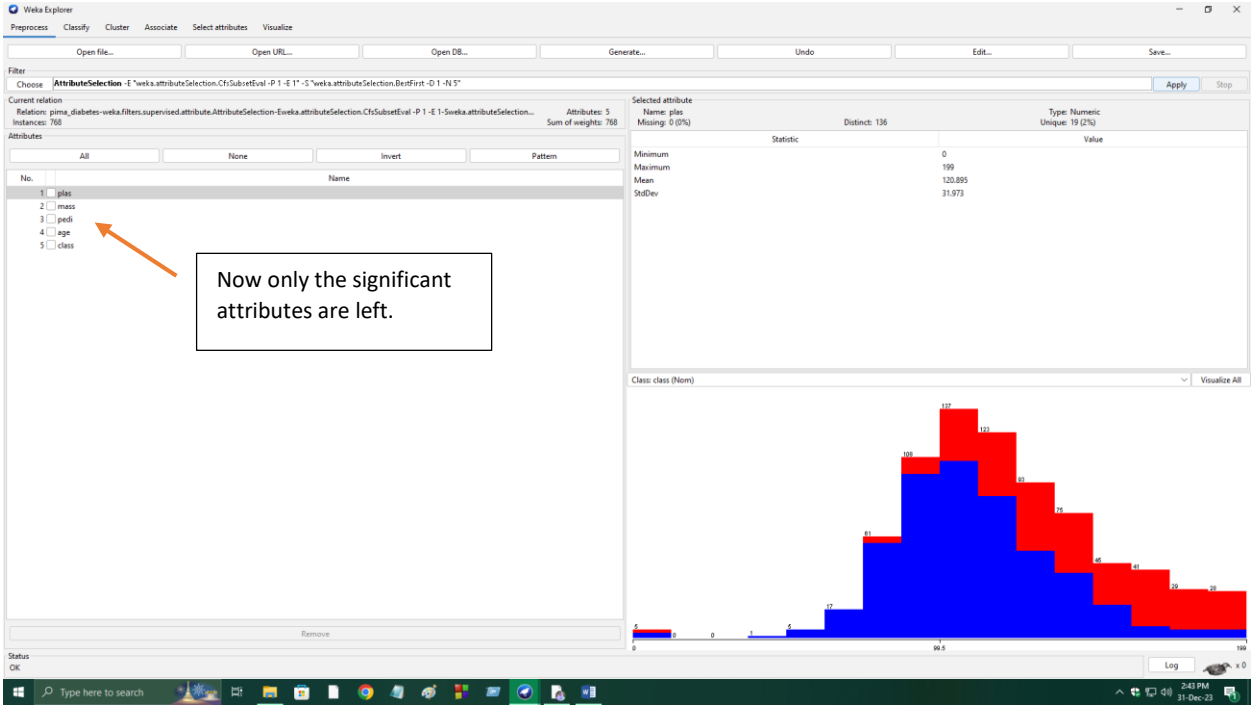


v) Now choose another algorithm called Logistic and click start.

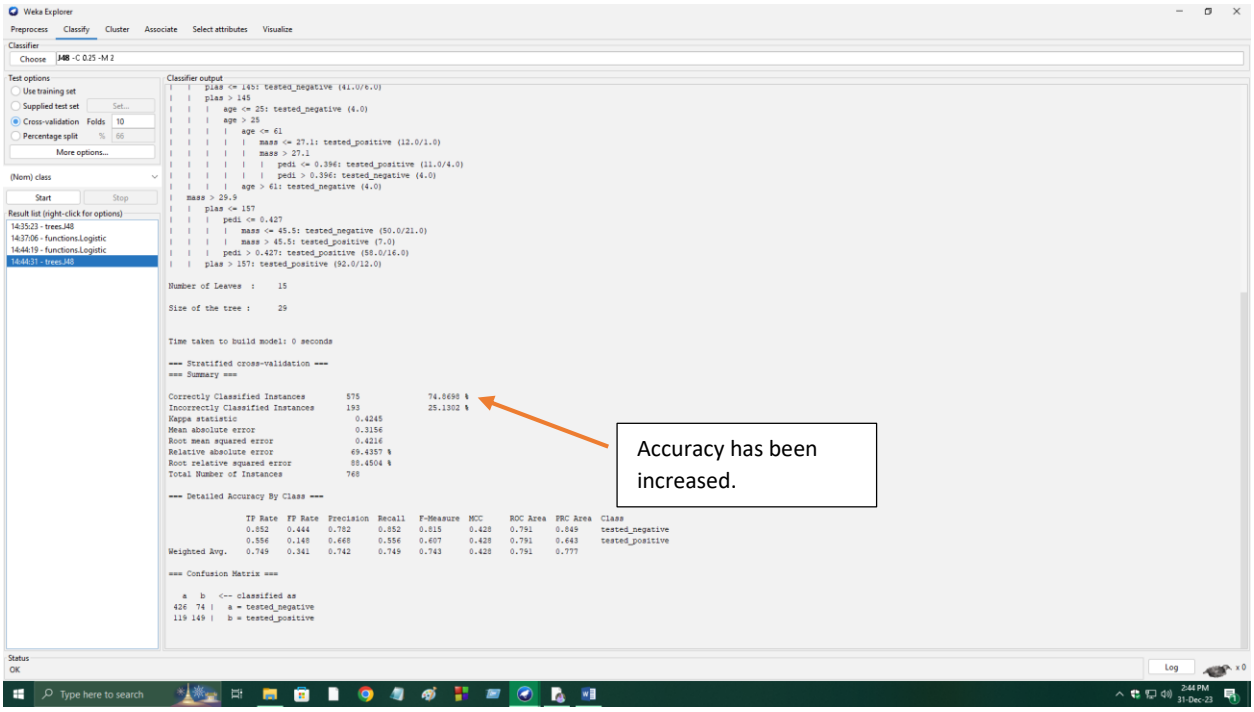


vi) Now we will filter out the less significant attributes by selecting a filter called AttributeSelection. Follow: Preprocess > Choose > Unsupervised > attribute > AttributeSelection > Apply.





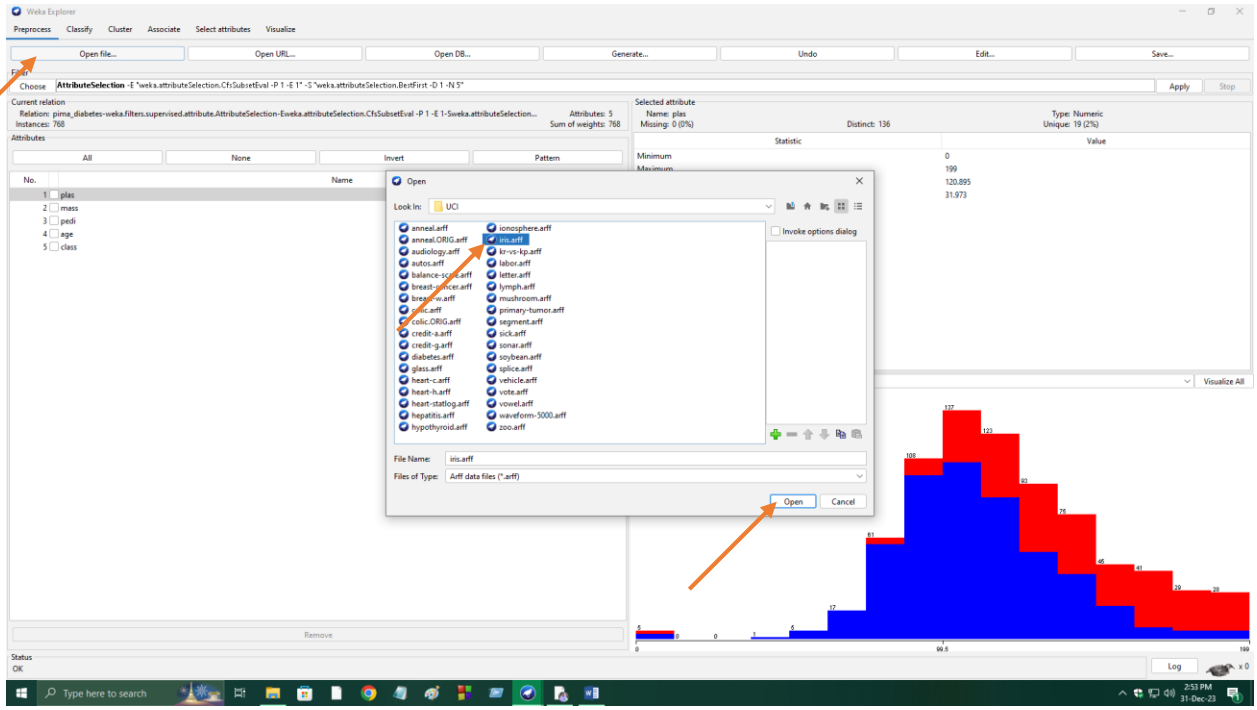
vii) Now again apply J48 from classify and check the accuracy of the decision tree.



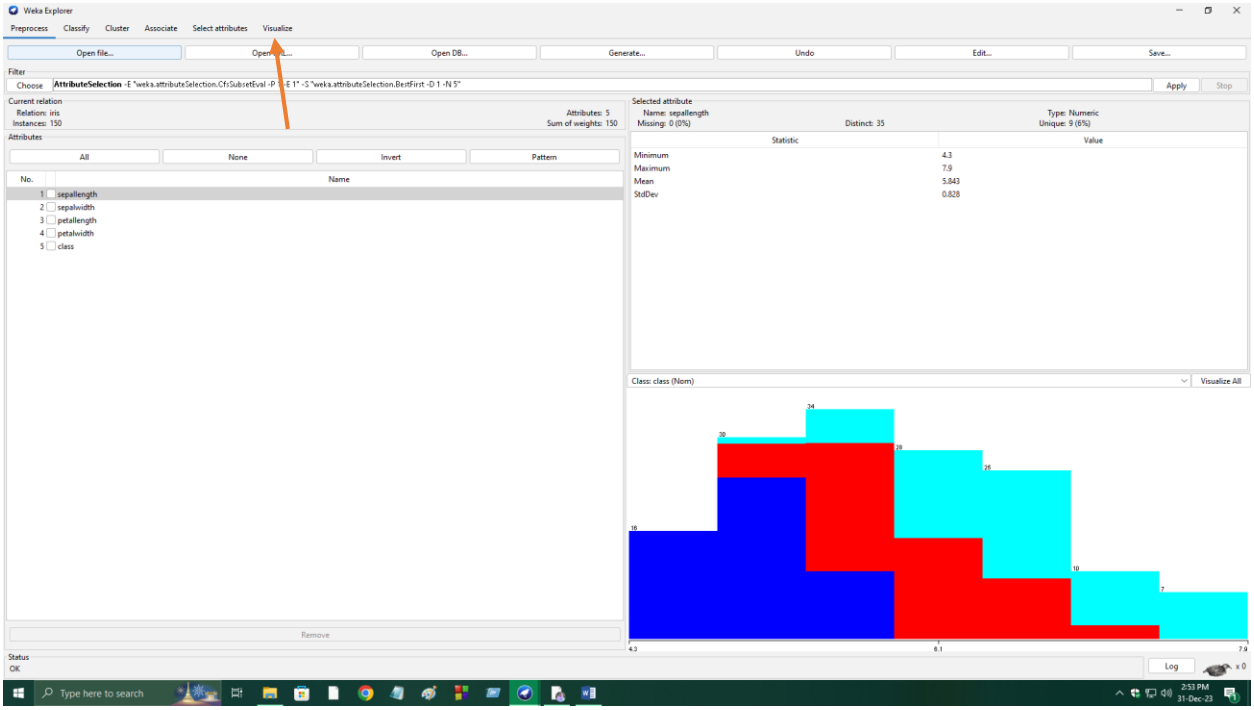
viii) Tutorial video link: https://www.youtube.com/watch?v=TF1yh5PKaqI

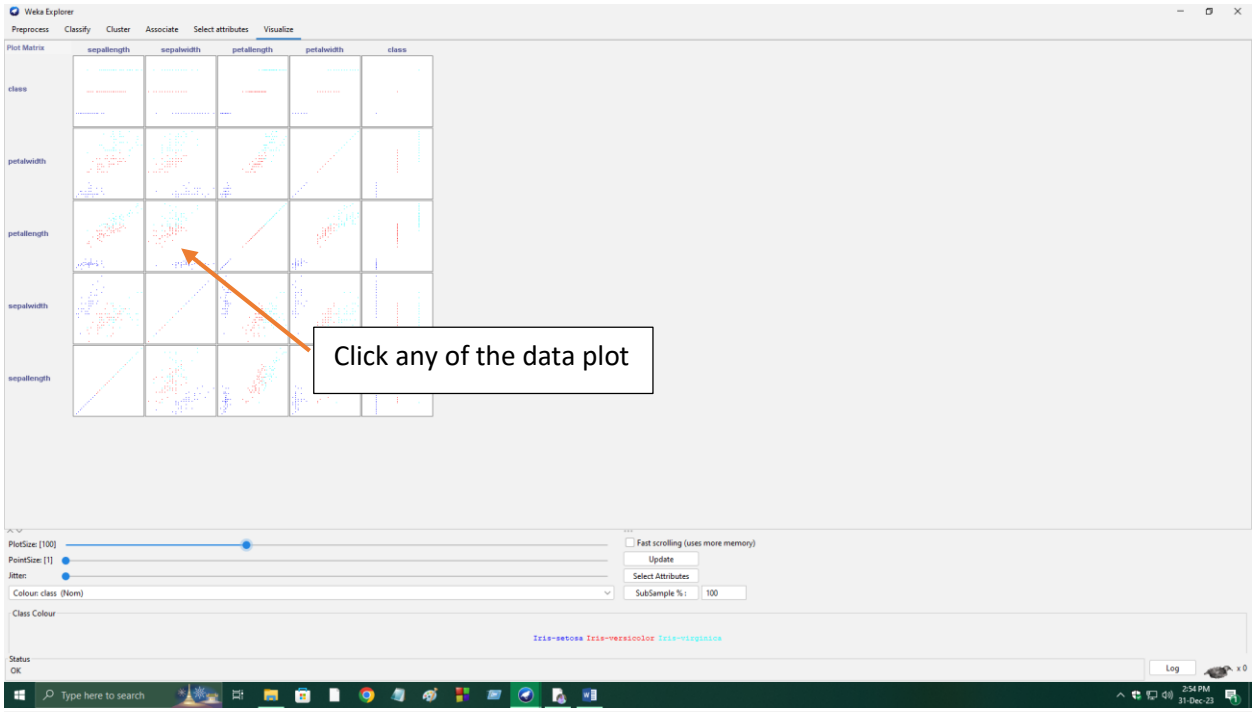
Step 3 Visualize your data from the Weka Dataset

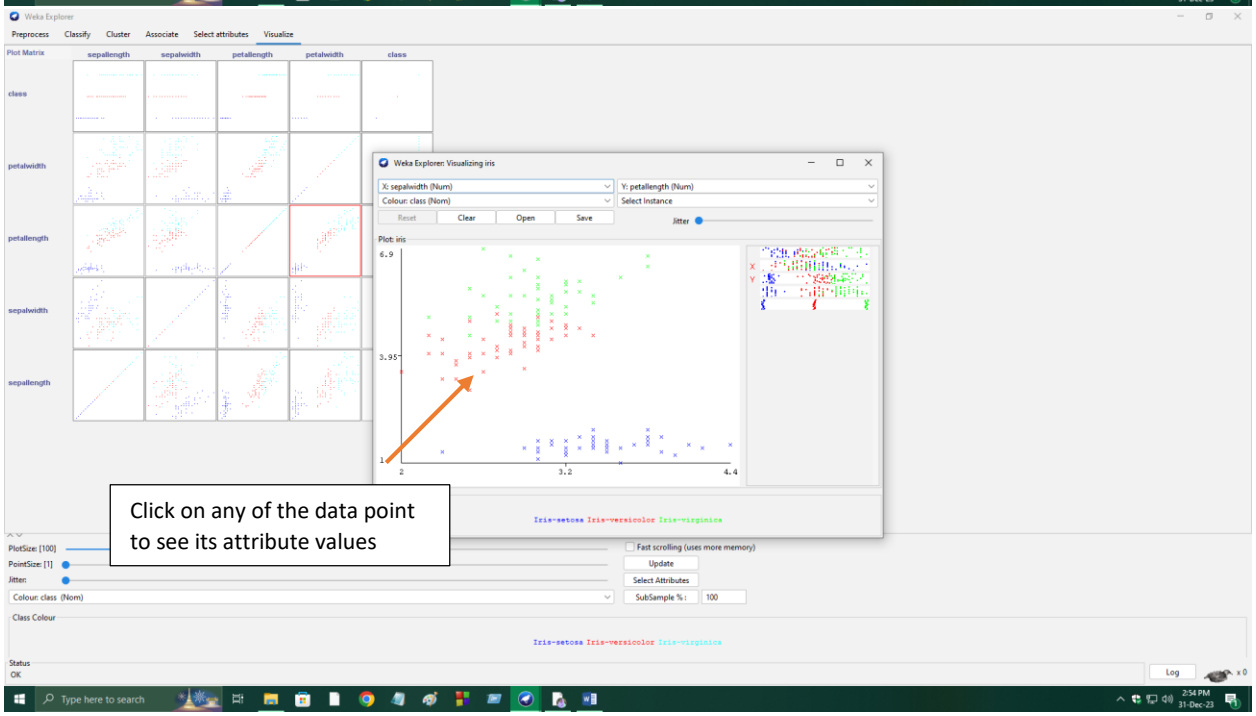
i) Open a dataset. Here we've opened iris.arff.

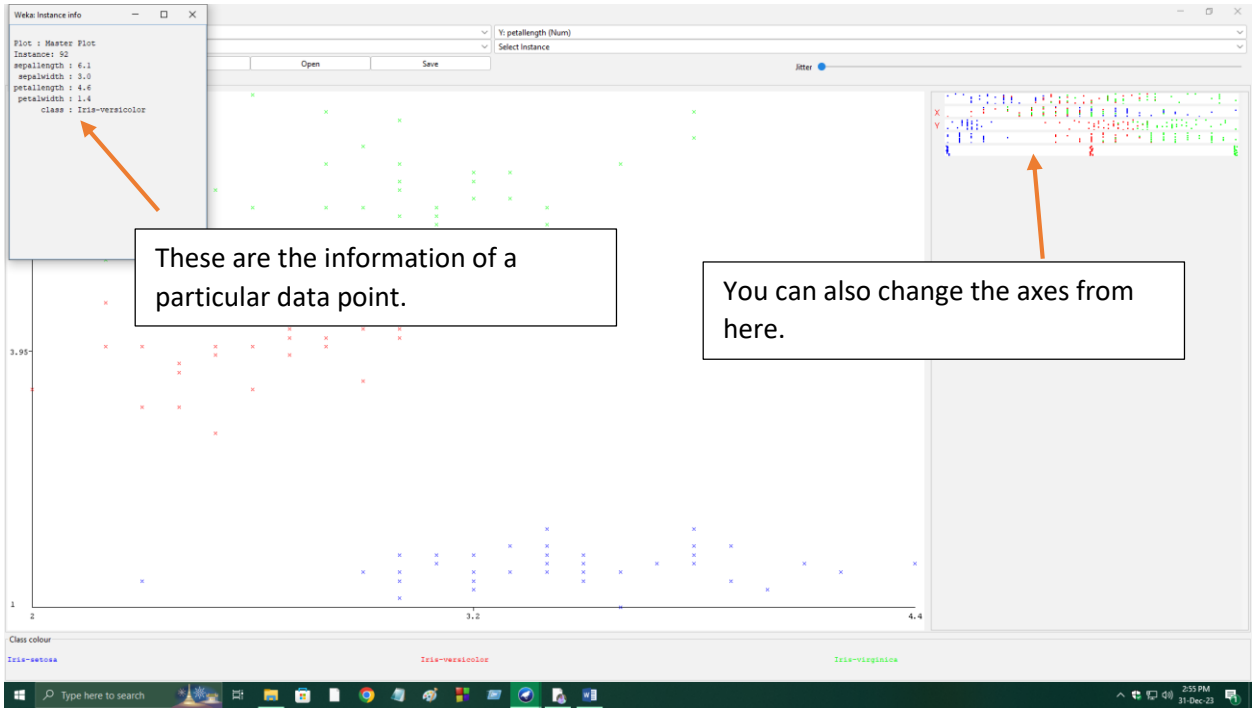


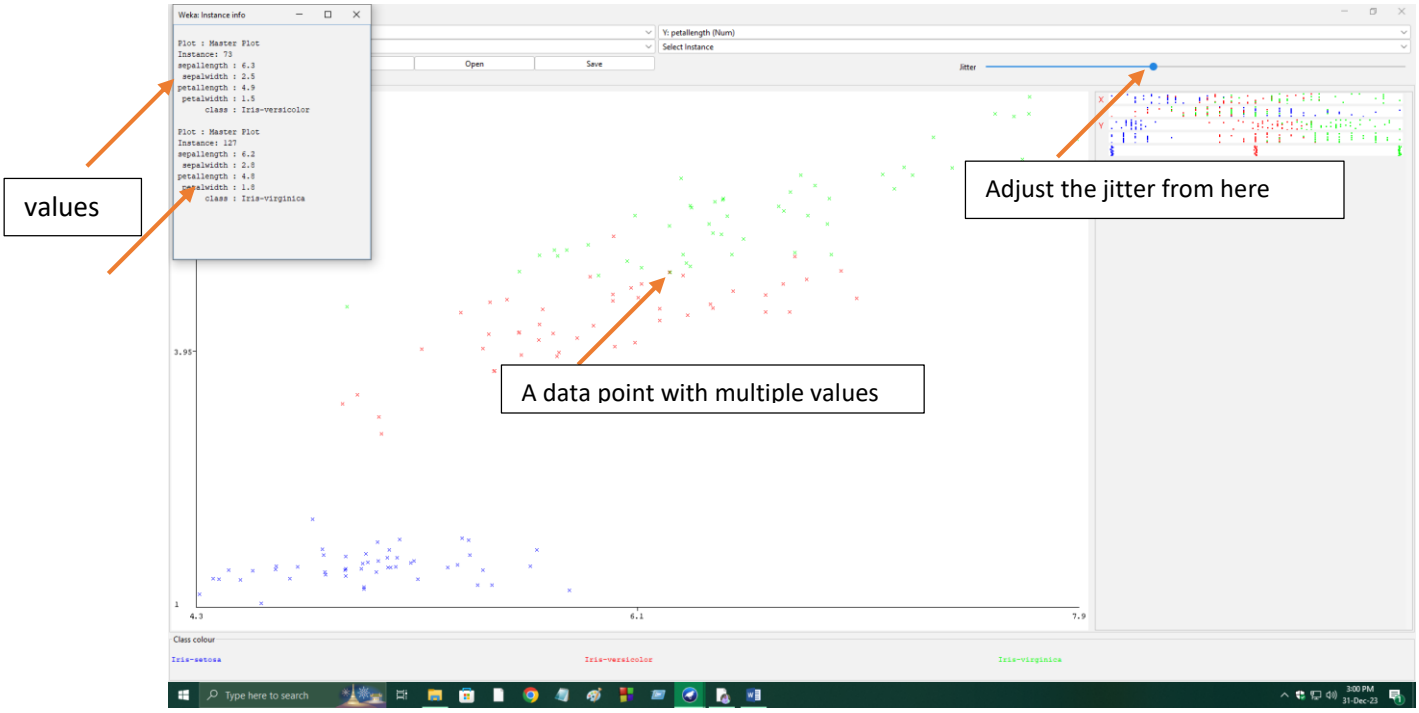
ii) Click on visualize.

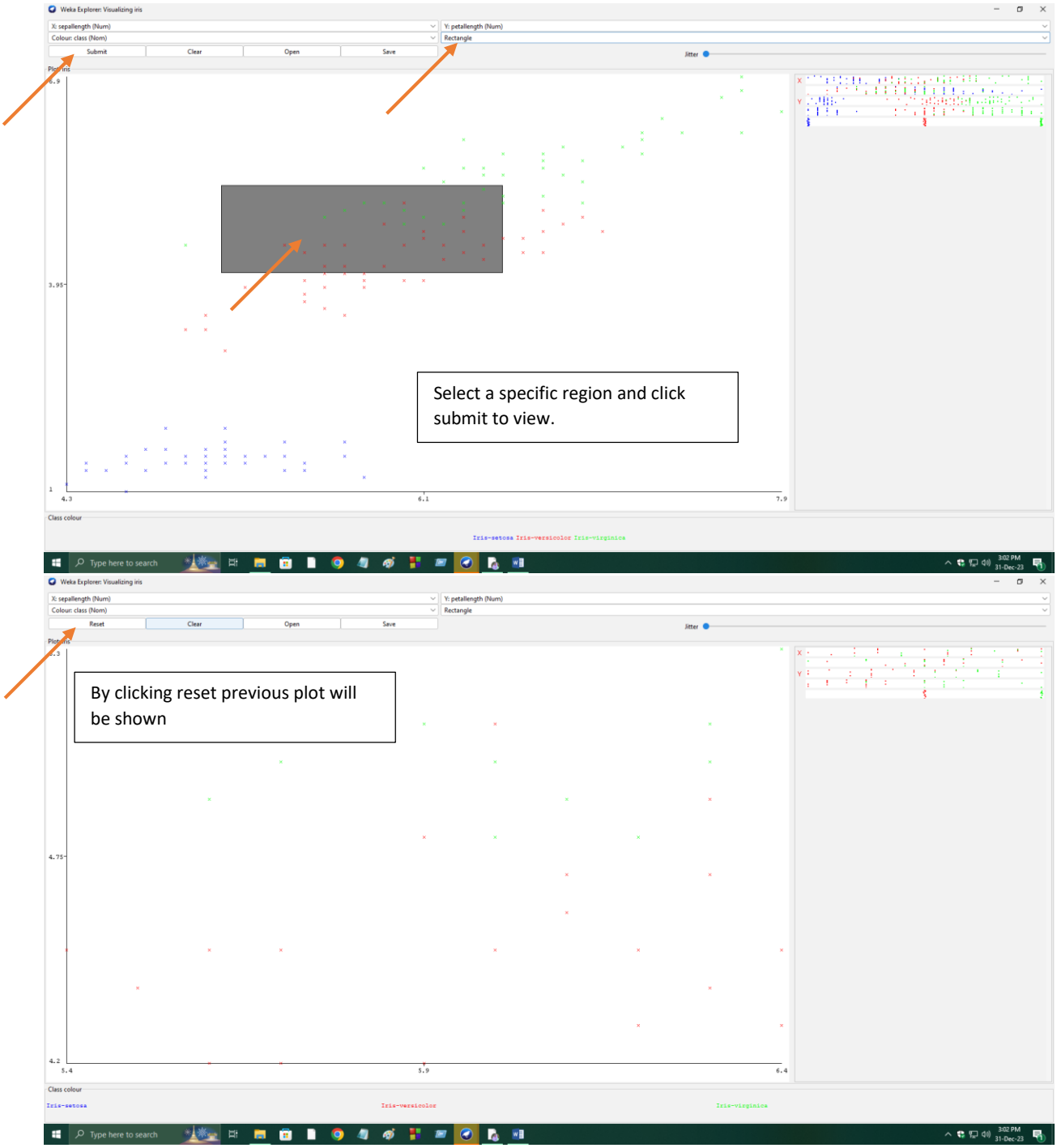



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
sepallength
sepalwidth
petallength
petalwidth
class
Click any of the data plot
PlotSize: [100]
PointSize: [1]
Jitter:
Colour: class (Nom)
Fast scrolling (uses more memory)
Update
Select Attributes
SubSample %:
100
Class Colour
Iris-setosa Iris-versicolor Iris-virginica
Status
OK
Log
Type here to search
2:54 PM
31-Dec-23

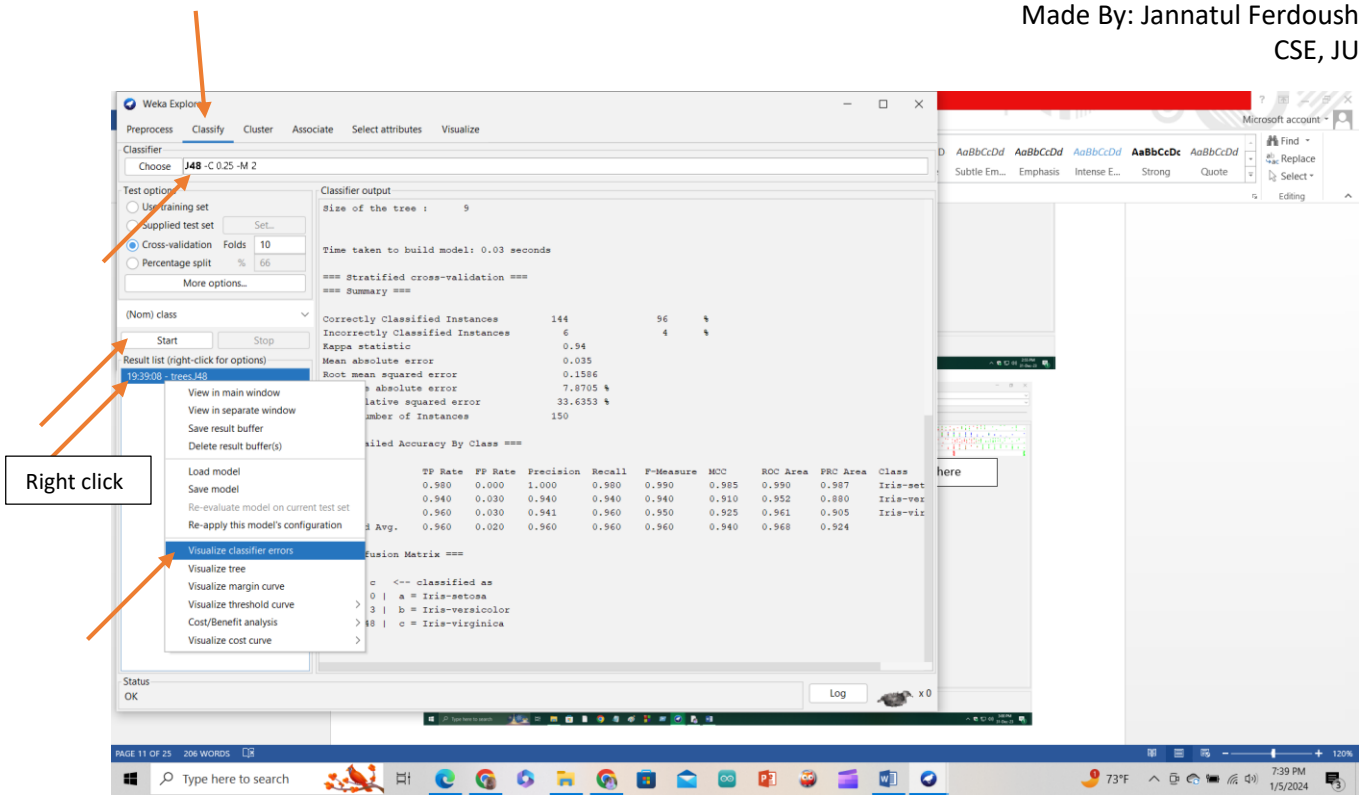

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
sepallength
sepalwidth
petallength
petalwidth
class
Weka Explorer: Visualizing iris
X: sepalwidth (Num)
Y: petallength (Num)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: iris
6.9
3.95
1
2
3.2
4.4
Click on any of the data point
to see its attribute values
PlotSize: [100]
PointSize: [1]
Jitter:
Colour: class (Nom)
Fast scrolling (uses more memory)
Update
Select Attributes
SubSample %:
100
Class Colour
Iris-setosa Iris-versicolor Iris-virginica
Status
OK
Log
Type here to search
2:54 PM
31-Dec-23

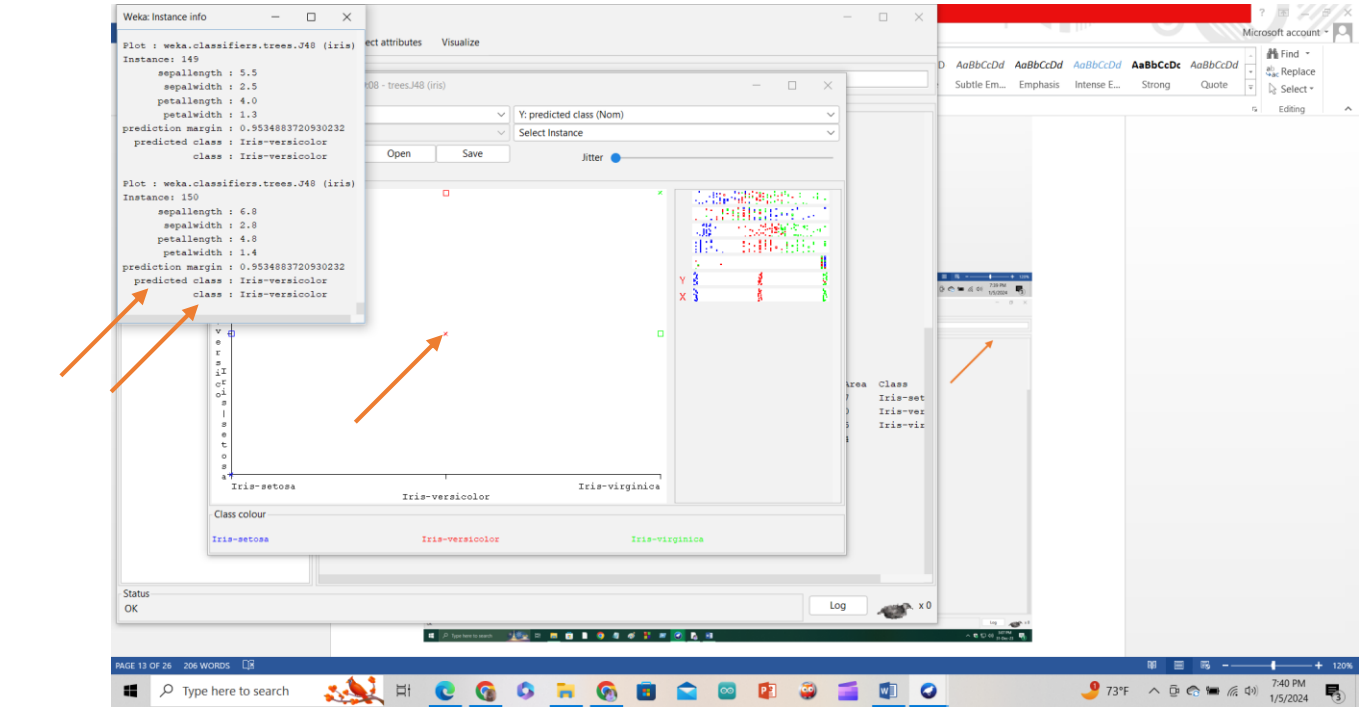

These are the information of a particular data point.
You can also change the axes from here.


values
Adjust the jitter from here
A data point with multiple values

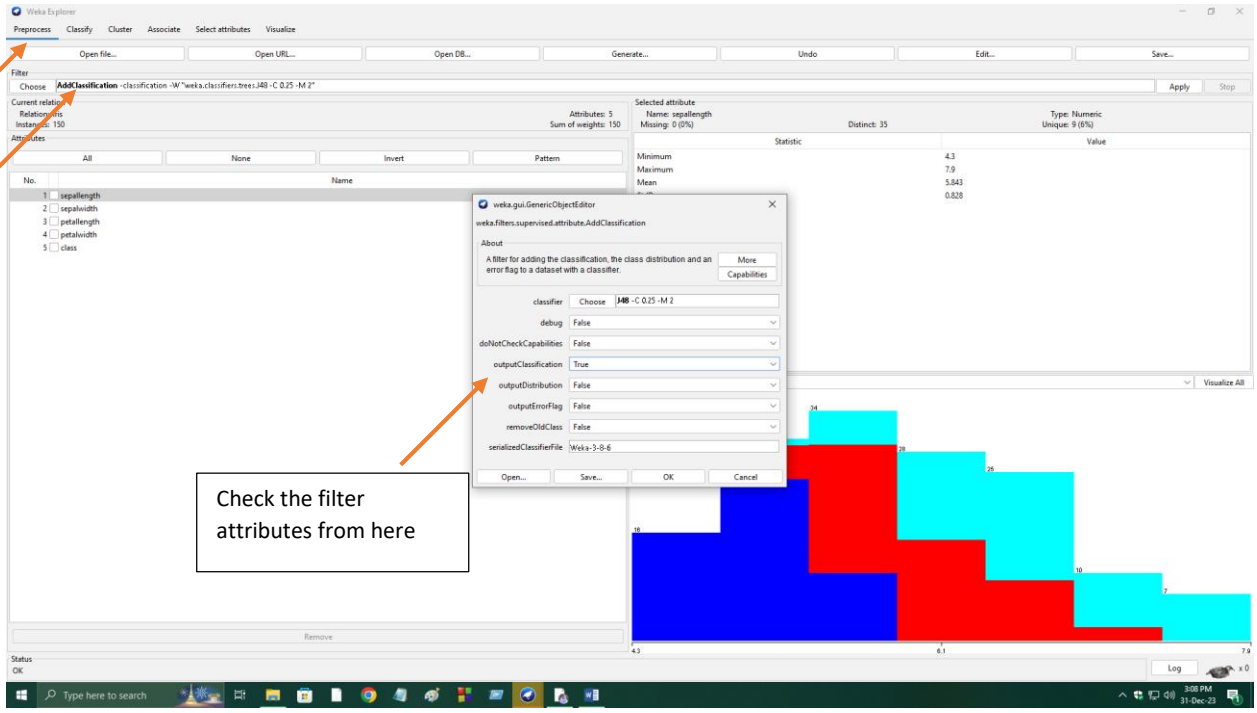


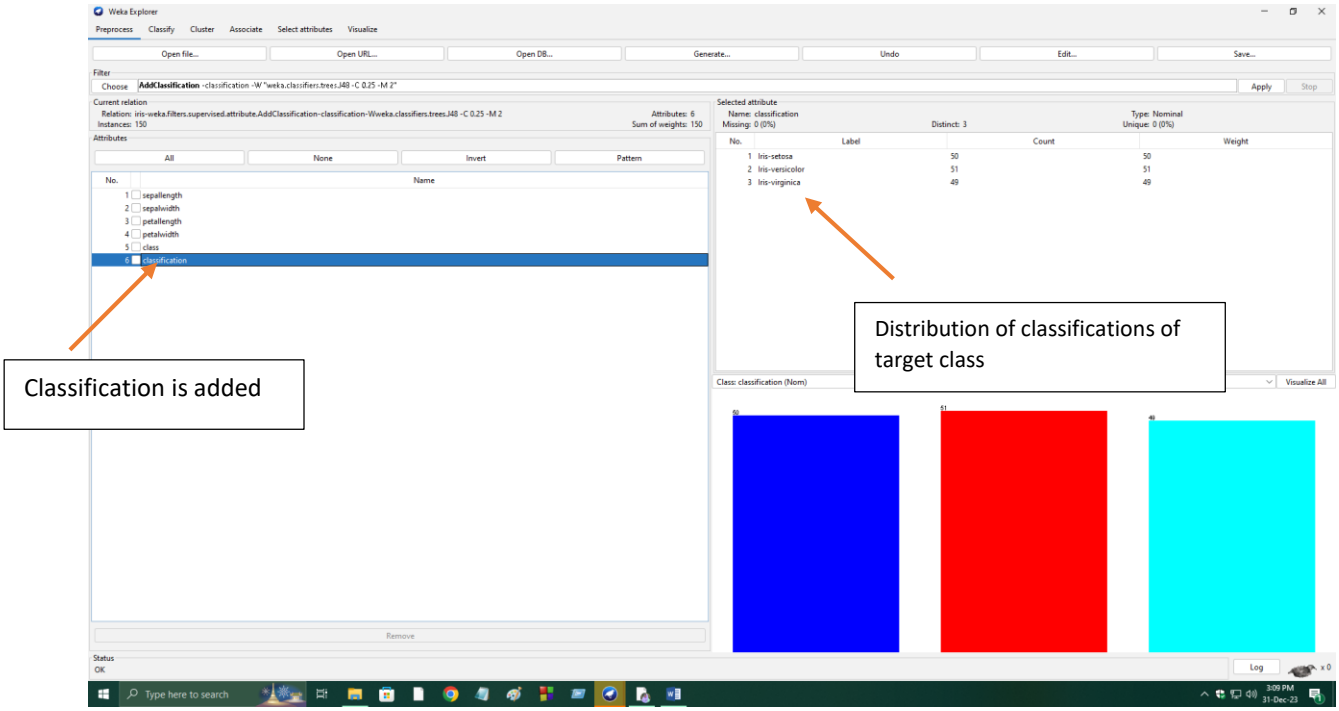
iii) Now build a J48 decision tree and visualize the errors of detecting target class.





iv) Now add classification by going: Preprocess > filter > Choose > unsupervised> attribute > AddClassification.

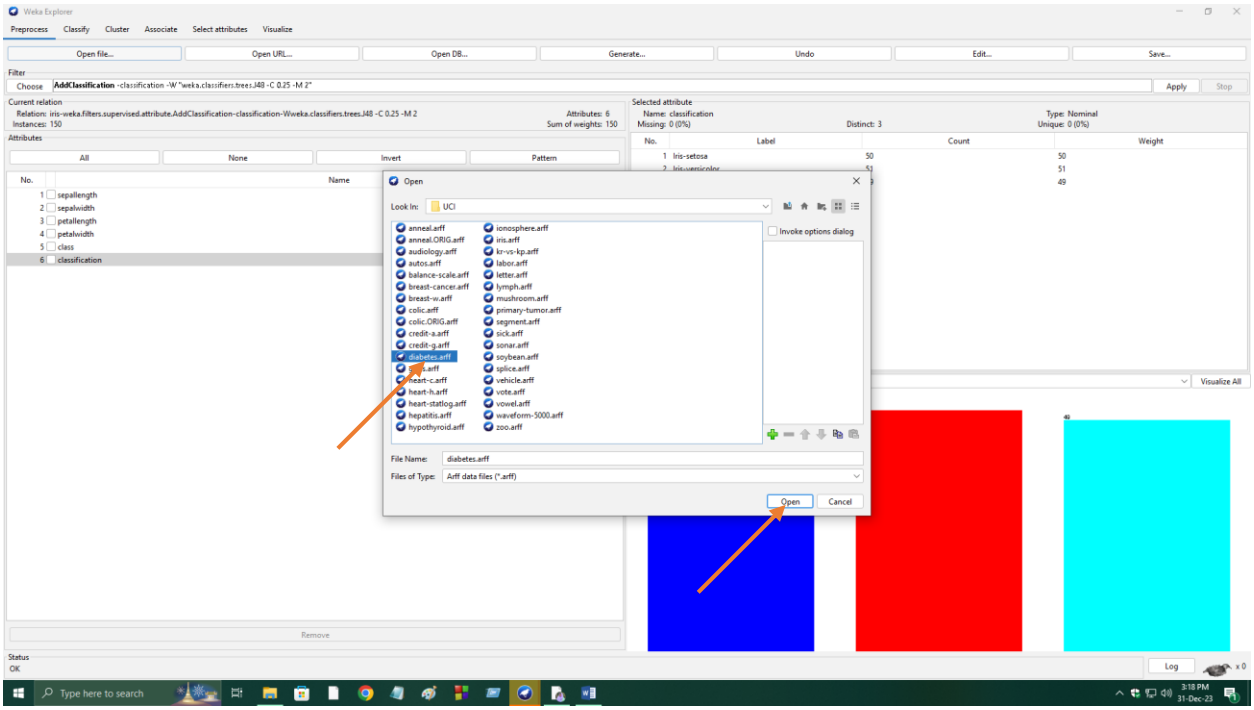




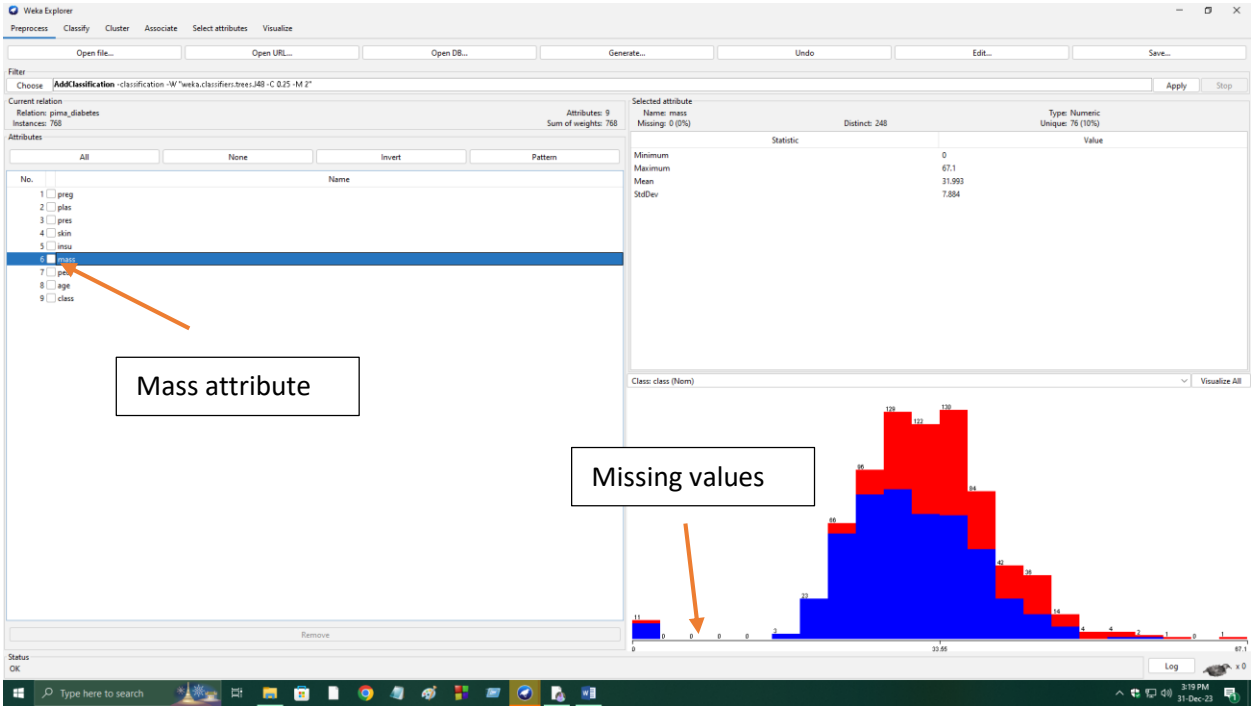
v) Tutorial video link: https://www.youtube.com/watch?v=U-1sTxmHE5U

Step 4 Handle missing values

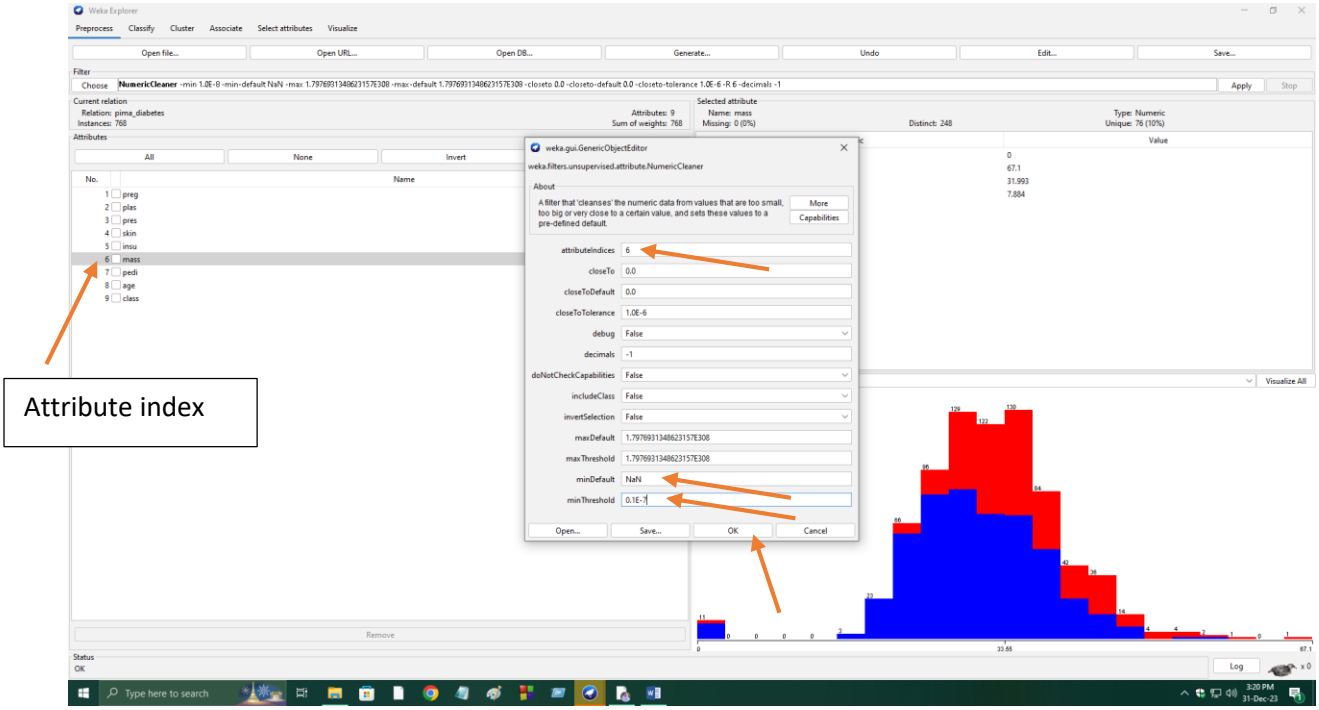
i) Open the diabetes.arff file.



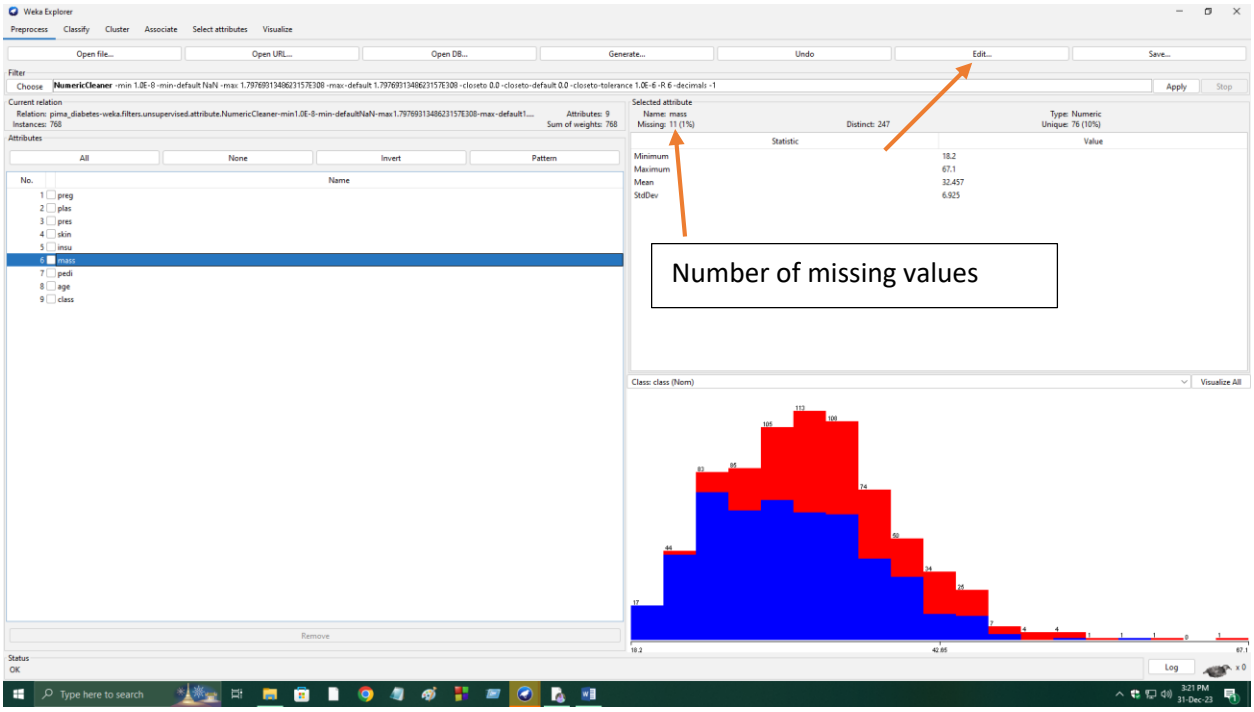
ii) Viewing an attribute named mass that has missing values as shown.



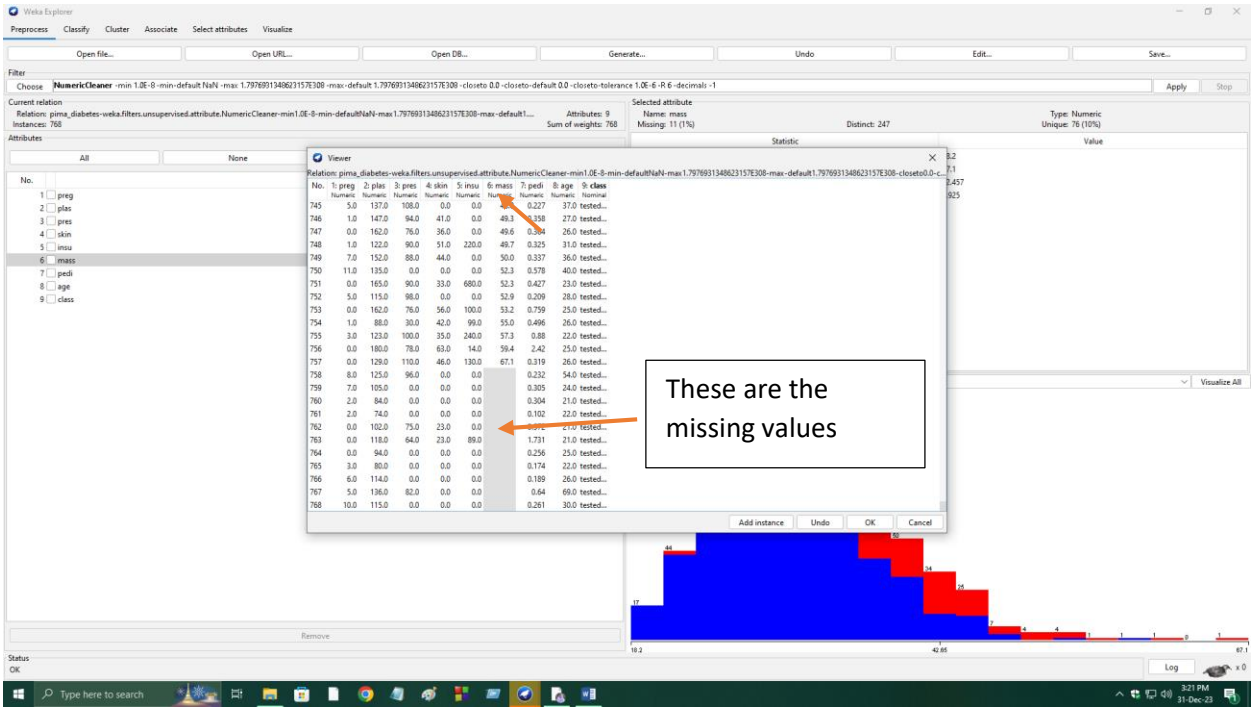
iii) To find the missing values follow: preprocess > filter > Choose > unsupervised > attribute > NumericCleaner. The click on the filter and change the fields as shown.



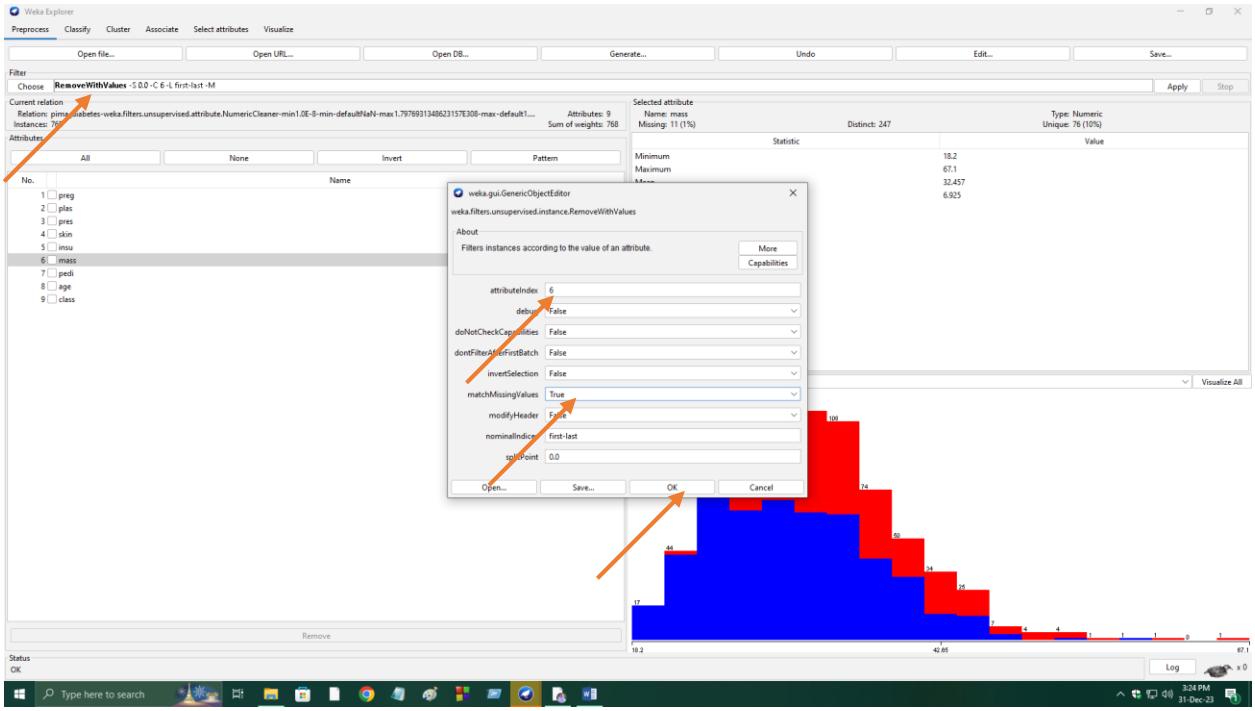
iv) Click on edit to see the missing values.



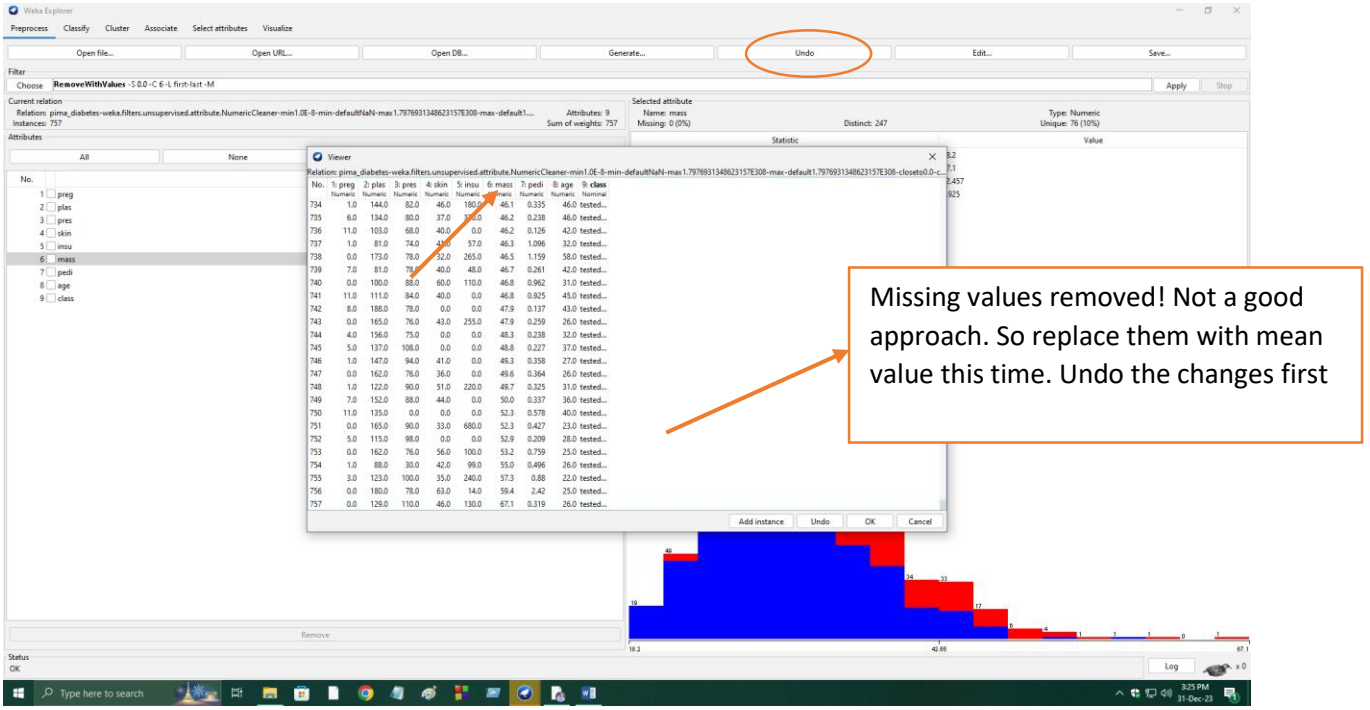
v) Select mass field and scroll down.



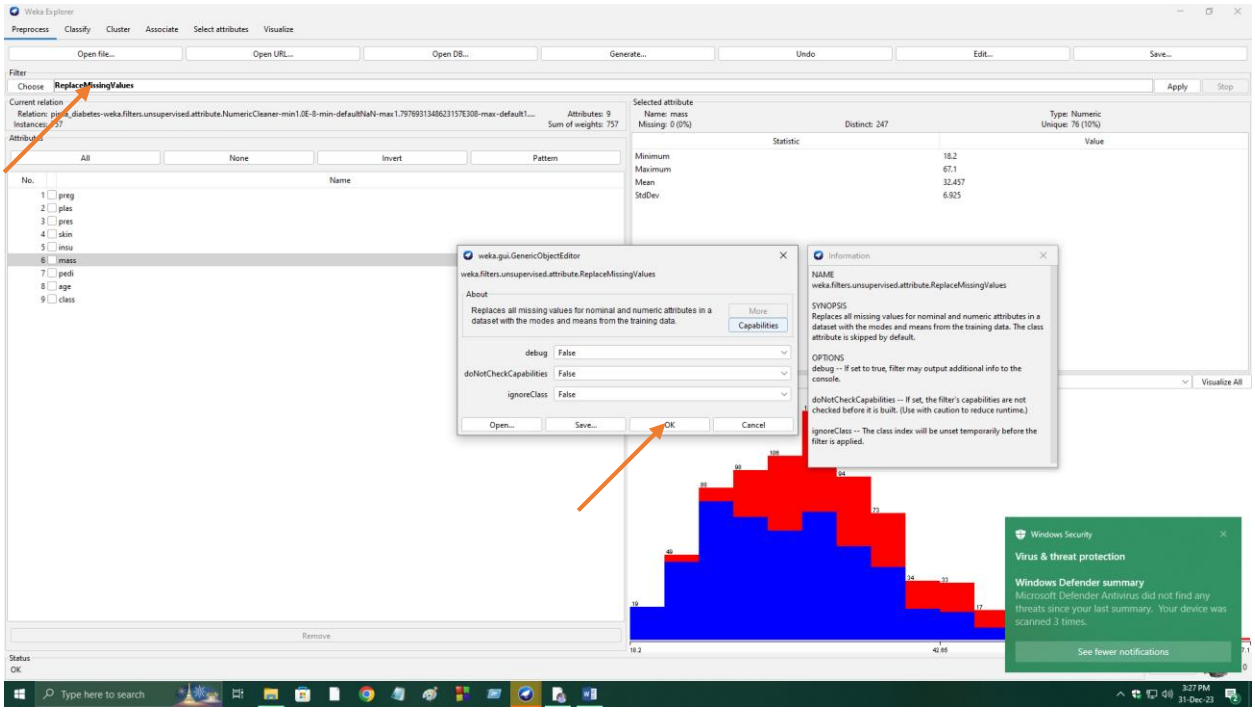
vi) Now to remove the missing values follow: preprocess > filter > choose > unsupervised > instance > RemoveWithValues. Click on the filter and modify the fields as shown. Click OK.



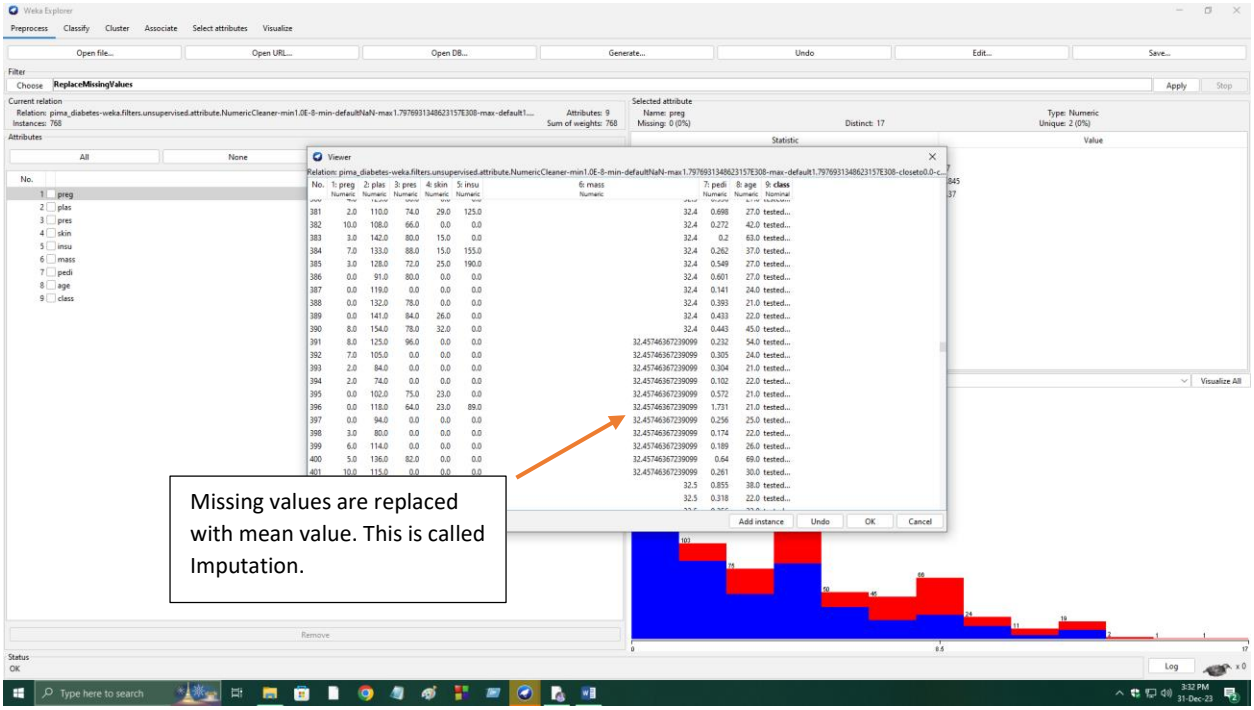
vii) Again click edit to see the missing values. This time you won't see the records containing missing values as they are removed but this is not a good approach so now will replace the missing values. So click undo to get the missing values again.



viii) We will replace the missing values with mean value. So follow: preprocess > filter > choose > unsupervised > attribute > ReplaceMissingValues.



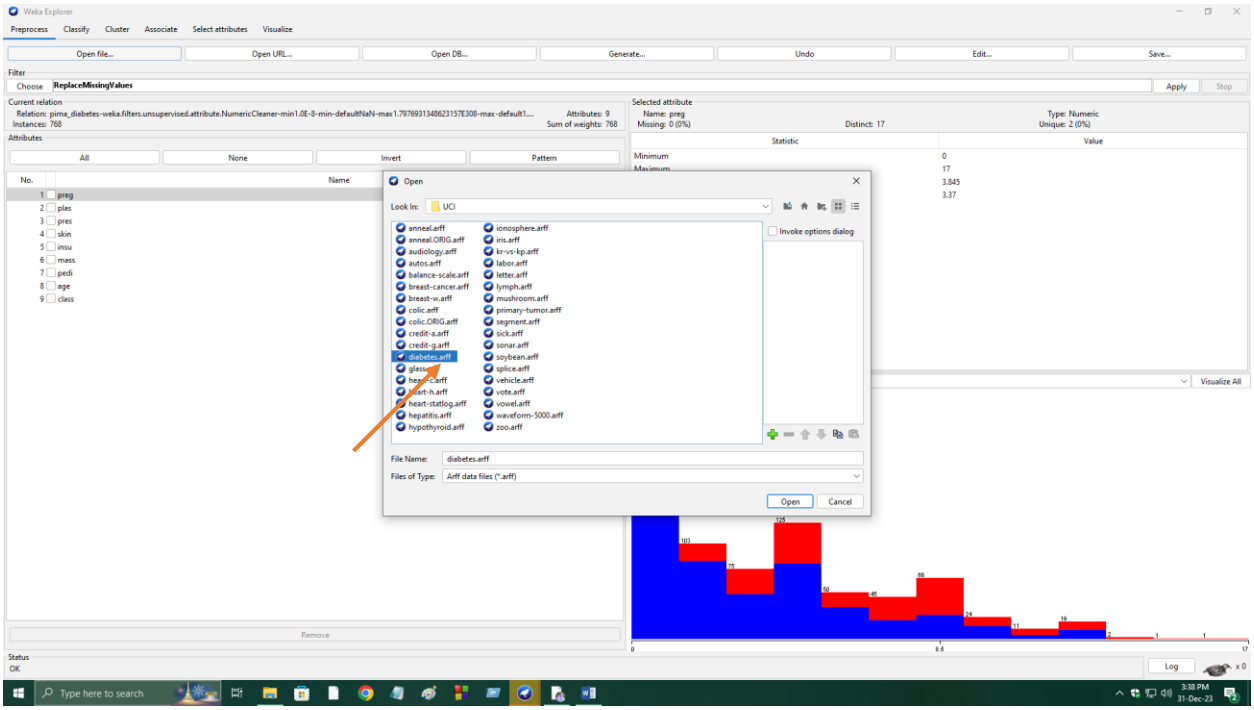
ix) Now again click on edit to view the missing values. This time after scrolling you will see the values are replaced with mean value.



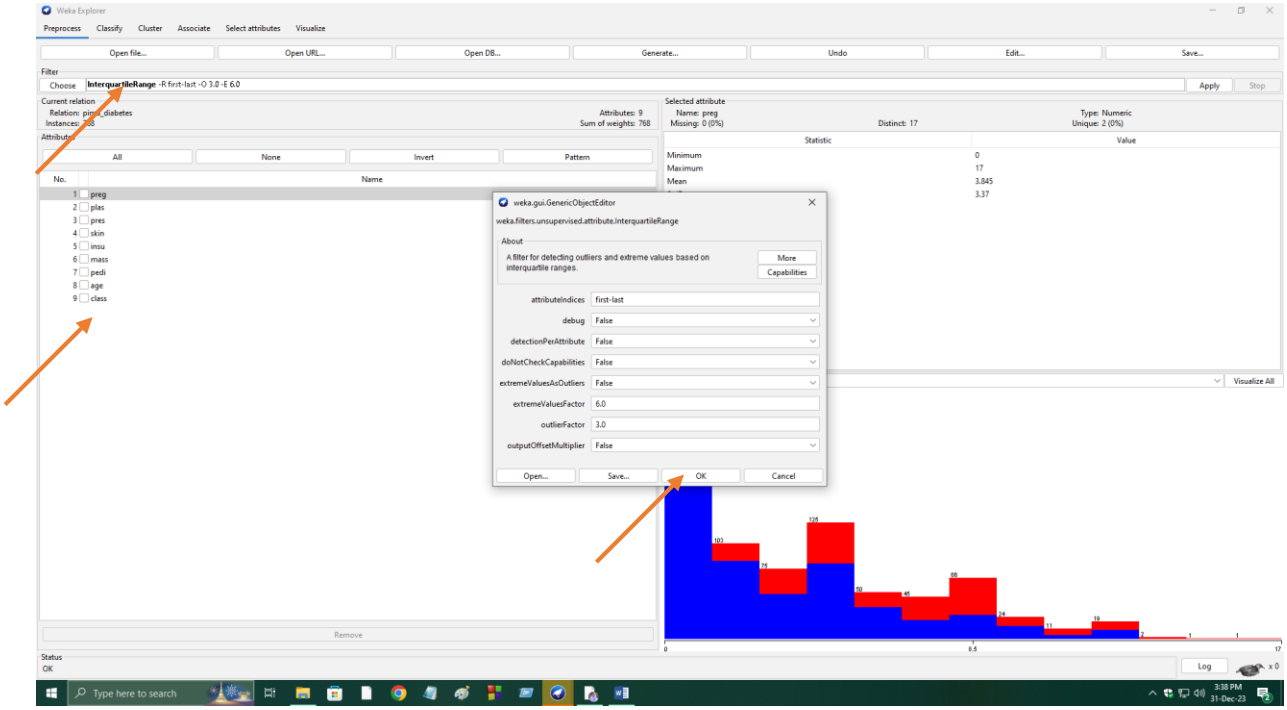
x) Tutorial video link: https://www.youtube.com/watch?v=Qw6yEwzaDJQ

Step 5 Handle Outliers and Extreme values

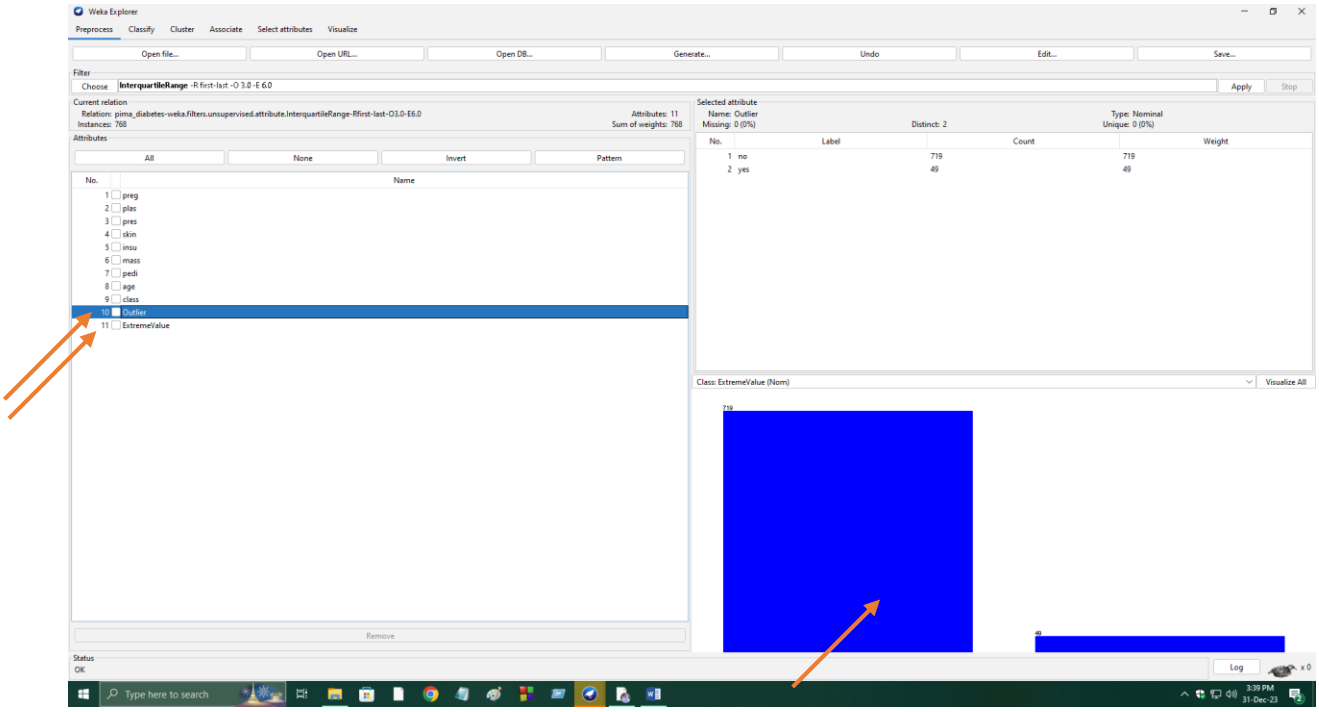
i) Open the diabetes.arff file.



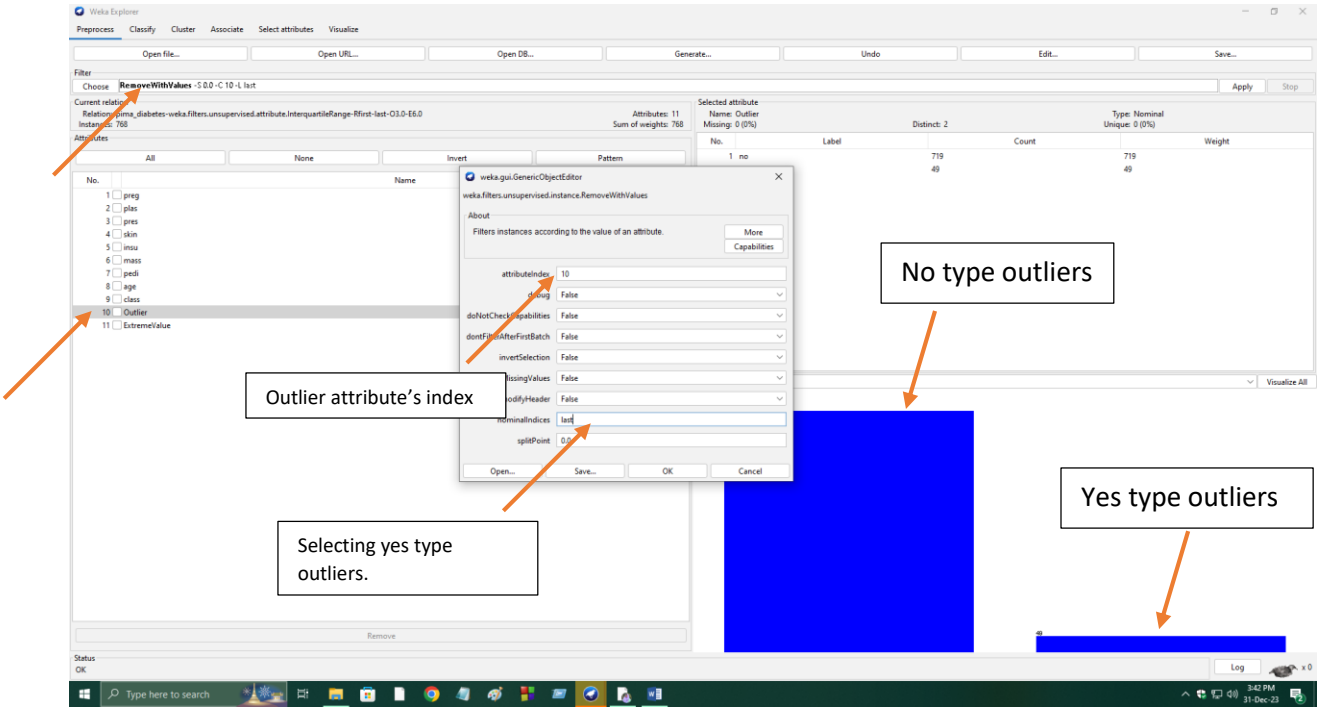
ii) To find the outliers and extreme values follow: preprocess > filter > choose > unsupervised > attribute > InterquartileRange.

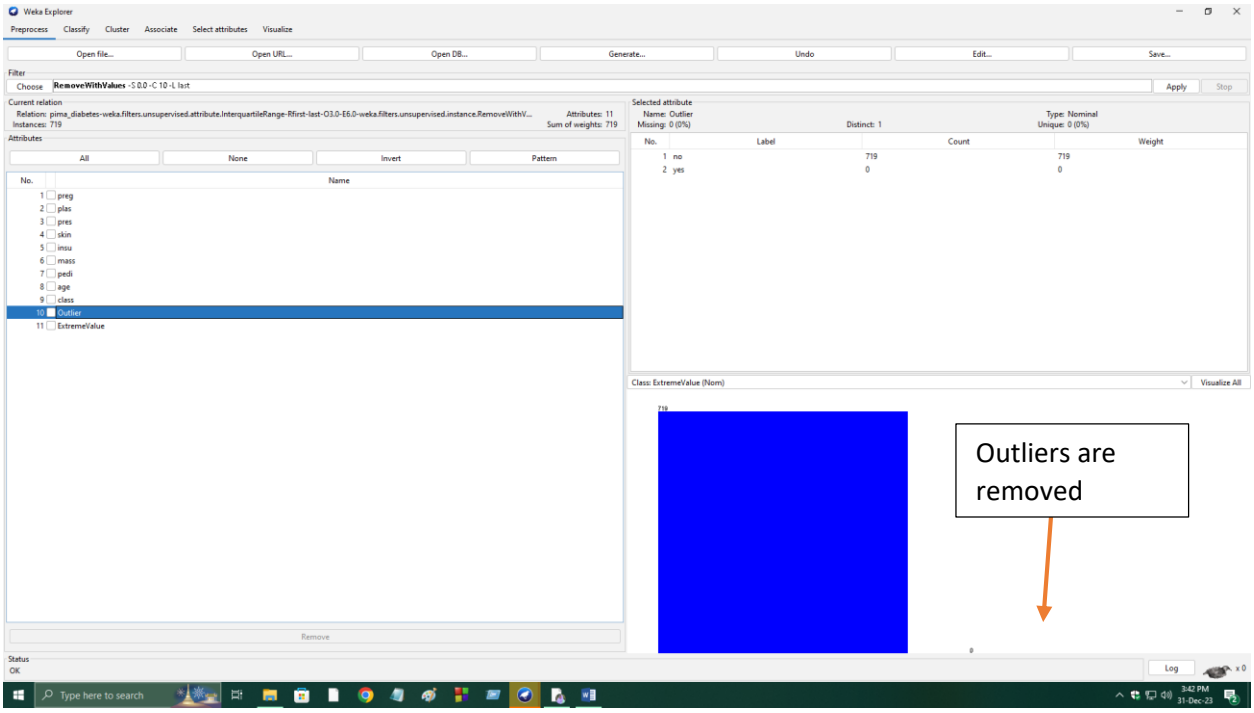


iii) Now you can see the outlier and extreme values.



iv) There are two types of outliers (no and yes). We want to remove the yes type outliers. So follow: preprocess > filter > choose > unsupervised > instance > RemoveWithValues. Click on the filter. And modify the fields as shown.



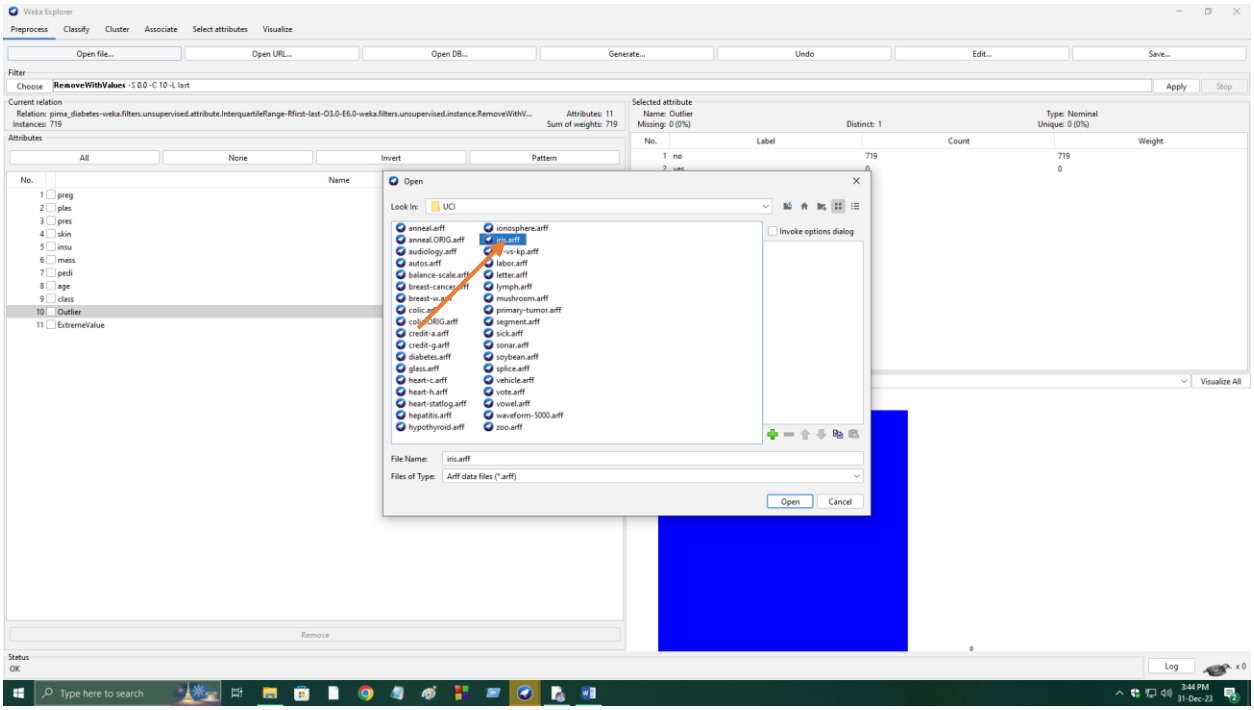


v) Following the same way you can also remove the extreme values.

vi) Tutorial video link: https://www.youtube.com/watch?v=WrjpO7CmUoQ

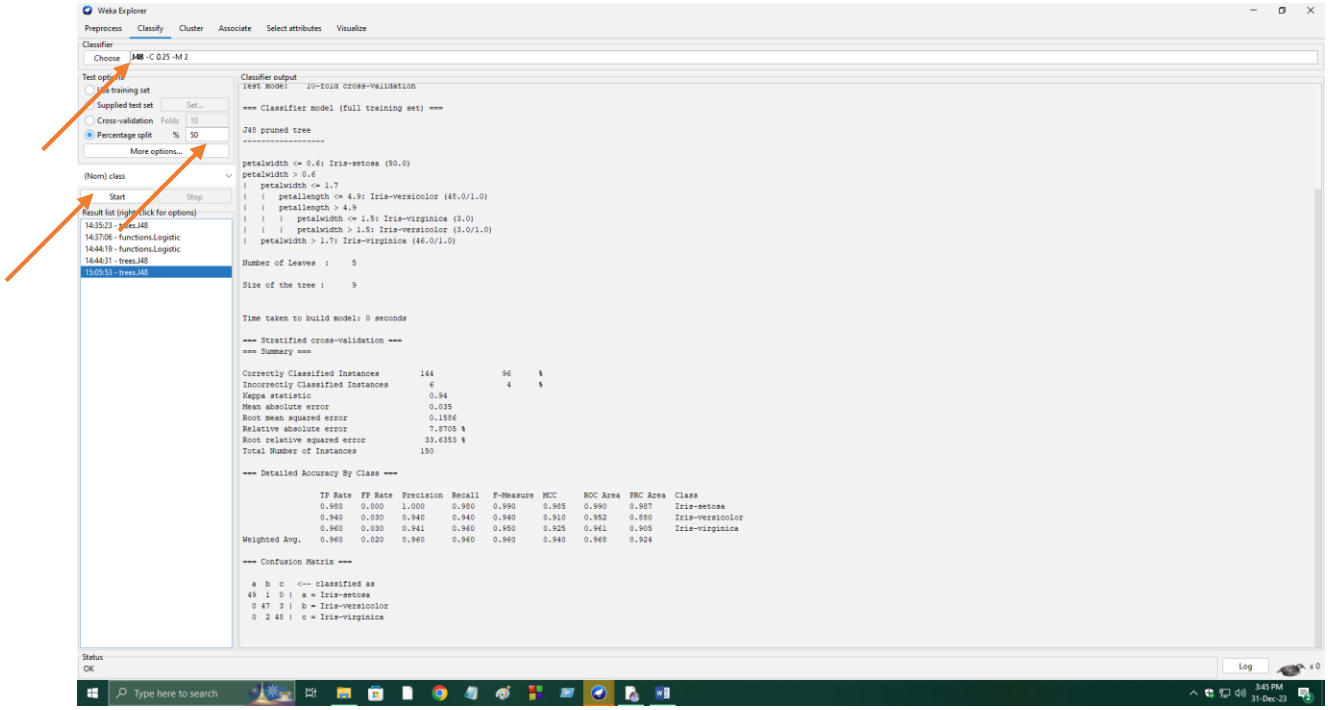
Step 6. Classification using Weka

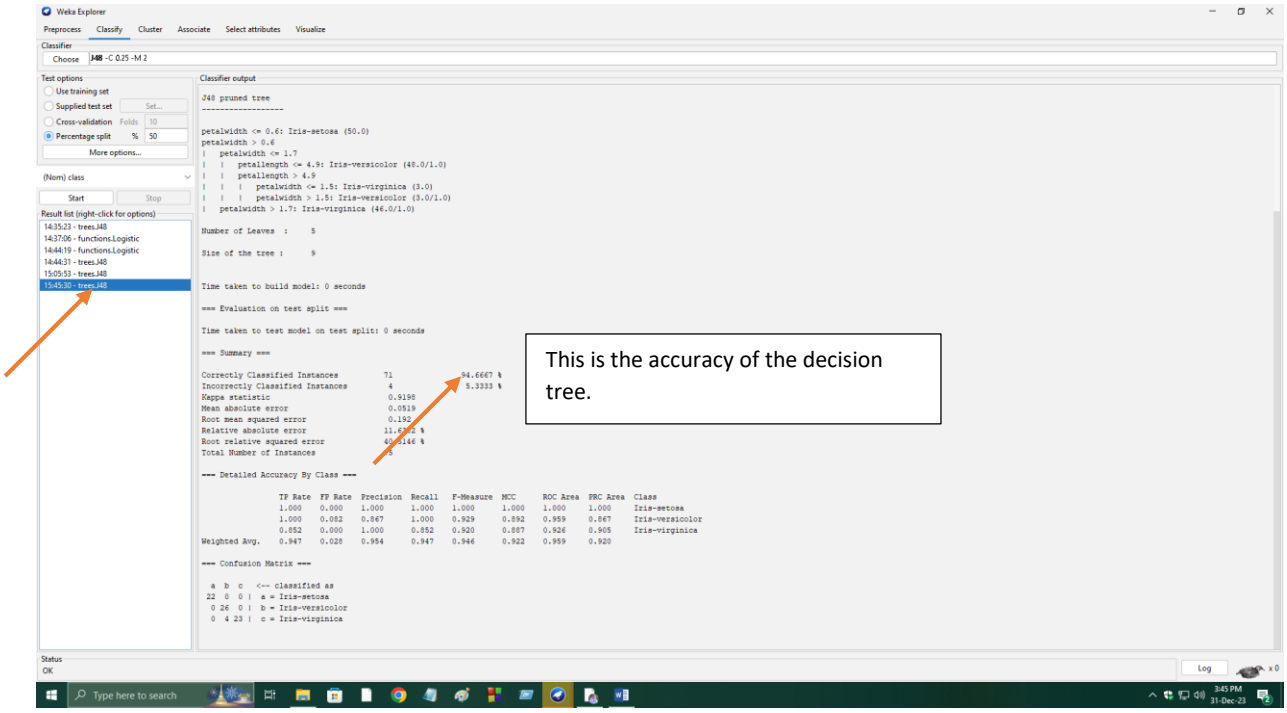
i) Open a file. Here iris.arff is opened.



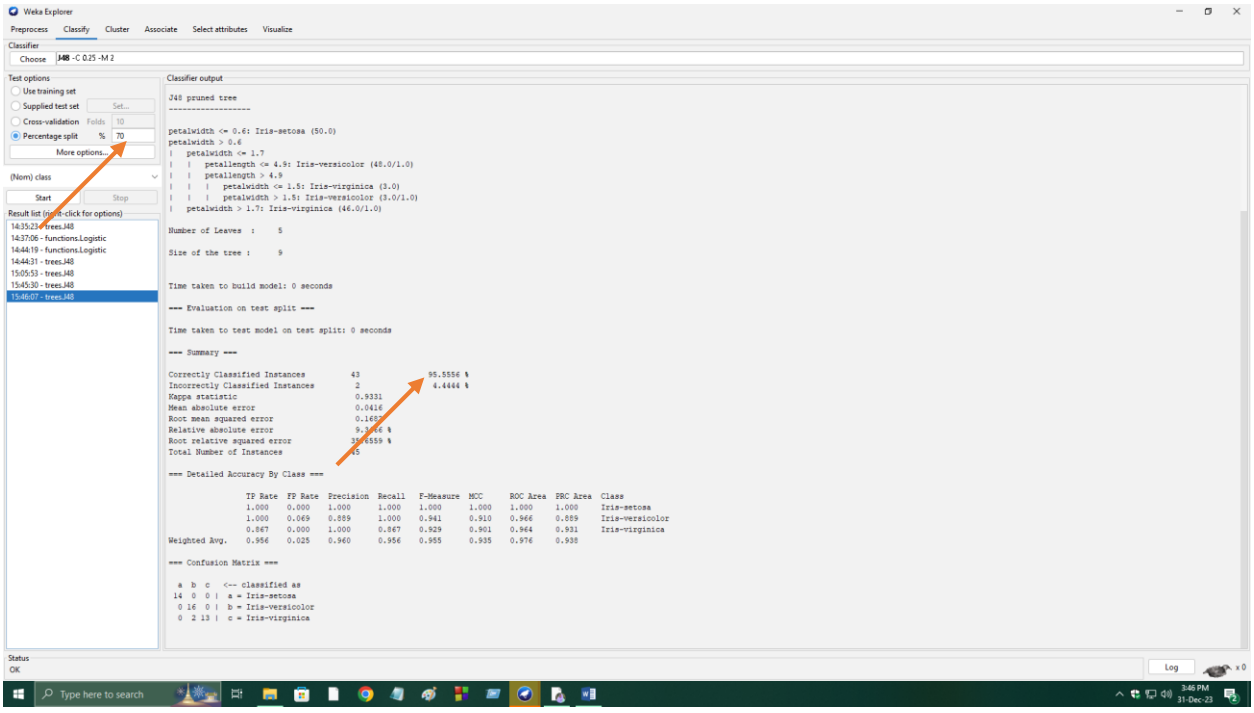
ii) Follow: Classify > Classifier > Choose > tree > J48.

iii) Choose split percentage of your choice. Then click Start.





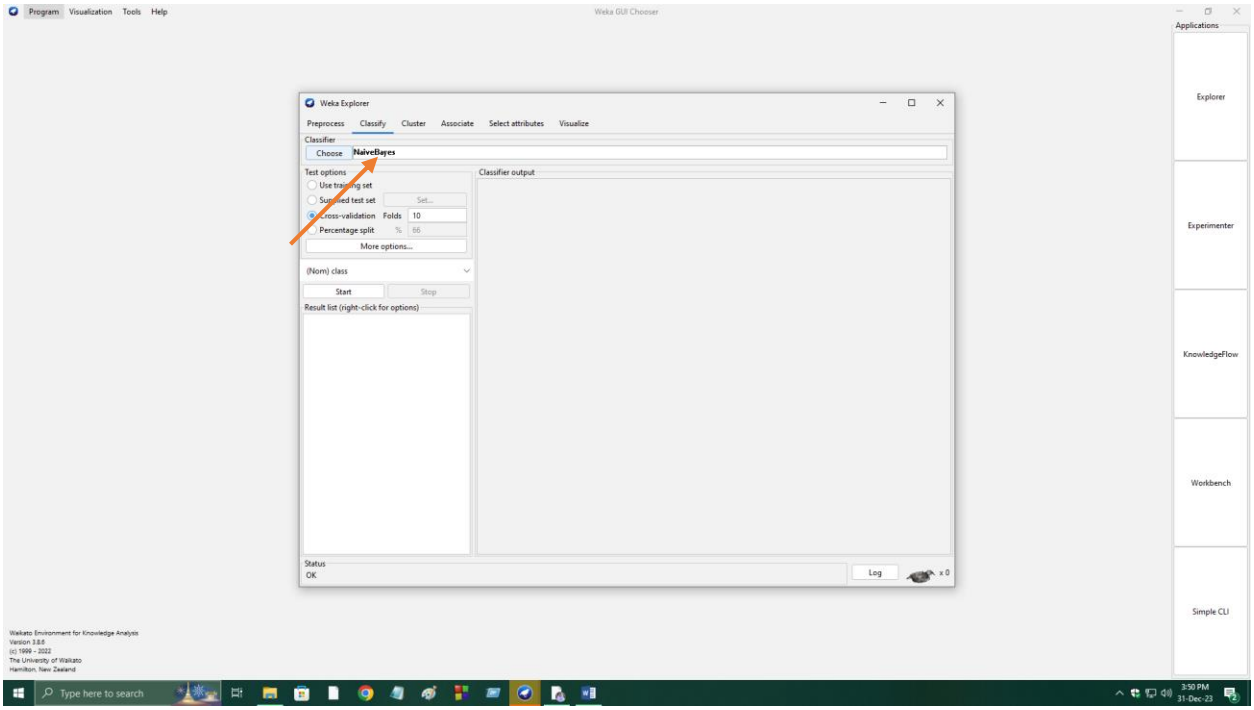
iv) Now change the Percentage split and check the accuracy.



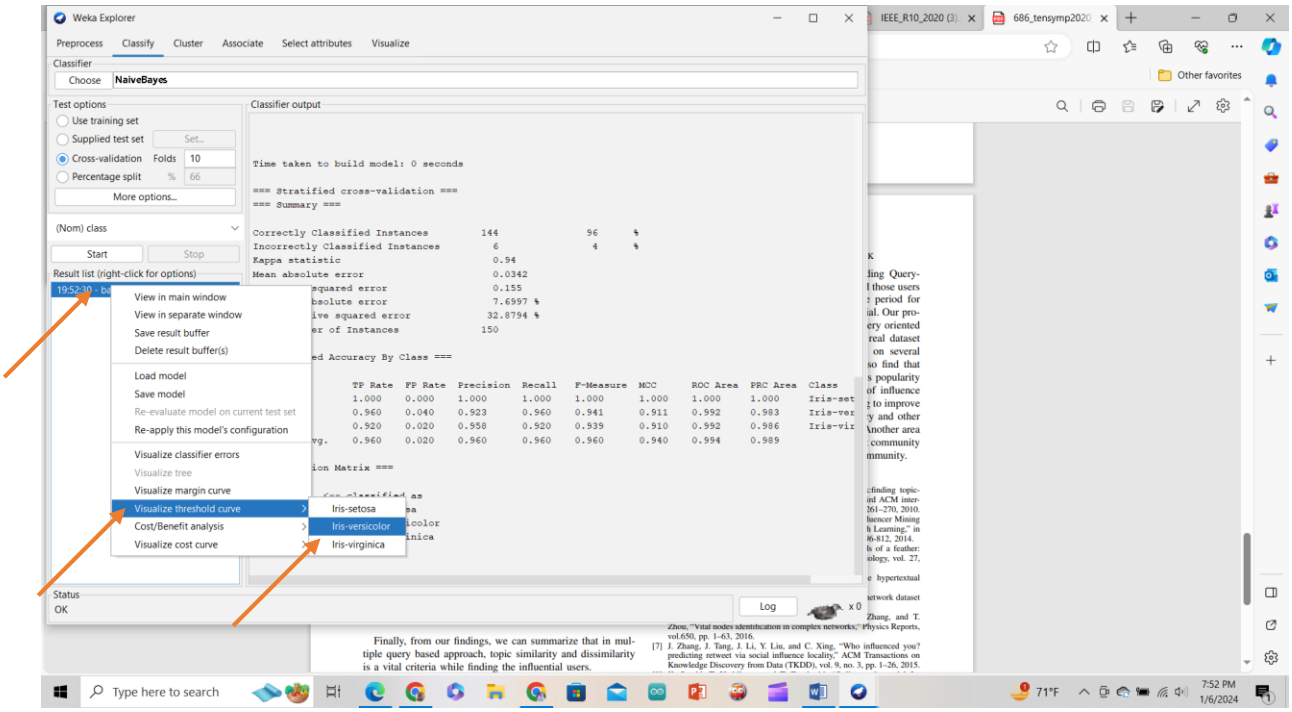
v) Tutorial video link: https://www.youtube.com/watch?v=guf93Gq3wAI

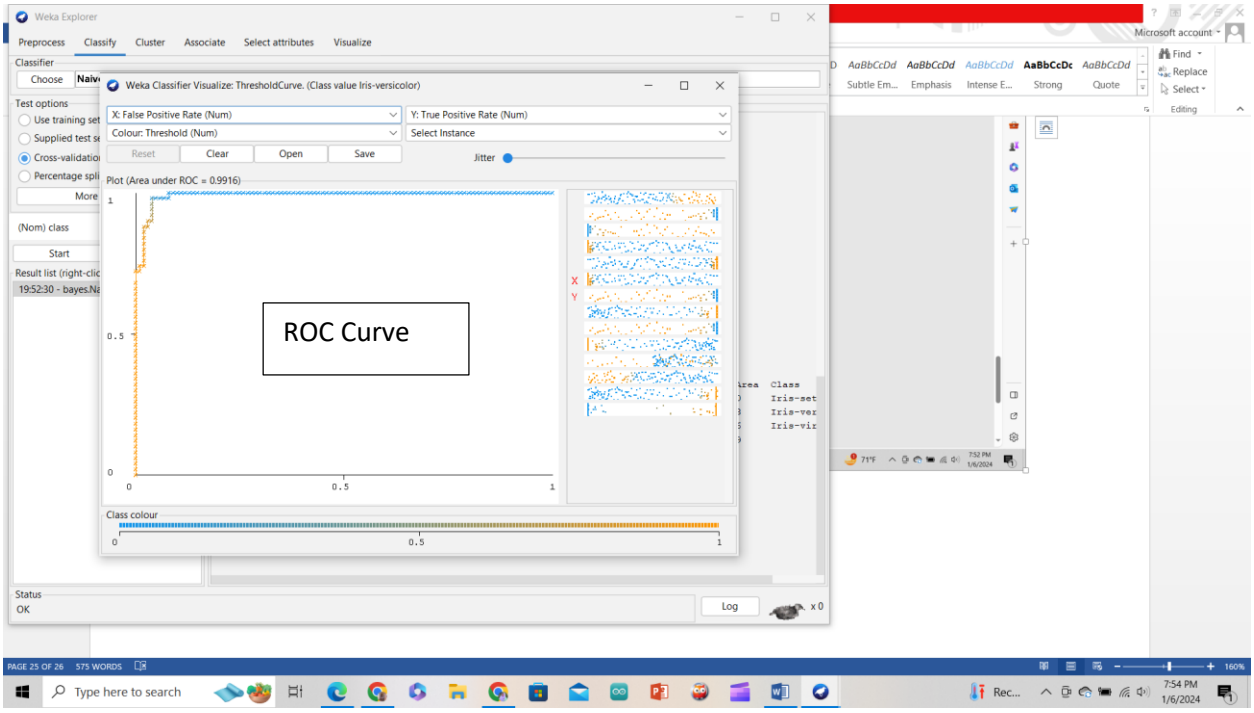
Step 7. Visualize result in Weka

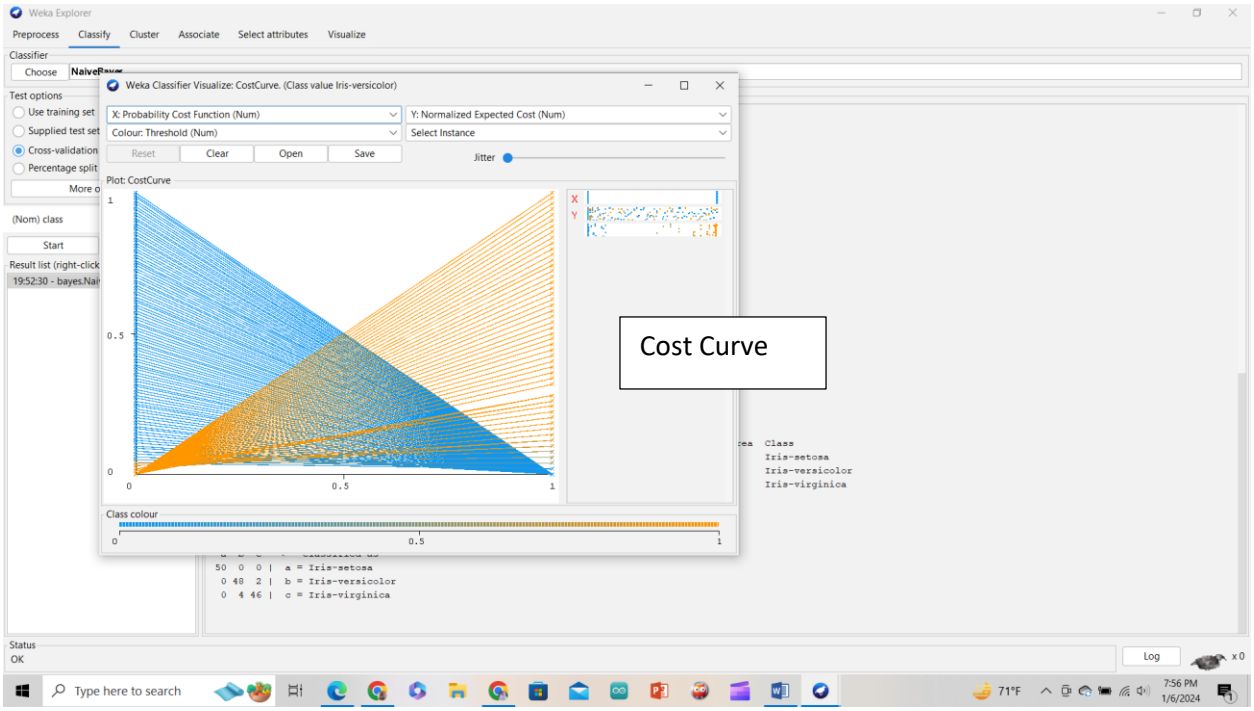
i) Open a file and go to Classify > Classifier > bayes > NaiveBayes > Start.



ii) Follow these steps to view the ROC curve. Also you can select Cost curve.







iii) Tutorial video link: https://www.youtube.com/watch?v=j97h_-b0gvw