

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.

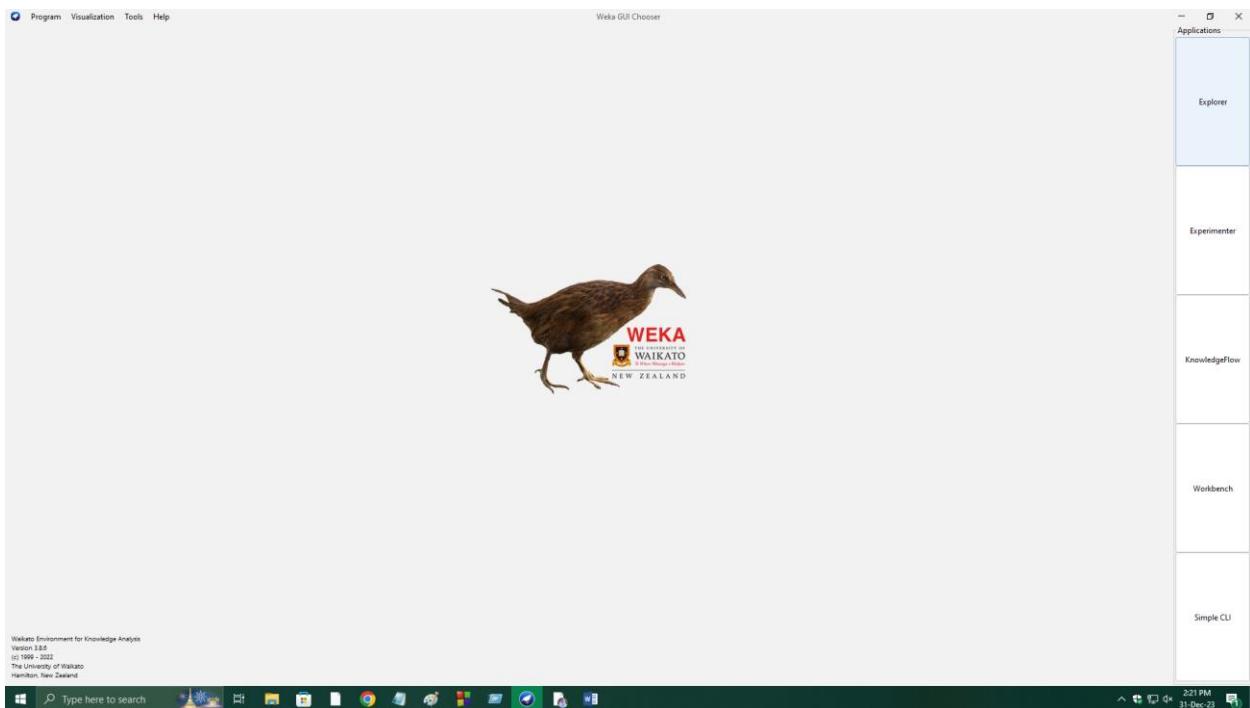


Fig 1: After launching WEKA

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/>

Weka Wiki

Datasets

Table of contents

Miscellaneous collections of datasets

Bioinformatics datasets

Some example datasets for analysis with Weka are included in the Weka distribution and can be found in the data folder of the installed software.

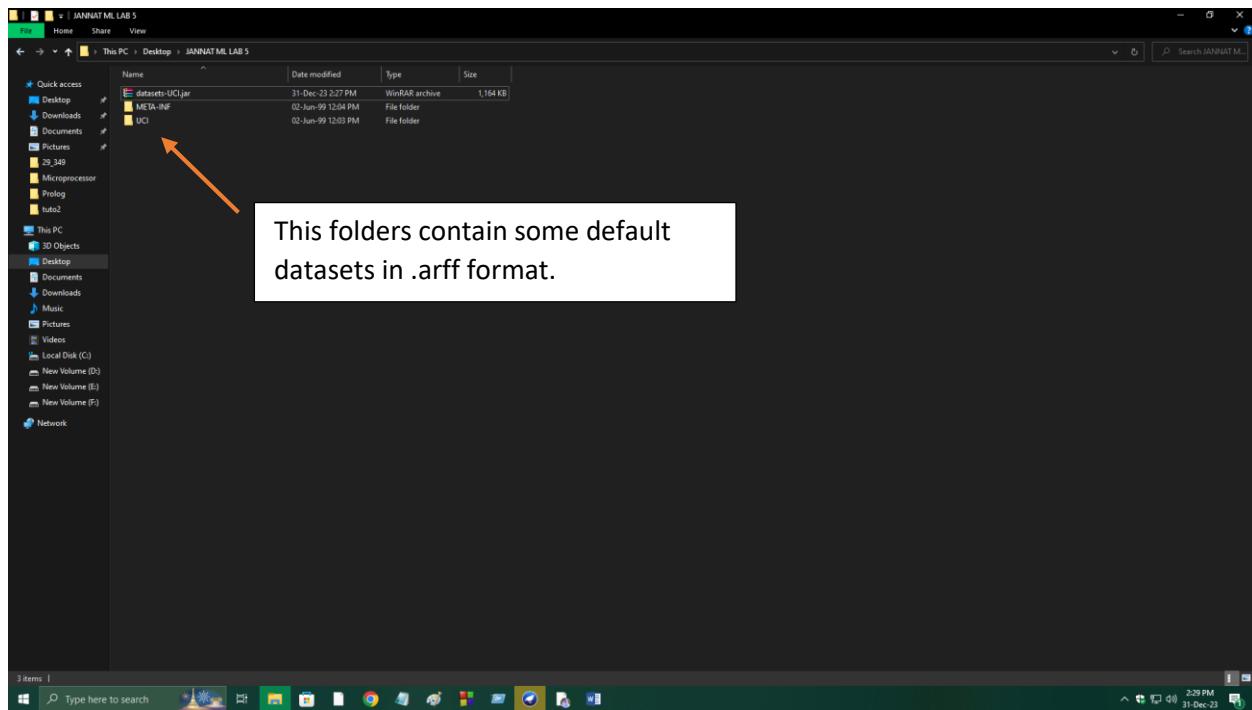
Miscellaneous collections of datasets

- A jarfile containing 37 classification problems originally obtained from the UCI repository of machine learning datasets ([datasets-UCI.jar](#), 1,190,961 Bytes).
- A jarfile containing 37 regression problems obtained from various sources ([datasets-numeric.jar](#), 169,344 Bytes).
- A jarfile containing 6 agricultural datasets obtained from New Zealand ([agrdatasets.jar](#), 31,200 Bytes).
- A jarfile containing 30 regression datasets collected by the University of Wisconsin ([datasets.jar](#), 10,090,266 Bytes).
- A gzipped tar containing UCI ML and UCI KDD datasets ([uci_ml_kdd.tar.gz](#), 1,177,049 Bytes).
- A gzipped tar containing StatLib datasets ([statlib-20050214.tar.gz](#), 12,785,582 Bytes).
- A gzipped tar containing ordinal, real-world datasets donated by Professor Arie Ben David ([datasets-arie_ben_david.tar.gz](#), 11,348 Bytes).
- A zip file containing 19 multi-class (1-of-n) text datasets donated by Dr George Forman ([19MclassTextWc.zip](#), 14,084,828 Bytes).
- A bziped tar file containing the Reuters21578 dataset split into separate files according to the ModApte split ([reuters21578-ModApte.tar.bz2](#), 81,745,032 Bytes).
- A zip file containing 41 drug design datasets formed using the Adriana.Code software donated by Dr Mehmet Fatih Arasyalı ([Drug-datasets.zip](#), 11,376,153 Bytes).

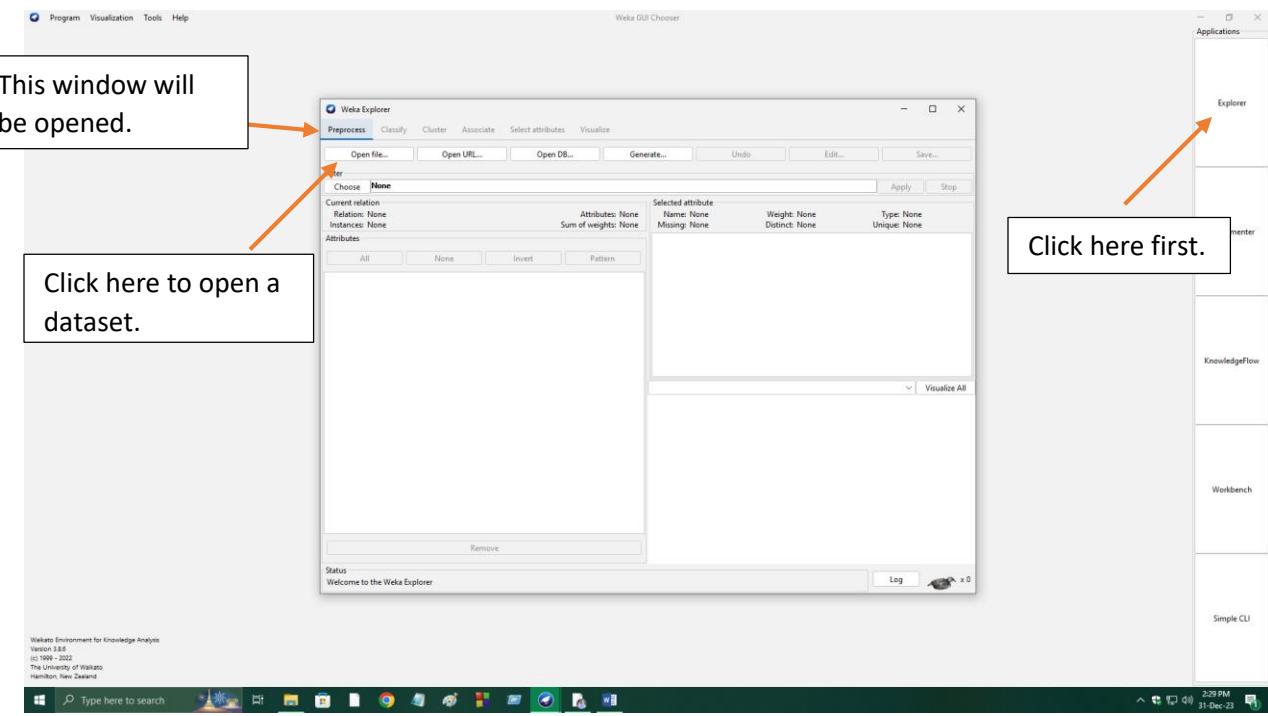
Downloads

This PC > Local Disk (C) > Users > Dell > Downloads

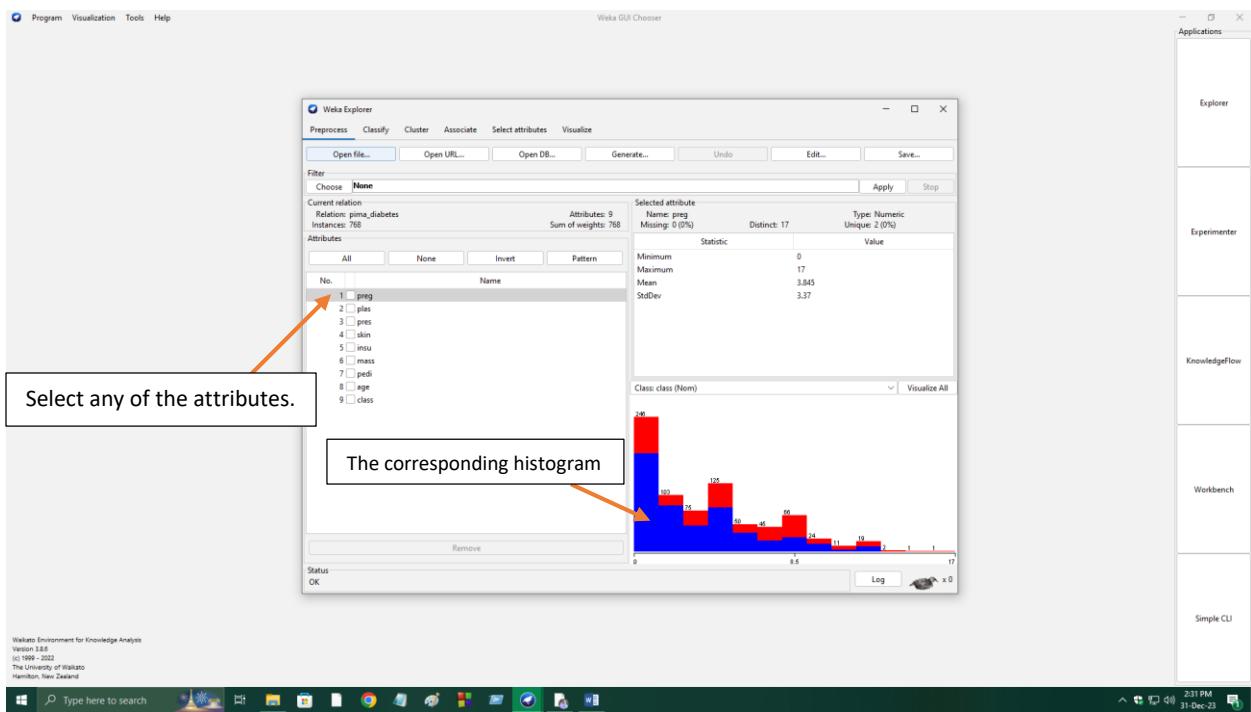
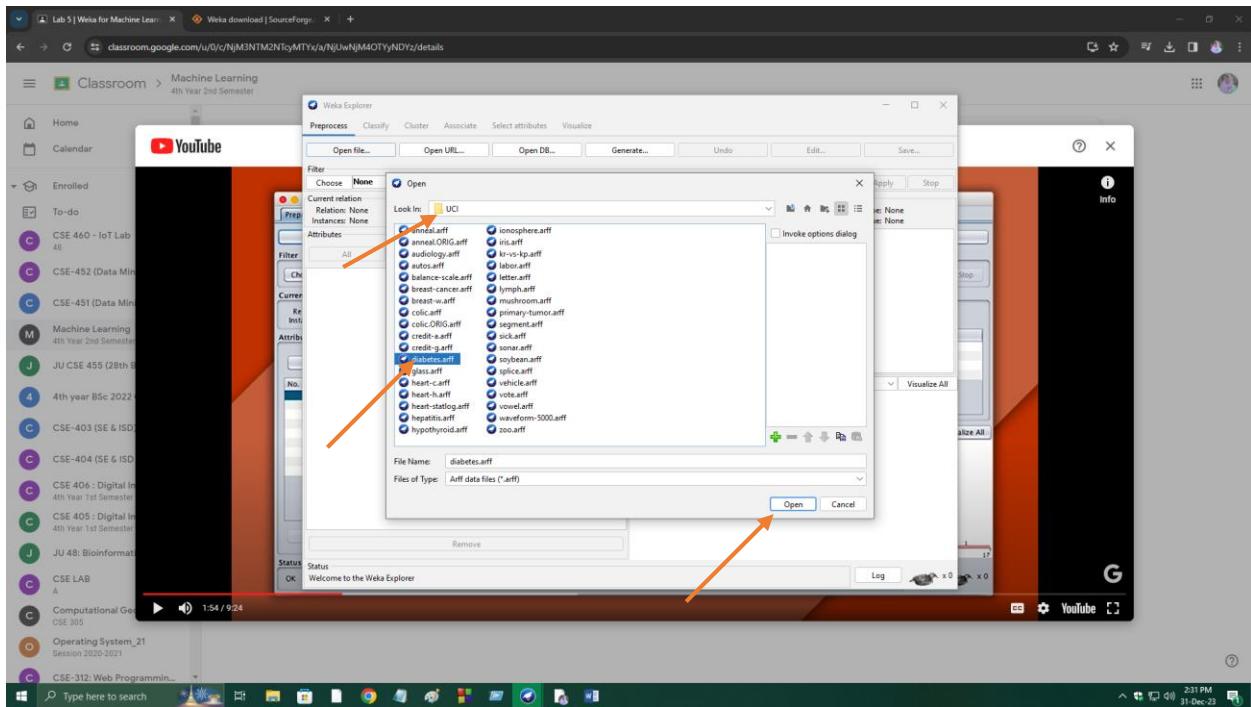
Name	Date modified	Type	Size
datasets-uci.jar	31-Dec-23 2:27 PM	JAR File	1,164 KB
dfp.pdf	20-Dec-23 11:52 AM	PL File	1 KB
dfp.pdf	20-Dec-23 11:52 AM	PL File	1 KB
Desktop Shortcut	20-Dec-23 11:52 AM	Shortcut	1 KB
Lab 8_CSE_360.pdf	20-Dec-23 11:36 AM	Microsoft Word P...	405 KB
LADBP (Telephony).pkt	19-Dec-23 8:36 PM	Microsoft Excel P...	0 KB
image_processing	08-Dec-23 10:45 PM	Folder	0 KB
music (1).zip	05-Dec-23 3:39 PM	Microsoft Edge P...	2,433 KB
Ezp2348_Eva_Grp13_Exp2.pdf	04-Dec-23 9:27 AM	Image	14 KB
matrice.jpg	04-Dec-23 9:27 AM	Microsoft Excel C...	494 KB
spam.csv	03-Dec-23 4:22 PM	Text	0 KB
naive_bayes (1).pynb	03-Dec-23 4:20 PM	Jupyter Source File	62 KB
naive_bayes.ipynb	03-Dec-23 4:20 PM	Jupyter Source File	62 KB
z1.py	03-Dec-23 2:32 PM	PY File	1 KB
1141.docx	30-Nov-23 6:27 PM	Microsoft Word 9...	255 KB
Why Yu-Charles Maruth-Assembly-Lang...	29-Nov-23 8:00 PM	Microsoft Edge P...	14,691 KB
Database.png	26-Nov-23 9:41 AM	PNG File	3 KB
Screencast 2023-11-22 194350.png	22-Nov-23 7:43 PM	PNG File	274 KB
swig-0.9.4.1-x64.exe	22-Nov-23 10:17 AM	Application	12,855 KB
Microprocessors&Microcontroller.jpg	31-Nov-23 9:04 AM	Image	7 KB
adminSetup-2018.msi	16-Nov-23 7:11 PM	Windows Installer ...	35,331 KB
u8BuildTools.exe	13-Nov-23 23:28 PM	Application	3,859 KB
rules.sql	13-Nov-23 6:45 PM	SQL Text File	9 KB
Postman-win64-Setup.exe	13-Nov-23 5:33 PM	Application	128,345 KB
databaseChange.png	13-Nov-23 4:43 PM	PNG File	283 KB
Earlier this year (22)			
-1 syllabus.pdf	20-Sep-23 7:45 PM	Microsoft Edge P...	205 KB
ch11.pdf	19-Sep-23 4:21 PM	Microsoft PowerP...	2,627 KB
project_C_Report(roup-12).pdf	10-Sep-23 2:27 PM	Microsoft Edge P...	1,087 KB
graphics-questions(26-27).pdf	10-Sep-23 2:46 AM	Microsoft Edge P...	4,159 KB
sead_3_R more_on_props.pdf	12-Aug-23 10:44 PM	Microsoft Edge P...	349 KB
exact_1.pdf	20-Jun-23 6:17 PM	Microsoft Edge P...	25 KB
exact_1.pdf	18-Jun-23 12:51 PM	C++ source file	5 KB
exact_1.pdf	18-Jun-23 13:20 PM	C++ source file	1 KB
exact_1_SlideChapter03Full.pdf	09-Jun-23 10:12 AM	Microsoft PowerP...	2,702 KB
exact_1_SlideChapter01Full.pdf	09-Jun-23 9:32 AM	Microsoft PowerP...	1,355 KB
Class Routine drawing.pdf	08-Jun-23 12:37 PM	Microsoft Edge P...	59 KB

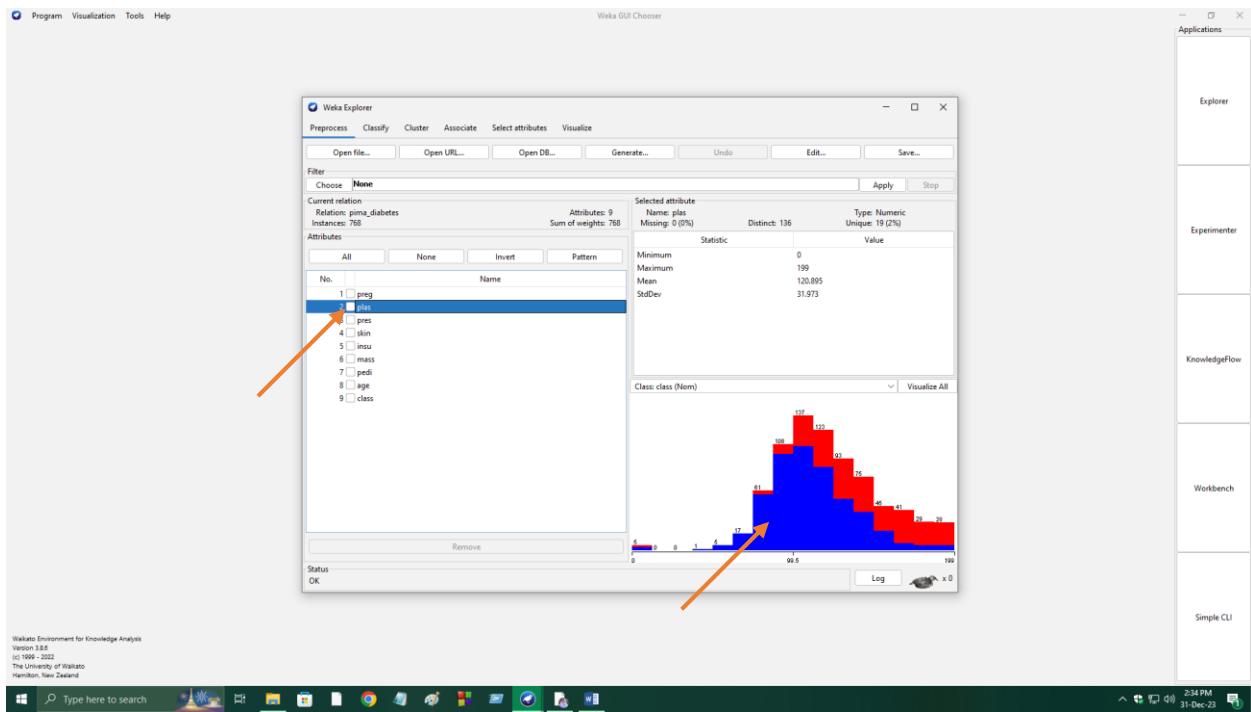


ii) Now launch Weka and click on explorer.

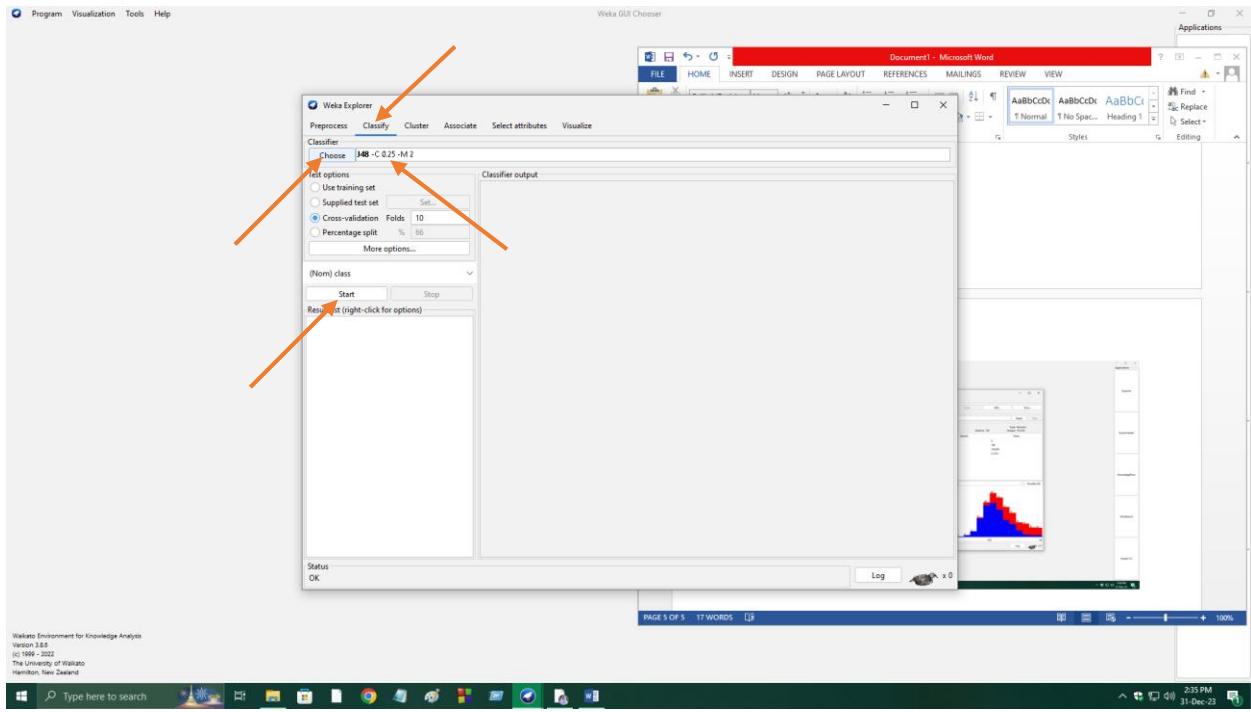


iii) Select diabetes.arff file to open. (You can open any file)





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



```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier Choose J48 -C 0.25 -M 2
Test options
  Use training set
  Supplied test set Set...
  Cross-validation Folds 10
  Percentage split % 66
  More options...
(Nom) class
Start Stop
Result list (right-click for options)
14.552 - trees.J48
Classifier output
|   |   age <= 25: tested_negative (4.0)
|   |   age > 25:
|   |       |   pres <= 41
|   |       |   pres > 41:
|   |           |   mass <= 27.1: tested_positive (12.0/1.0)
|   |           |   mass > 27.1:
|   |               |   pres <= 82
|   |               |   pres > 82: tested_negative (4.0)
|   |               |   pres > 82: tested_positive (8.0/1.0)
|   |               |   age > 61: tested_negative (4.0)
|   |               |   age > 61: tested_positive (4.0)
|   |       mass > 29.2:
|   |           pres <= 157
|   |           pres > 157: tested_positive (15.0/1.0)
|   |       pres <= 61
|   |           |   age <= 30: tested_negative (40.0/13.0)
|   |           |   age > 30: tested_positive (60.0/17.0)
|   |           |   pres > 157: tested_positive (92.0/12.0)
Number of Leaves : 20
Size of the tree : 39
Time taken to build model: 0.01 seconds
--- Stratified cross-validation ---
--- Summary ---
Correctly Classified Instances      567          73.6283 %
Incorrectly Classified Instances    201          26.1719 %
Root mean square error            0.4164
Mean absolute error                0.3155
Root mean squared error           0.4463
Relative absolute error            69.4941 %
Root relative squared error        93.6293 %
Total Number of Instances         768
--- Detailed Accuracy By Class ---
          TP Rate FP Rate Precision Recall F-Measure MCC ROC Area EER Area Class
          0.814   0.403   0.790   0.814   0.802   0.417   0.751   0.611   tested_negative
          0.597   0.186   0.632   0.597   0.414   0.417   0.751   0.572   tested_positive
Weighted Avg.                      0.738   0.327   0.735   0.738   0.736   0.417   0.751   0.727
--- Confusion Matrix ---
a b  --> classified as
407 93  |  a = tested_negative
108 160 |  b = tested_positive

```

The decision tree is created. Remember the accuracy.

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

```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier Choose Logistic -R 1.0E-8 -M 1 -num-decimal-places:4
Test options
  Use training set
  Supplied test set Set...
  Cross-validation Folds 10
  Percentage split % 66
  More options...
(Nom) class
Start Stop
Result list (right-click for options)
14.552 - trees.J48
14.570 - functions.Logistic
Classifier output
pres      0.0133
skin     -0.0006
liso      0.0112
mass     -0.0097
pedi     -0.9452
age      -0.0149
Intercept 0.4047

Odds Ratios...
Variable   Class
preg      tested_negative
pres      tested_positive
plas      tested_negative
pres      tested_positive
liso      tested_negative
mass      tested_negative
pedi      tested_positive
age      tested_negative

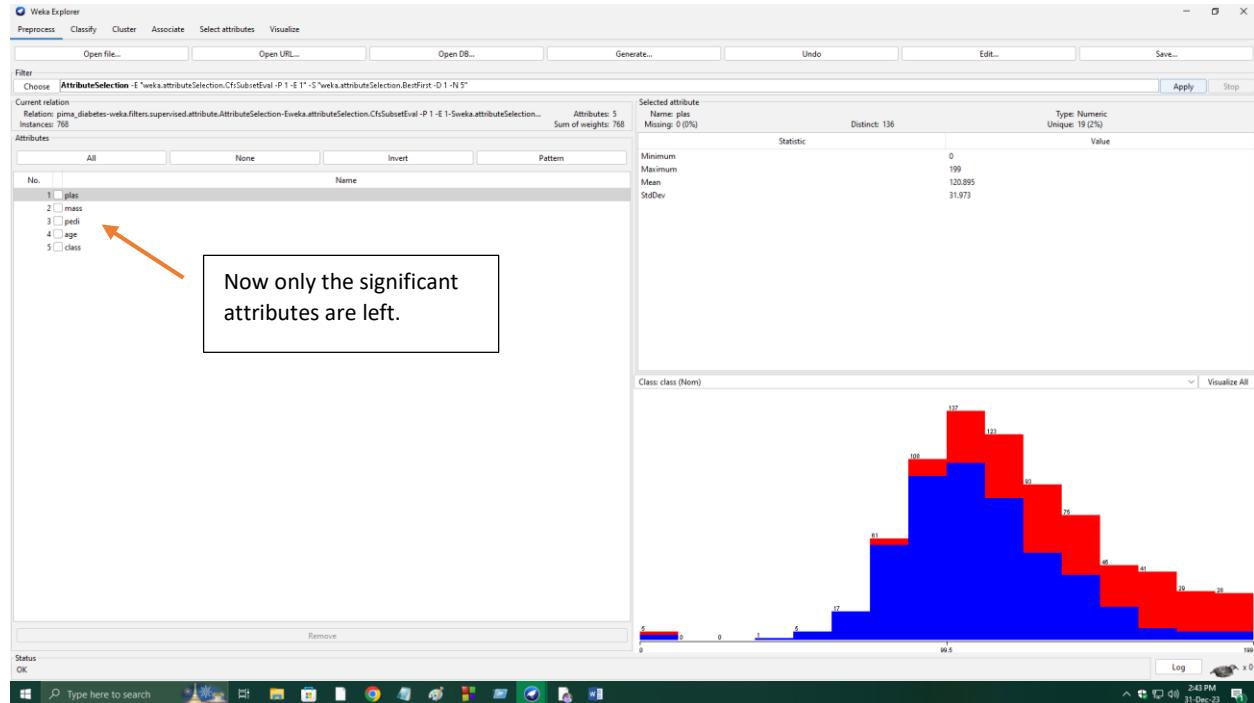
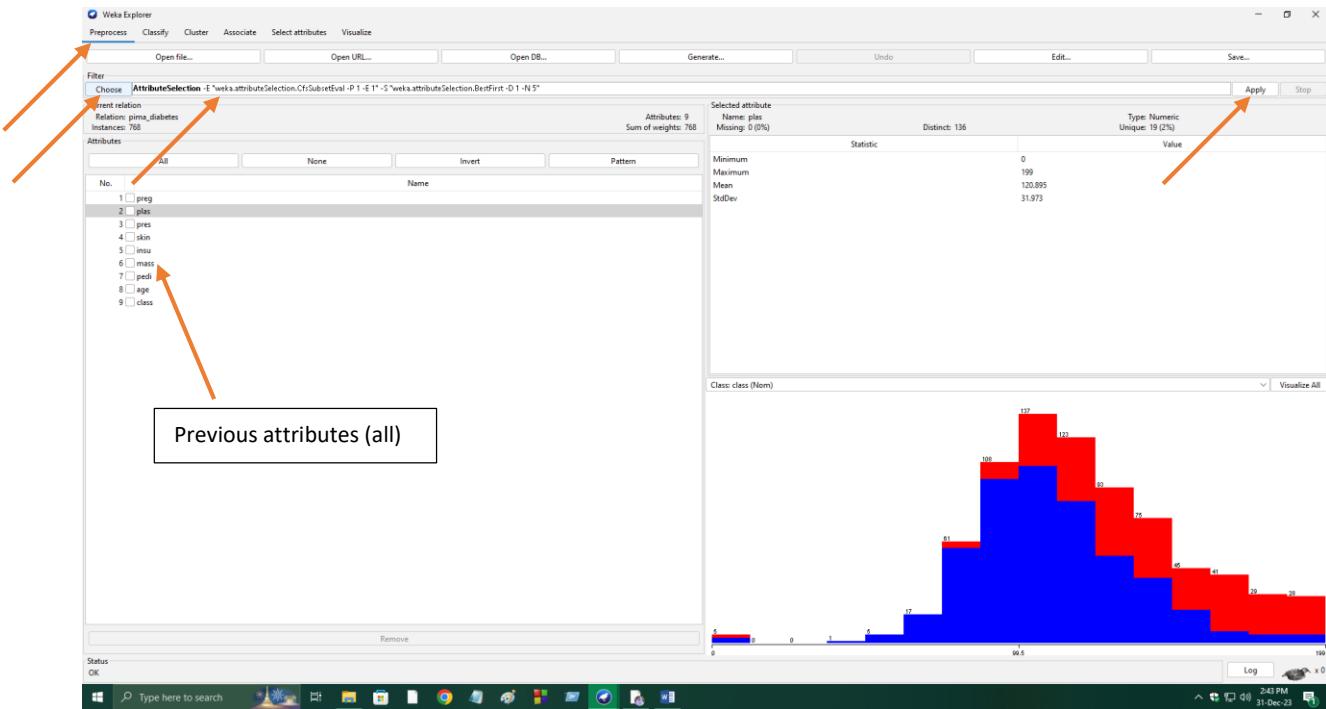
Time taken to build model: 0.03 seconds
--- Stratified cross-validation ---
--- Summary ---
Correctly Classified Instances      593          77.2135 %
Incorrectly Classified Instances    175          22.7865 %
Root mean square error            0.4734
Mean absolute error                0.3954
Root mean squared error           0.6058 %
Relative absolute error            68.0518 %
Root relative squared error        82.9651 %
Total Number of Instances         768
--- Detailed Accuracy By Class ---
          TP Rate FP Rate Precision Recall F-Measure MCC ROC Area EER Area Class
          0.880   0.429   0.793   0.880   0.834   0.480   0.832   0.682   tested_negative
          0.571   0.120   0.718   0.571   0.436   0.480   0.832   0.715   tested_positive
Weighted Avg.                      0.772   0.321   0.767   0.772   0.765   0.480   0.832   0.681
--- Confusion Matrix ---
a b  --> classified as
440 60  |  a = tested_negative
115 153 |  b = tested_positive

```

You can toggle these two

The tree is created.

- 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.

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier Choose **J48 - C 0.25 - M 2**

Test options

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 65
- [More options...](#)

(Nom) class

Start Stop

Result list (right-click for options)

- 143523 - trees.J48
- 143706 - functions.Logistic
- 144419 - functions.Logistic
- 144431 - trees.J48

```

Classifier output
1 - |> plas > 145: tested_negative (41.0/6.0)
   |   plas > 145
   |   |   age <= 25: tested_negative (4.0)
   |   |   |   age > 25
   |   |   |   |   age <= 41:
   |   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
   |   |   |   |   |   |   mass > 27.1
   |   |   |   |   |   |   |   pedi <= 0.396: tested_positive (11.0/4.0)
   |   |   |   |   |   |   |   pedi > 0.396: tested_negative (4.0)
   |   |   |   |   |   |   age > 41: tested_negative (4.0)
   |   |   mass > 29.9
   |   |   |   plas <= 157
   |   |   |   |   pedi <= 0.427
   |   |   |   |   |   mass <= 45.5: tested_negative (50.0/21.0)
   |   |   |   |   |   |   mass > 45.5: tested_positive (7.0)
   |   |   |   |   |   pedi > 0.427: tested_positive (58.0/16.0)
   |   |   |   plas > 157: tested_positive (92.0/12.0)

Number of Leaves : 15
Size of the tree : 29

Time taken to build model: 0 seconds

*** Stratified cross-validation ***
*** Summary ***

Correctly Classified Instances      575          74.0698 %
Incorrectly Classified Instances    193          25.1302 %

Kappa statistic                      0.4245
Standard Error                      0.0556
Root mean absolute error            0.4216
Relative absolute error              69.4357 %
Root relative squared error         88.4504 %
Total Number of Instances           768

*** Detailed Accuracy By Class ***



|               | TP Rate | FP Rate | Precision | Recall | F-Measure | NCC   | ROC Area | FRC Area | Class           |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-----------------|
| a             | 0.052   | 0.444   | 0.762     | 0.052  | 0.428     | 0.791 | 0.849    | 0.849    | tested_negative |
| b             | 0.785   | 0.147   | 0.866     | 0.596  | 0.807     | 0.428 | 0.791    | 0.777    | tested_positive |
| Weighted Avg. | 0.749   | 0.341   | 0.742     | 0.749  | 0.743     | 0.428 | 0.791    | 0.777    |                 |



*** Confusion Matrix ***



|   | a   | b   |
|---|-----|-----|
| a | 426 | 74  |
| b | 119 | 149 |


```

Status OK

Type here to search

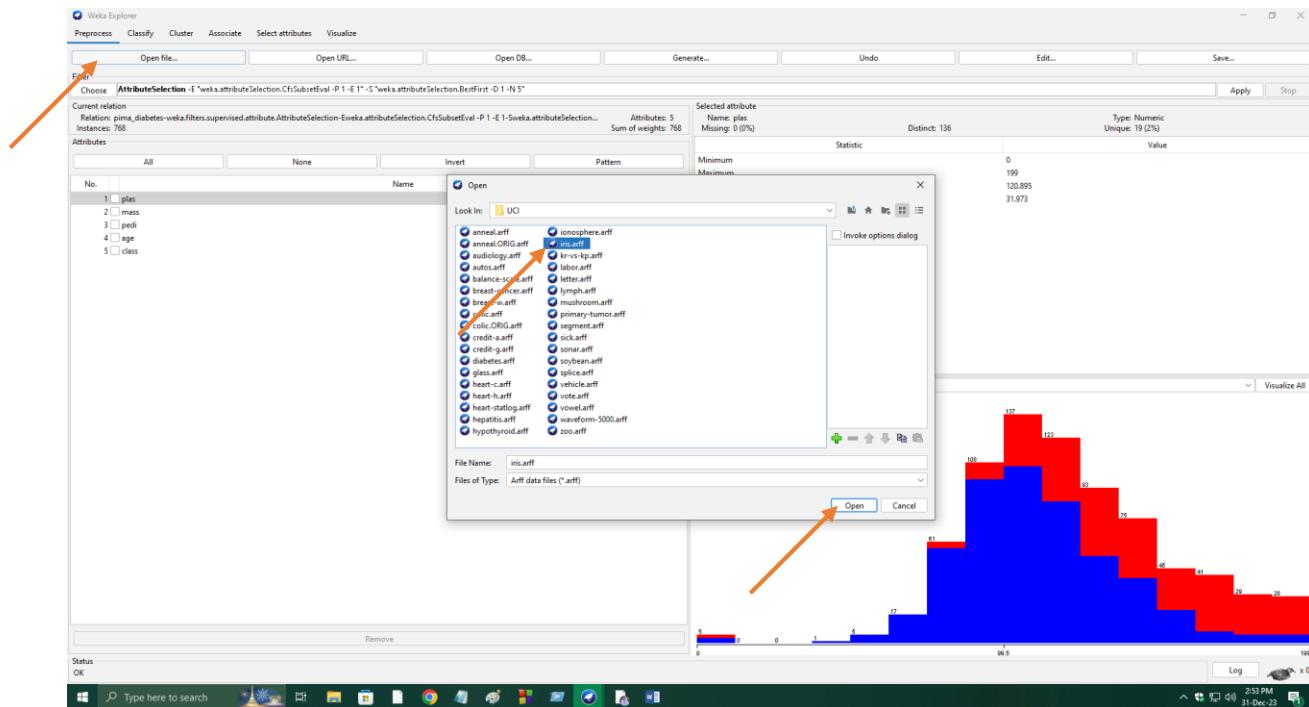
Log x 0

2:44 PM 31-Dec-23

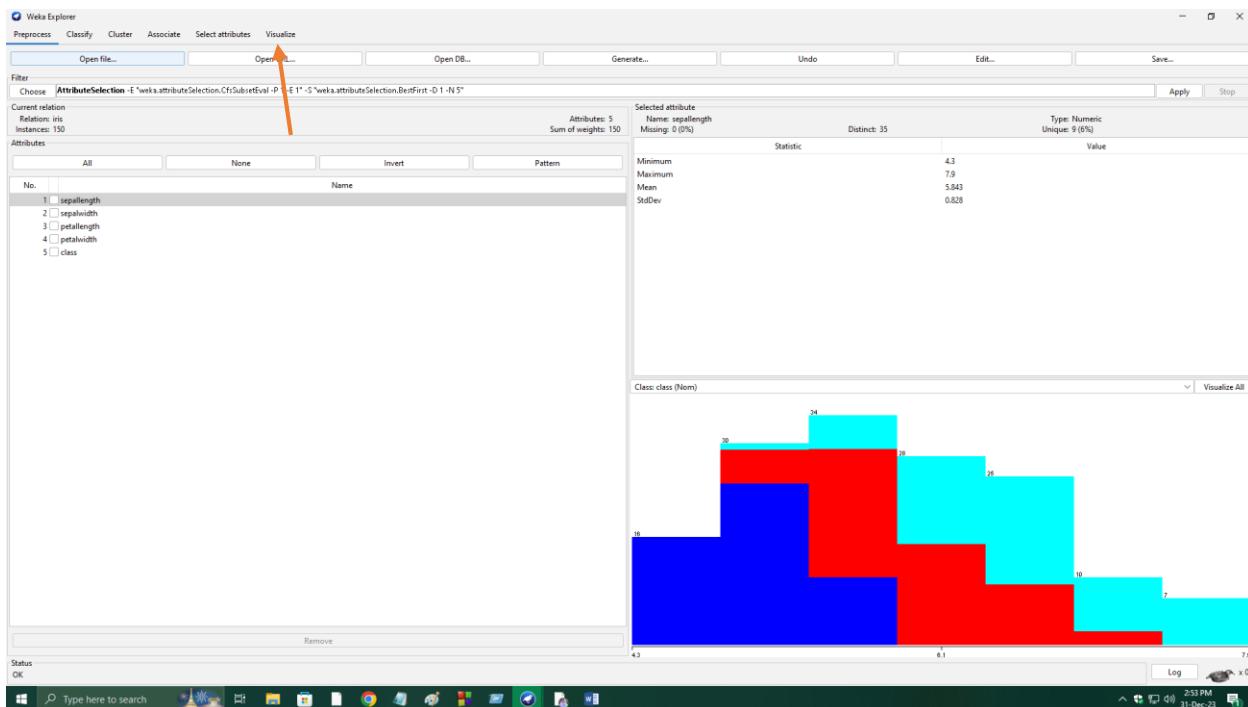
viii) Tutorial video link: <https://www.youtube.com/watch?v=TF1yh5PKaql>

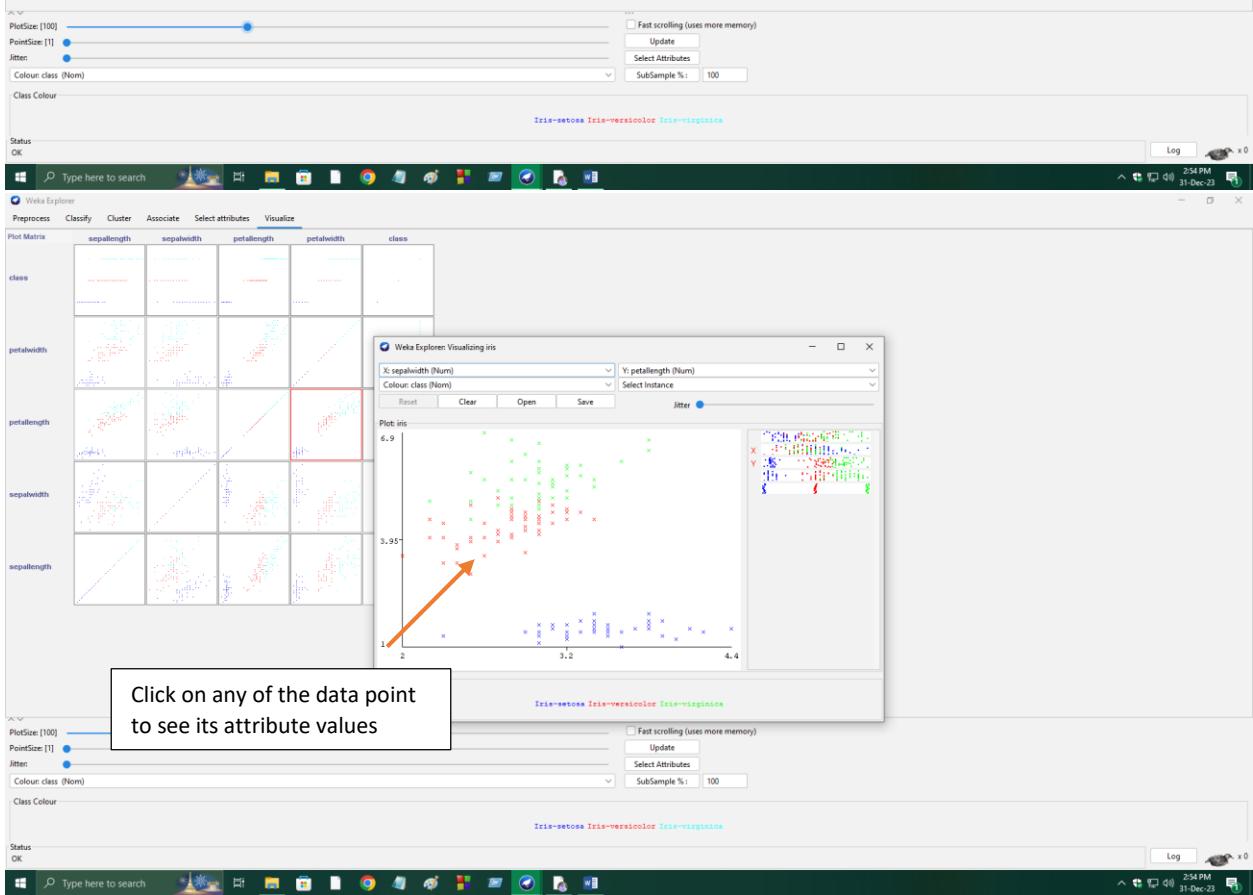
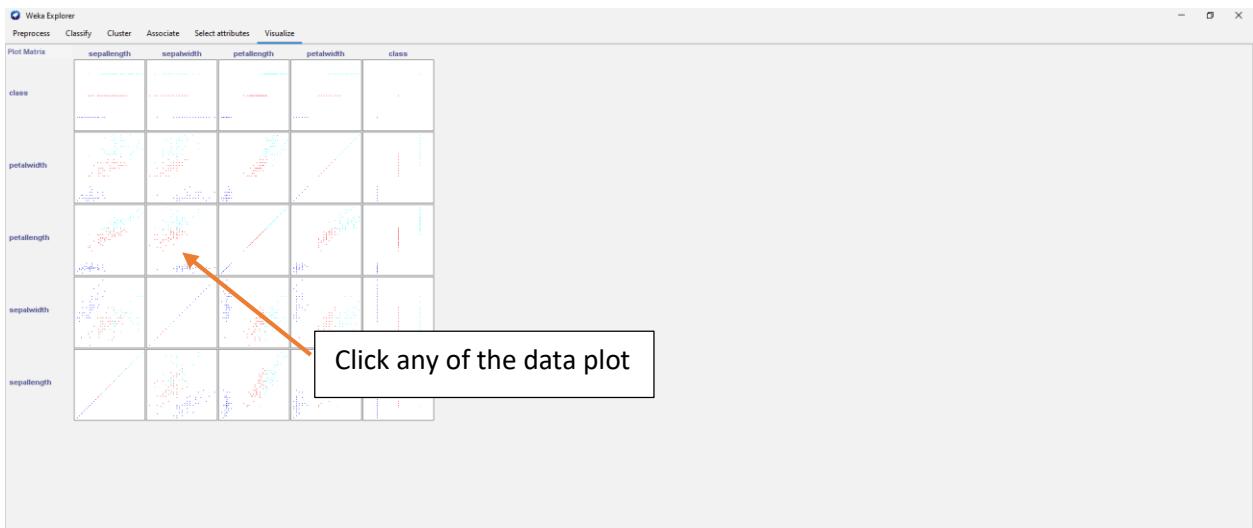
Step 3 Visualize your data from the Weka Dataset

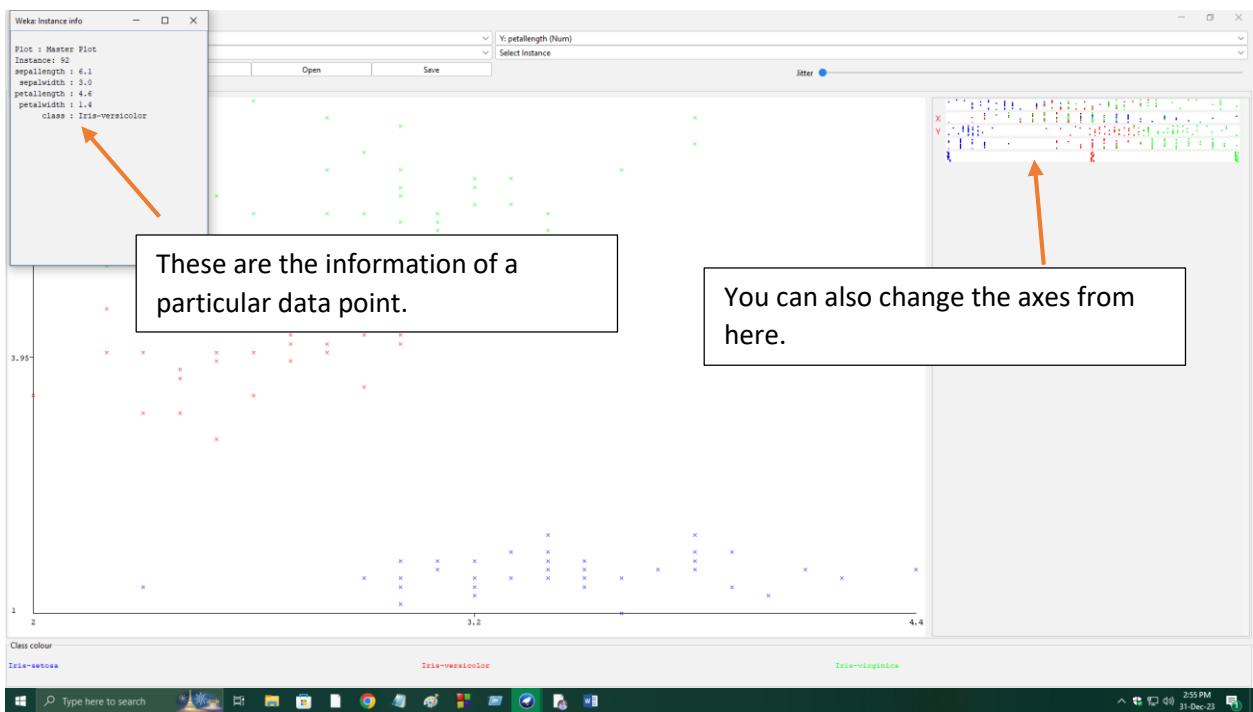
- Open a dataset. Here we've opened iris.arff.



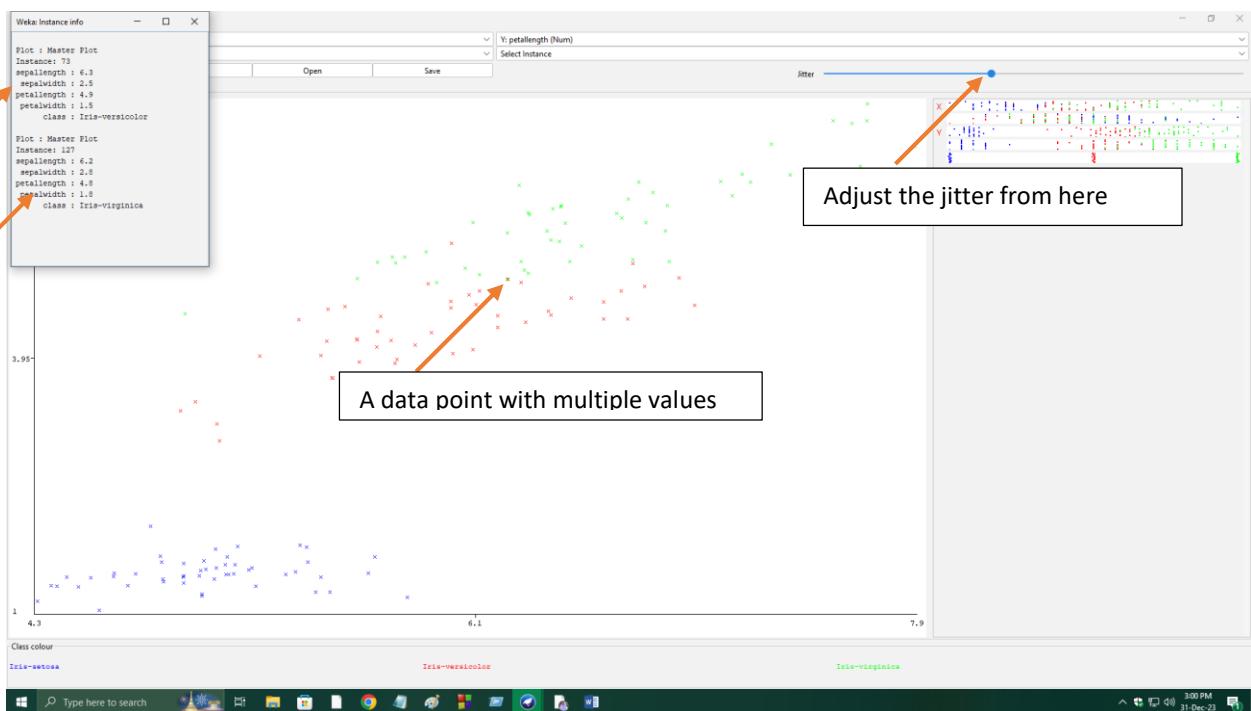
- Click on visualize.

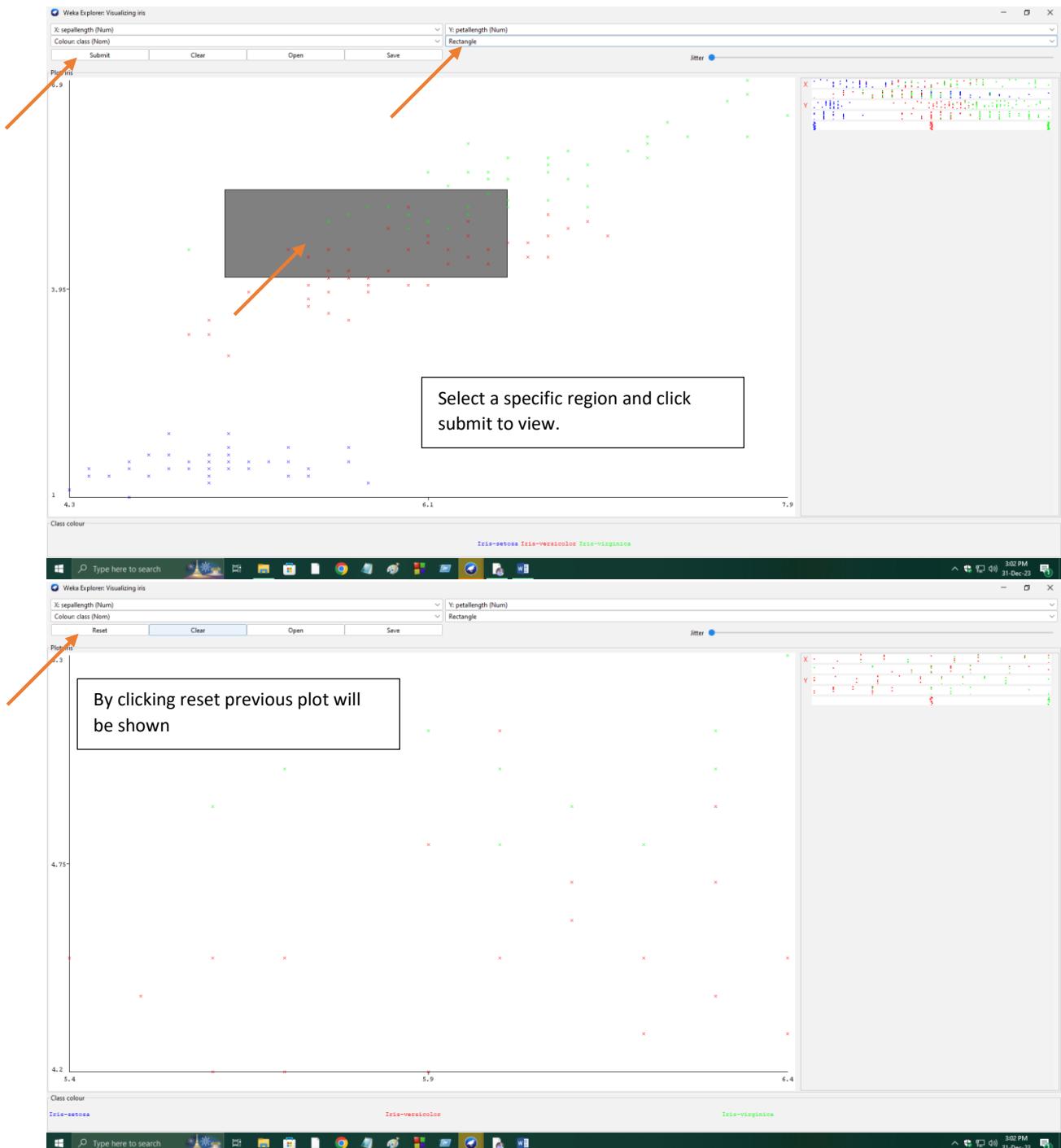




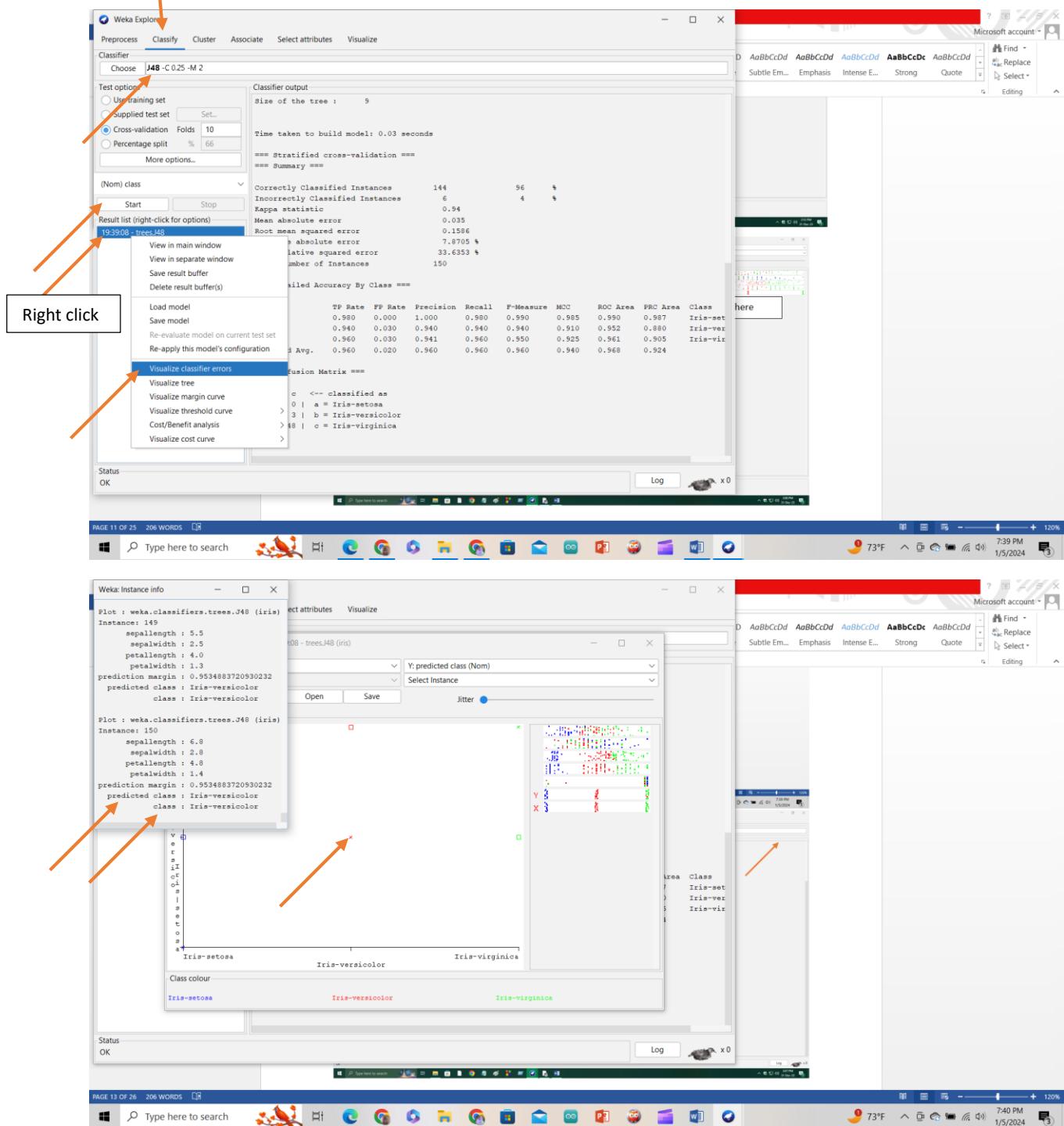


You can also change the axes from here.

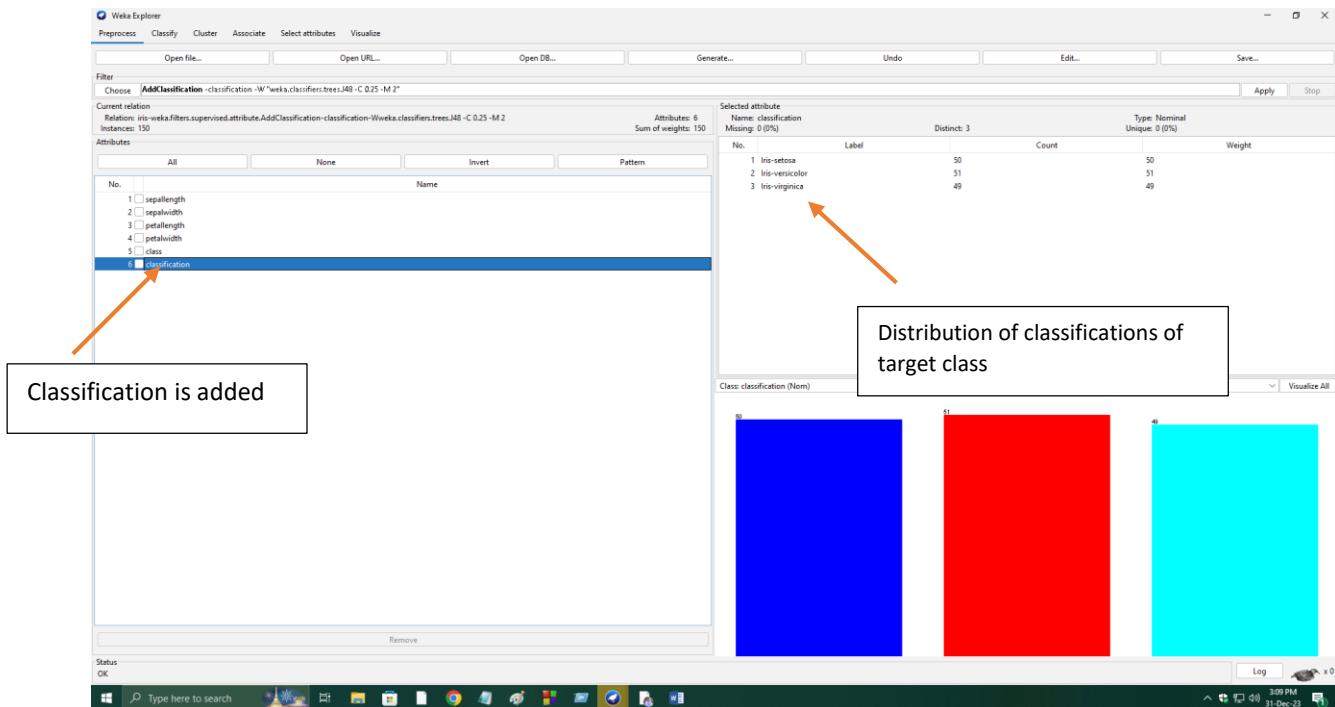
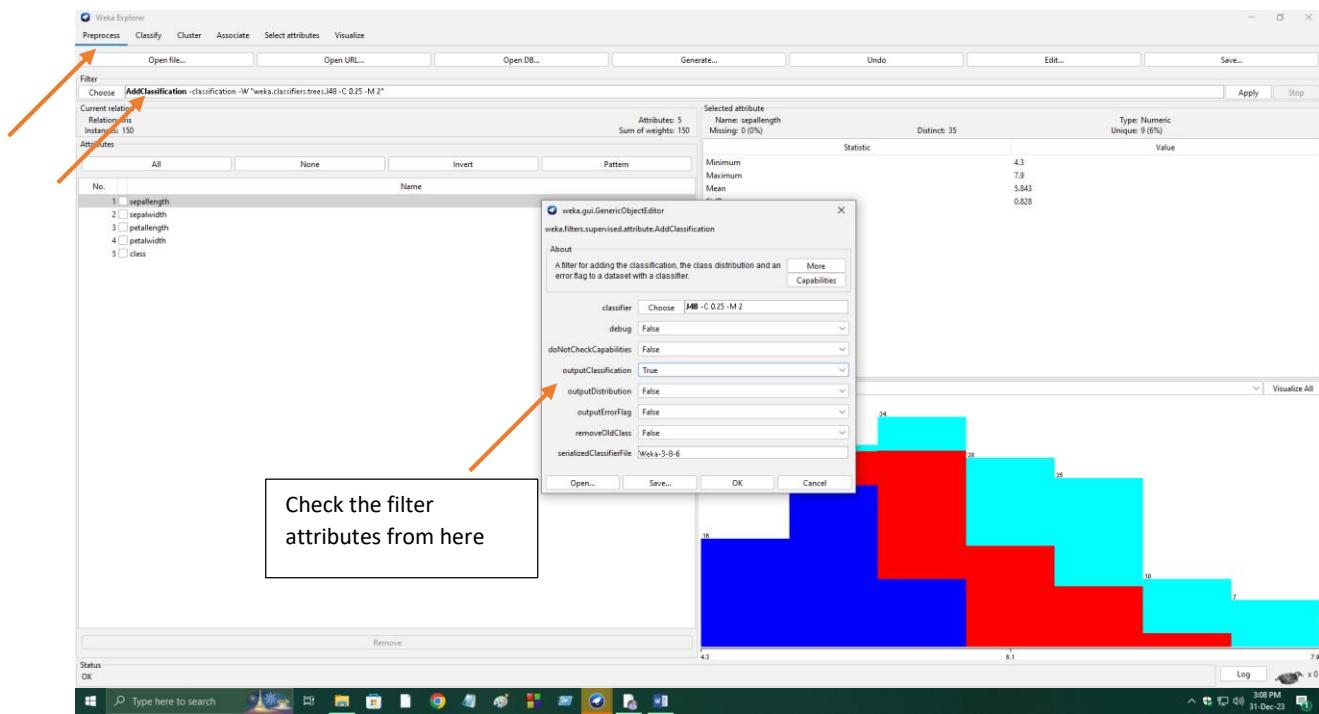




iii) Now build a J48 decision tree and visualize the errors of detecting target class.



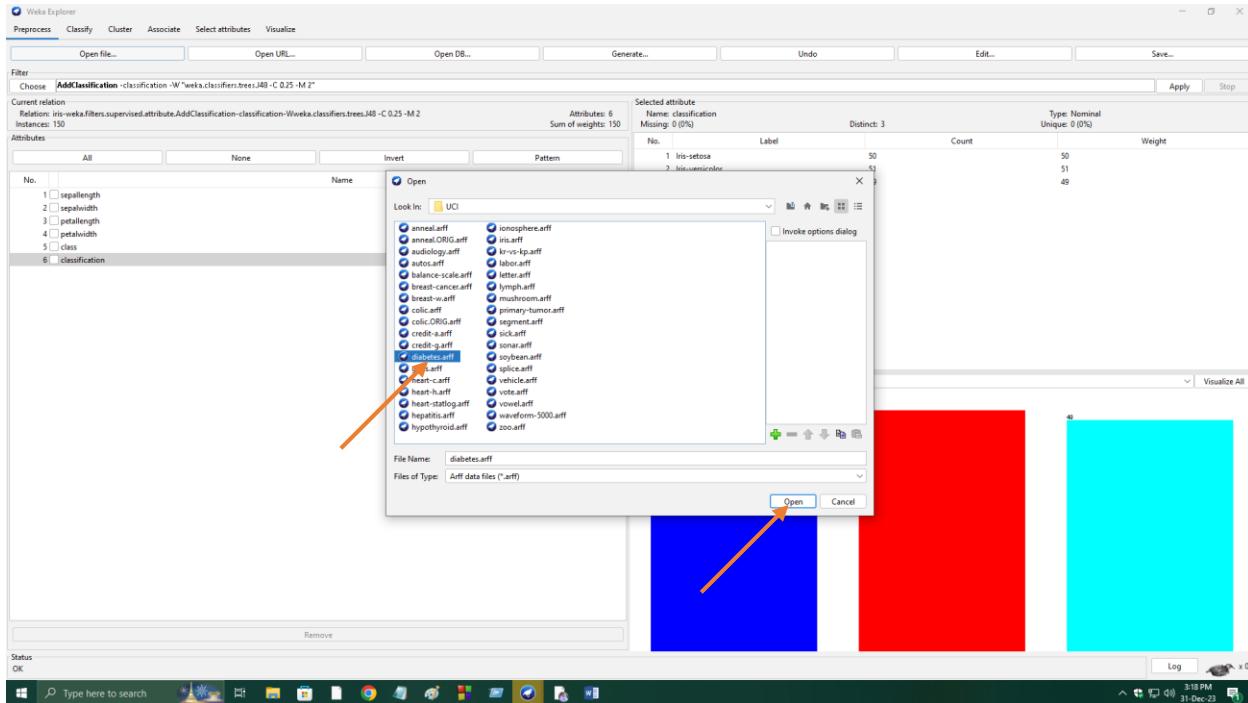
- iv) Now add classification by going: Preprocess > filter > Choose > unsupervised> attribute > AddClassification.



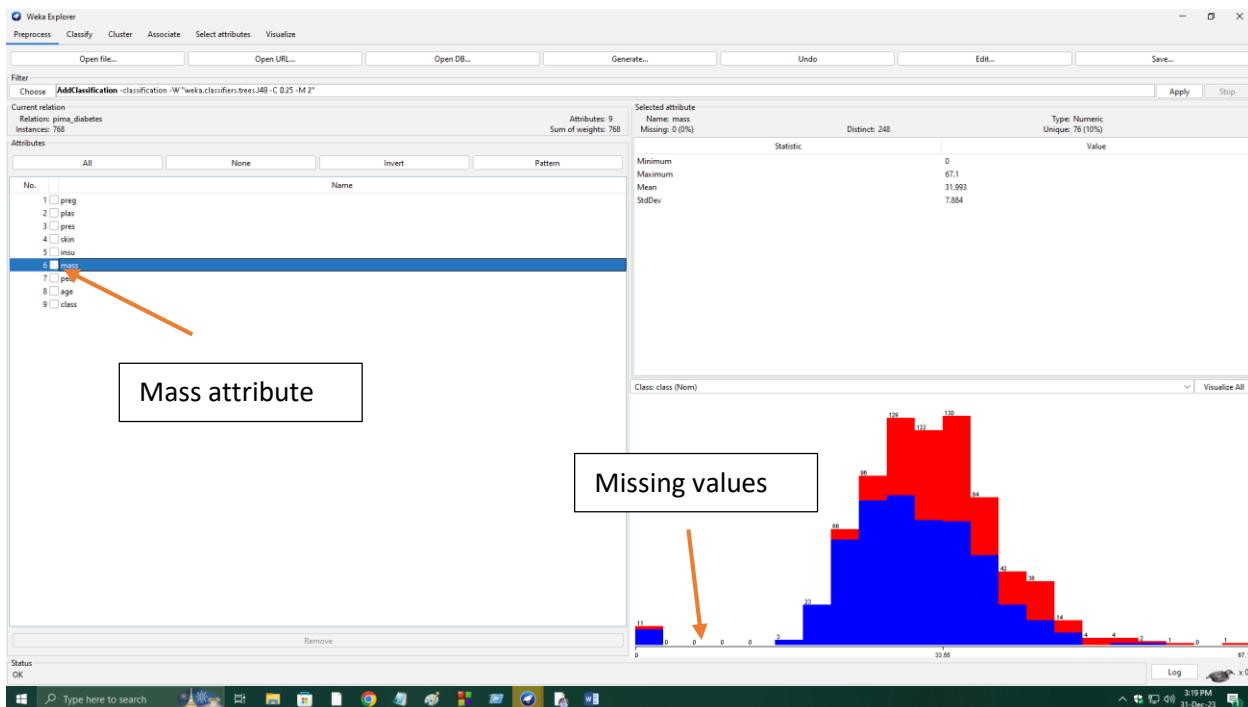
v) Tutorial video link: <https://www.youtube.com/watch?v=U-1sTxmHE5U>

Step 4 Handle missing values

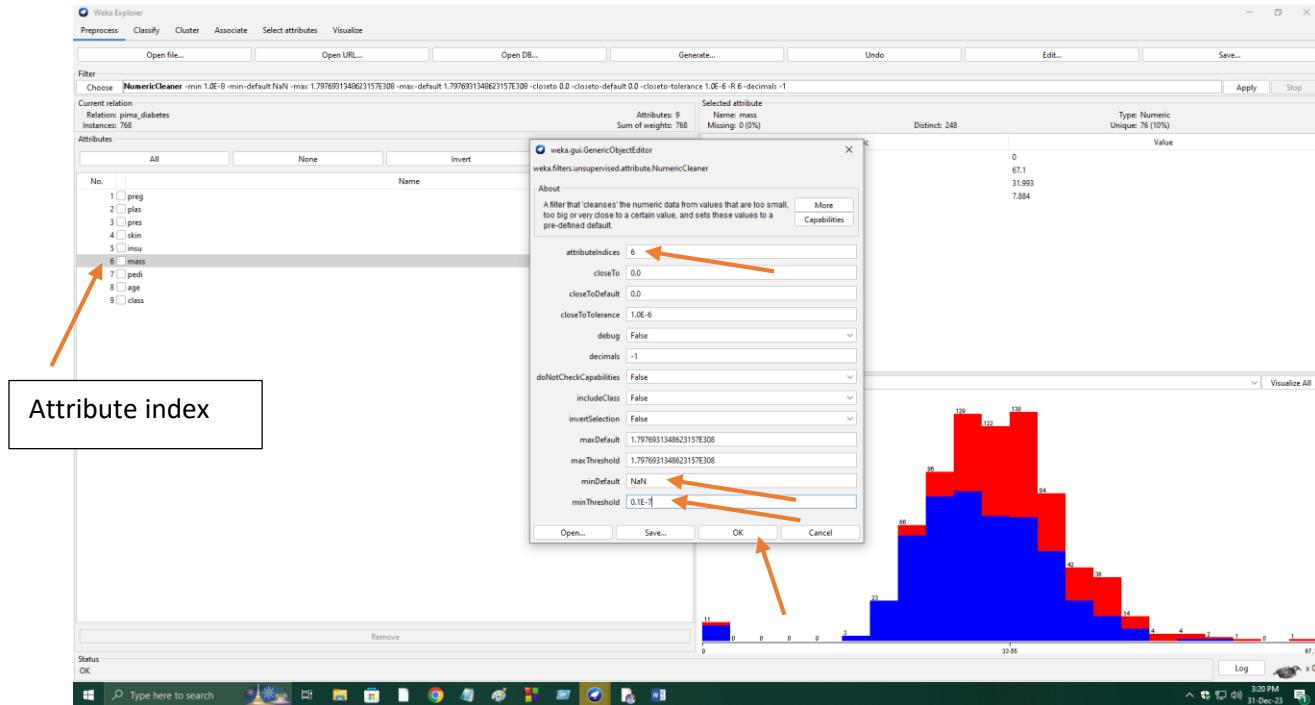
- Open the diabetes.arff file.



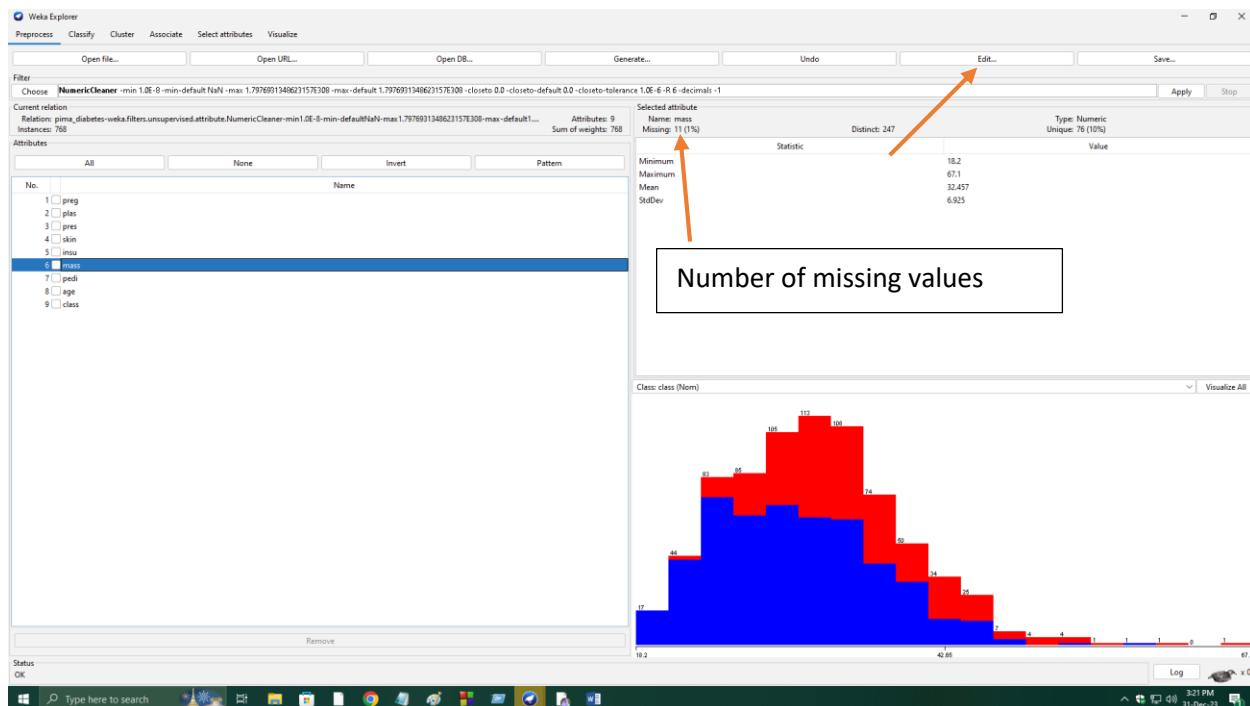
- Viewing an attribute named mass that has missing values as shown.



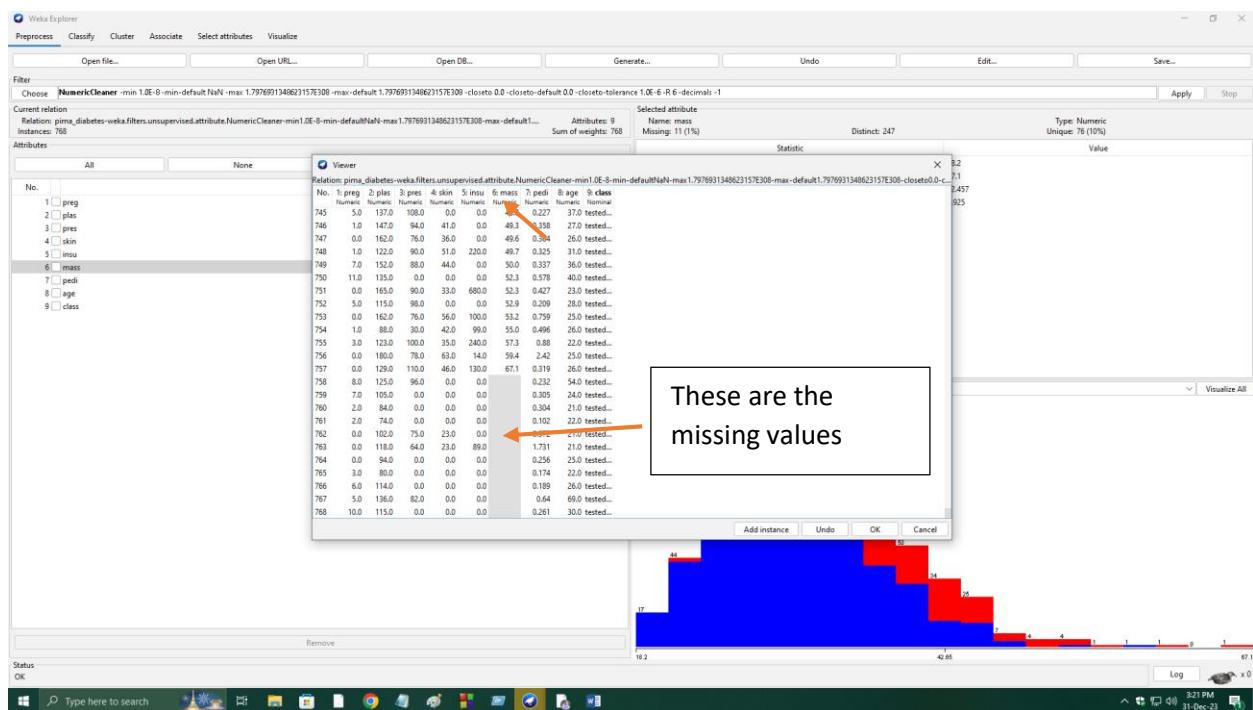
- iii) To find the missing values follow: preprocess > filter > Choose > unsupervised > attribute > NumericCleaner. The click on the filter and change the fields as shown.



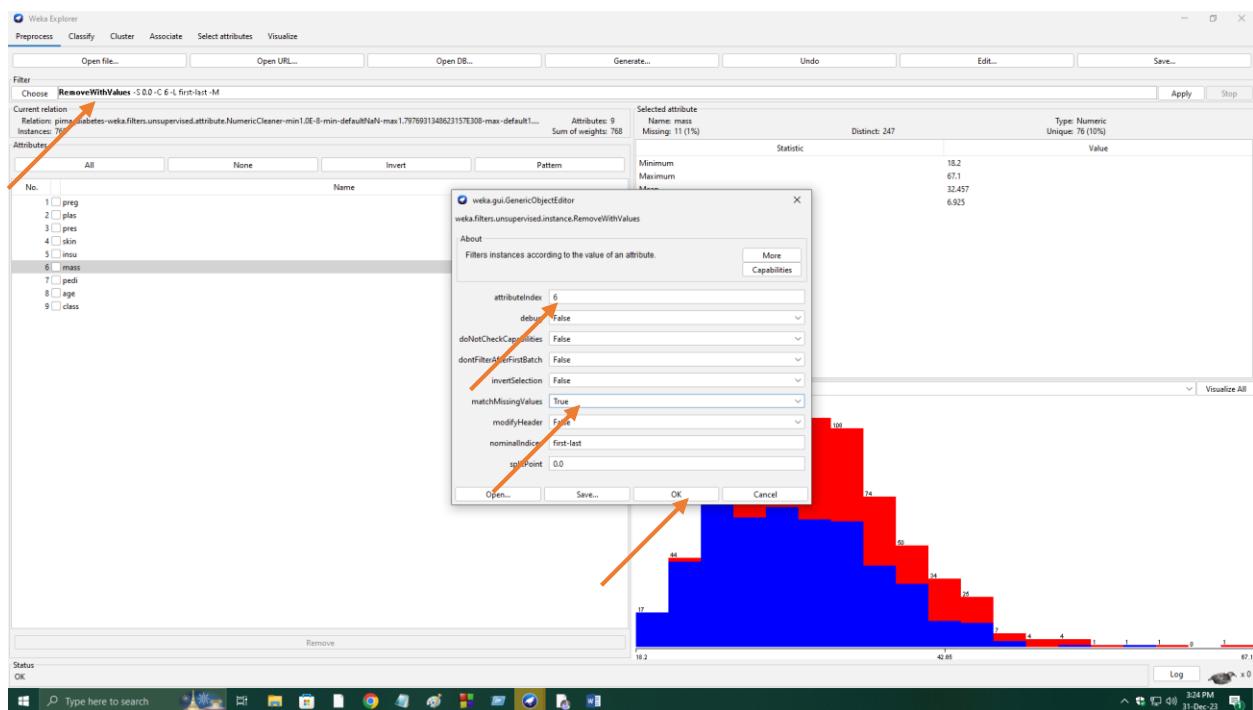
- iv) Click on edit to see the missing values.



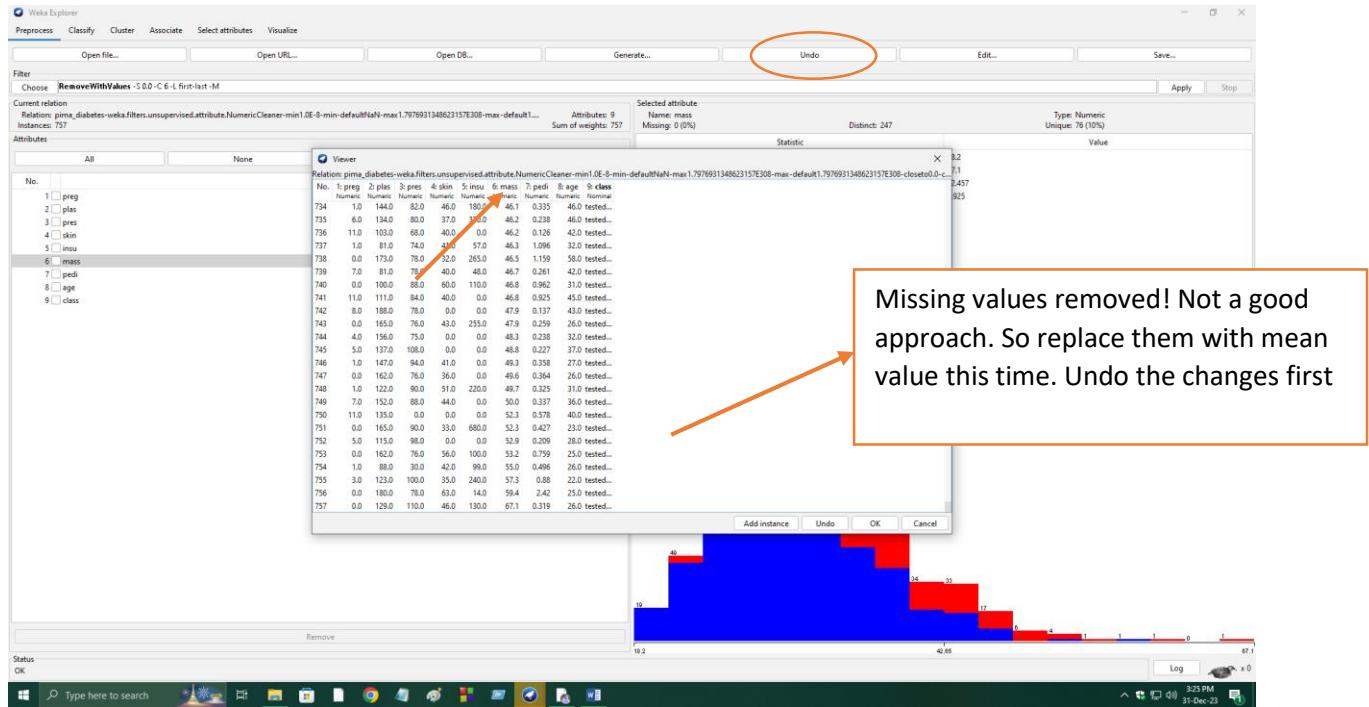
v) Select mass field and scroll down.



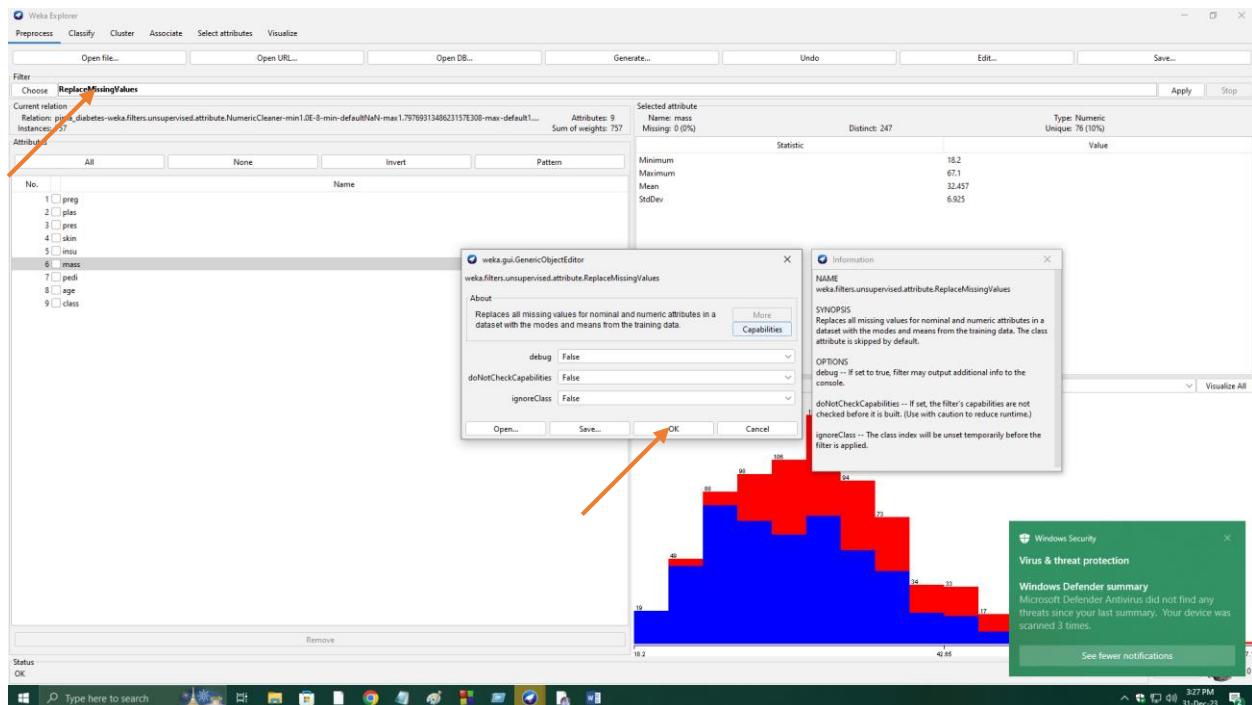
vi) Now to remove the missing values follow: preprocess > filter > choose > unsupervised > instance > RemoveWithValues. Click on the filter and modify the fields as shown. Click OK.



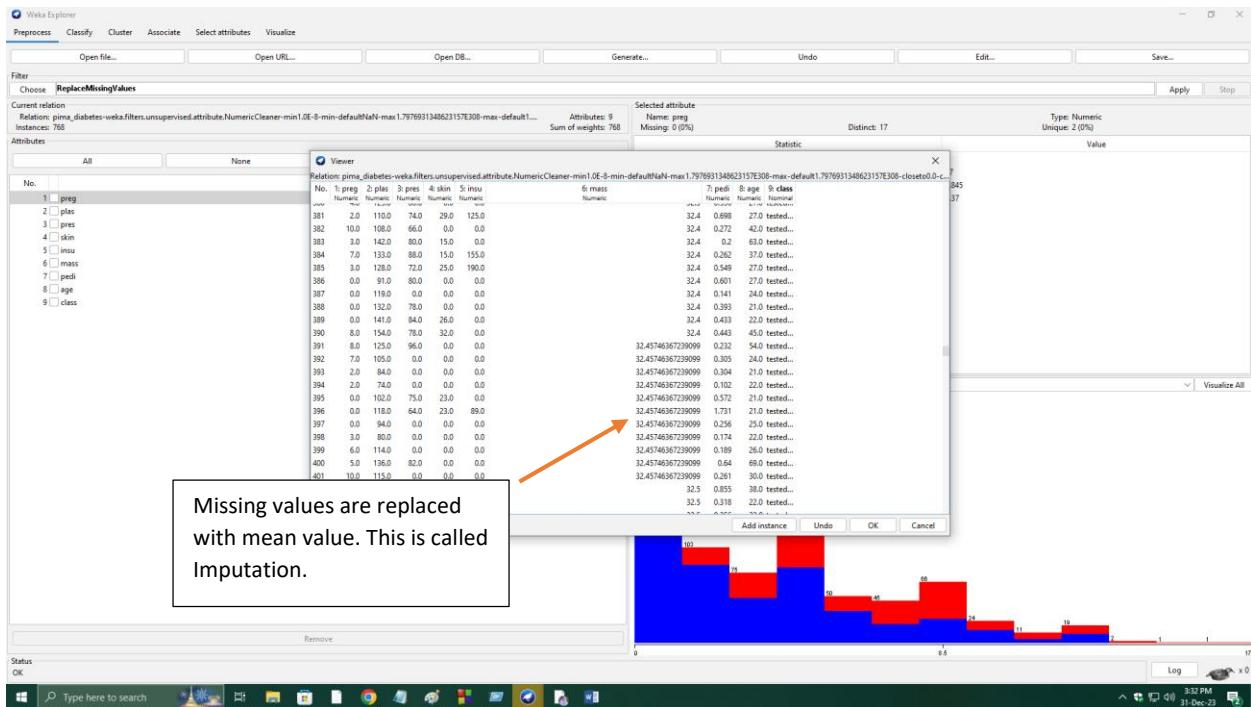
- vii) Again click edit to see the missing values. This time you won't see the records containing missing values as they are removed but this is not a good approach so now will replace the missing values. So click undo to get the missing values again.



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



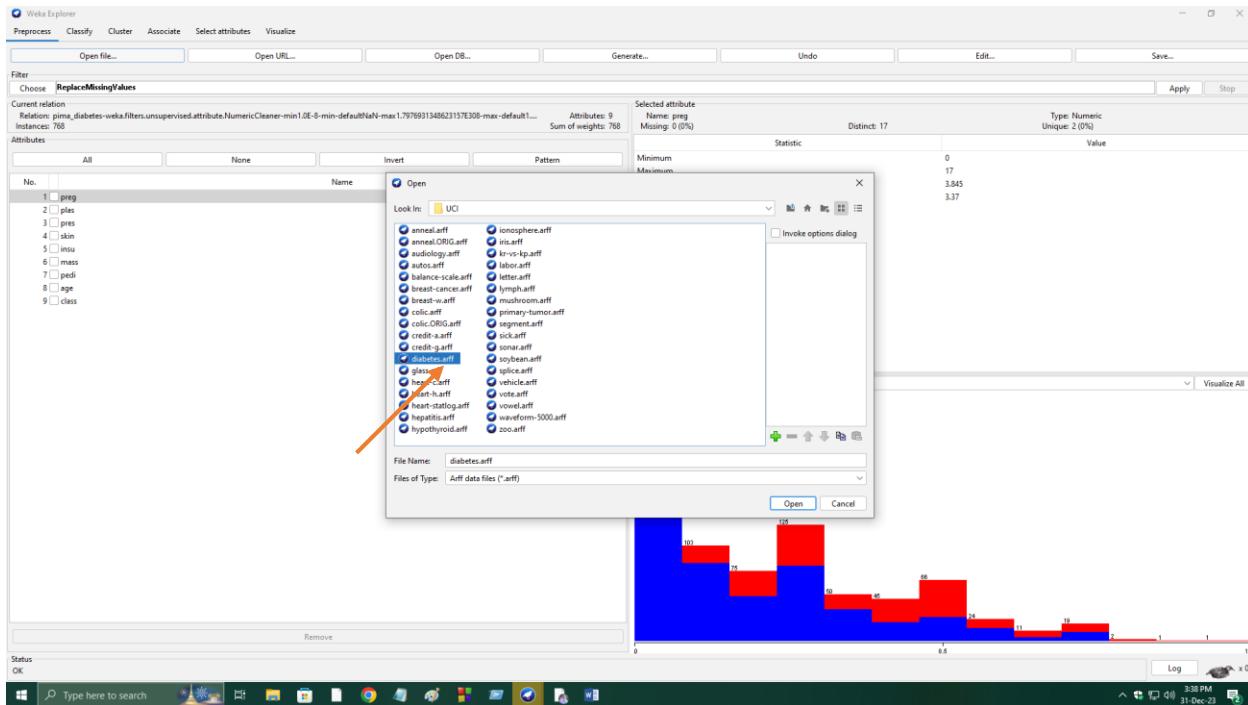
- ix) Now again click on edit to view the missing values. This time after scrolling you will see the values are replaced with mean value.



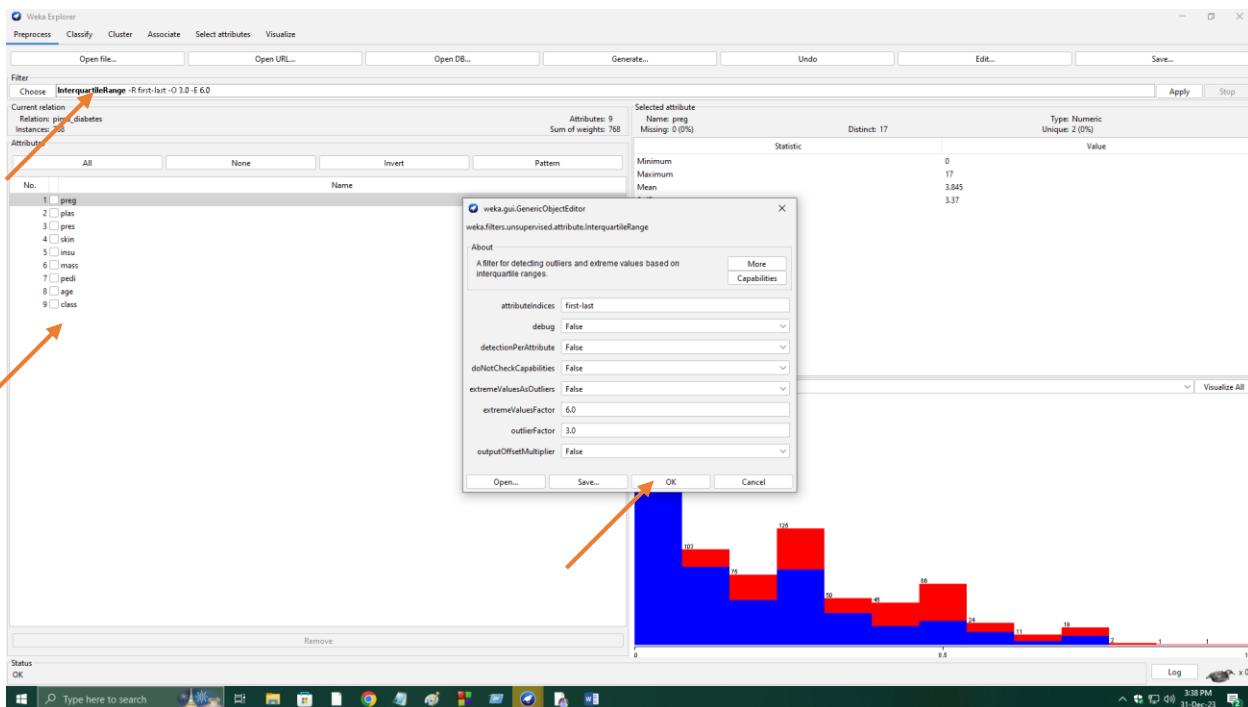
- x) Tutorial video link: <https://www.youtube.com/watch?v=Qw6yEwzaDJQ>

Step 5 Handle Outliers and Extreme values

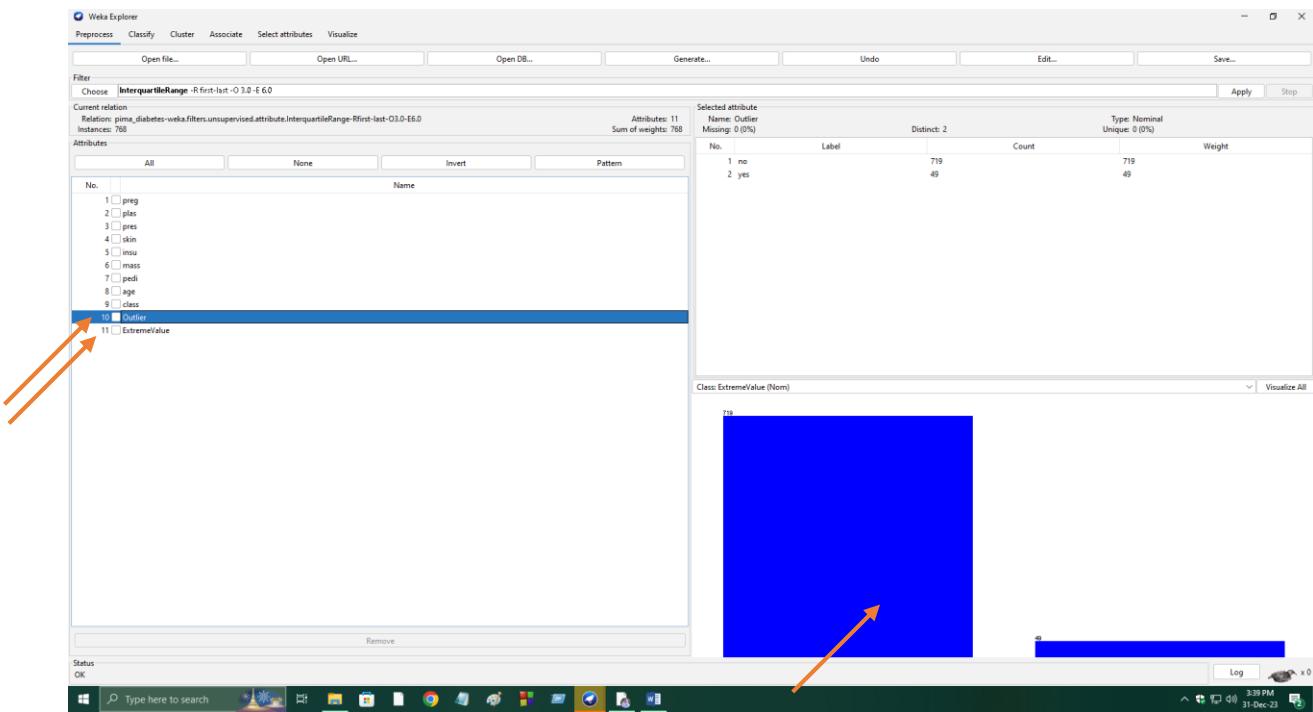
- Open the diabetes.arff file.



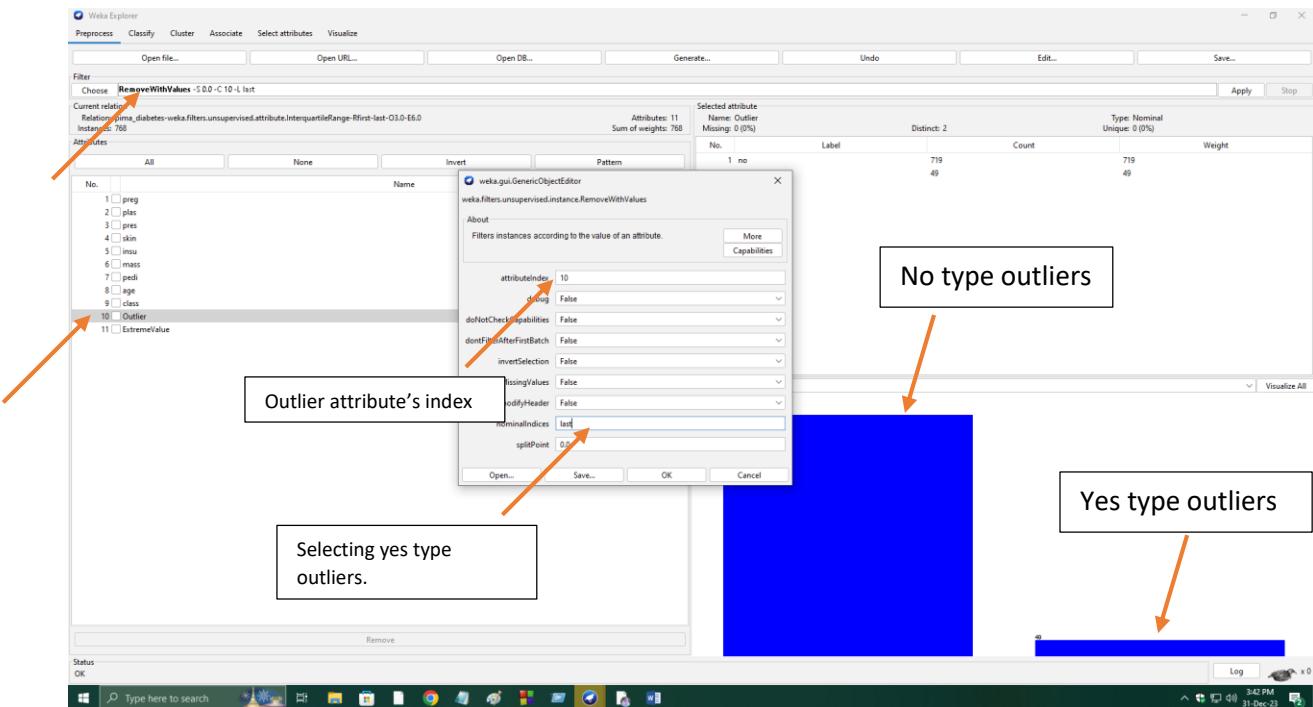
- To find the outliers and extreme values follow: preprocess > filter > choose > unsupervised > attribute > InterquartileRange.

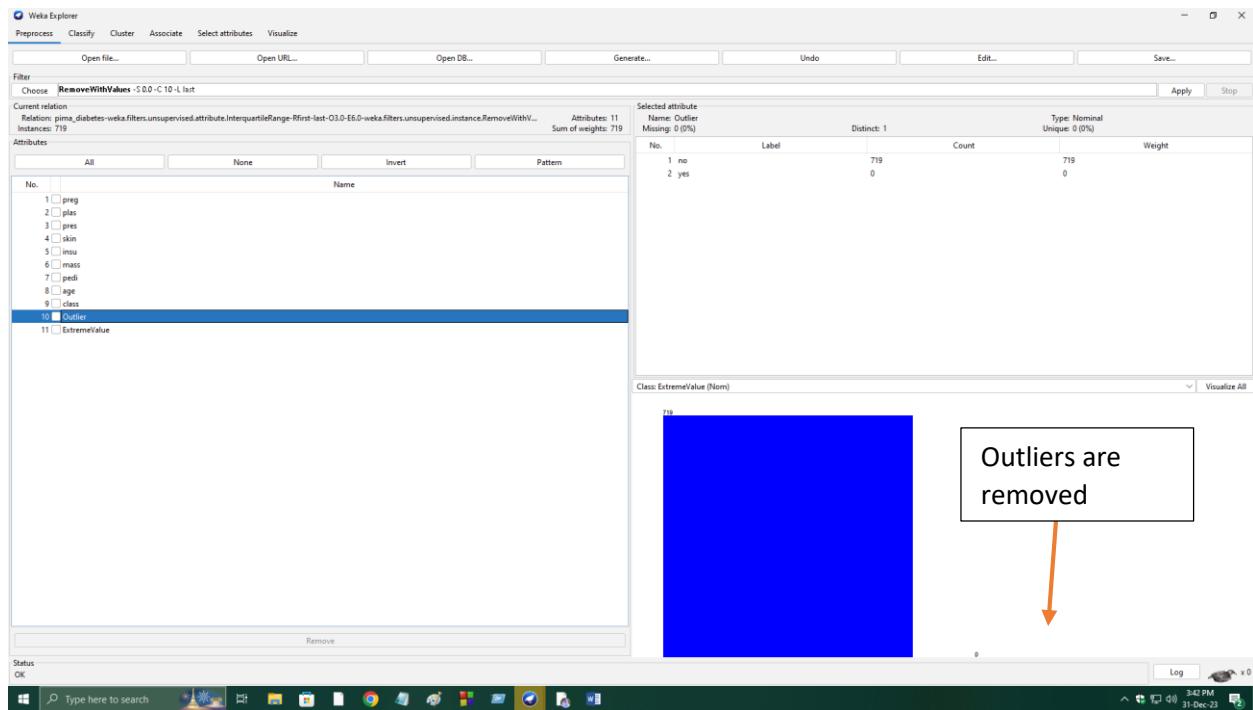


iii) Now you can see the outlier and extreme values.



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.

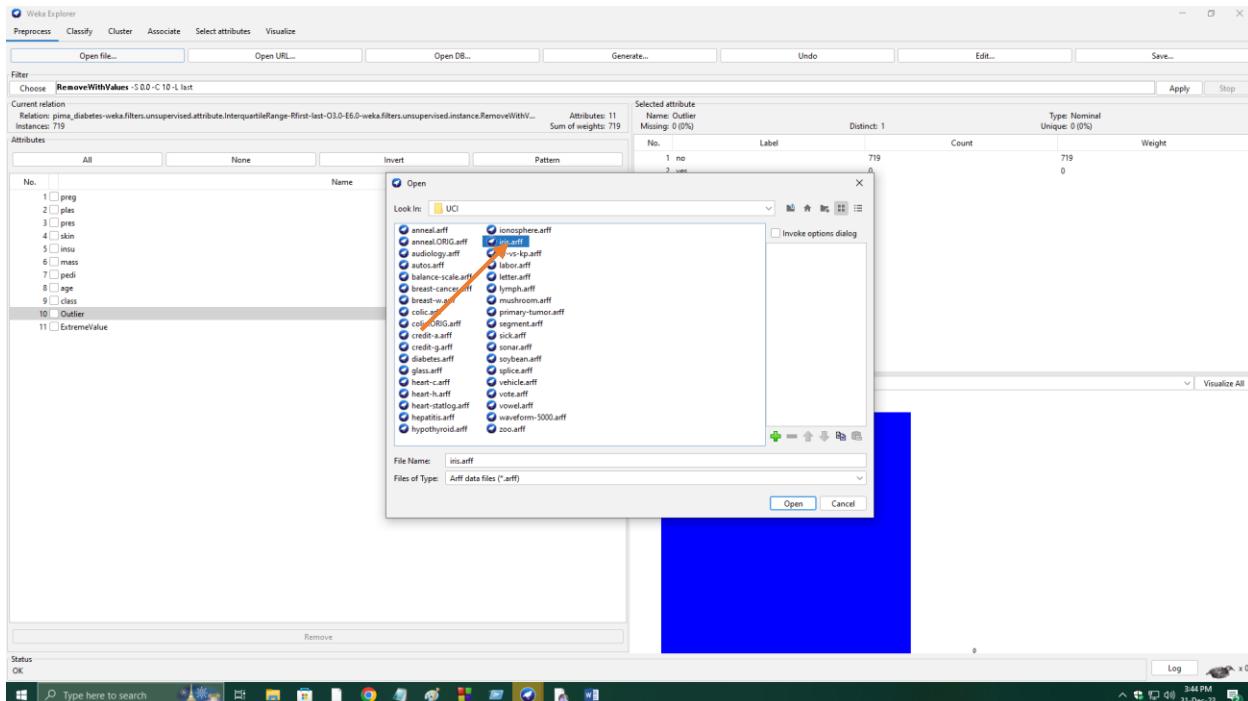




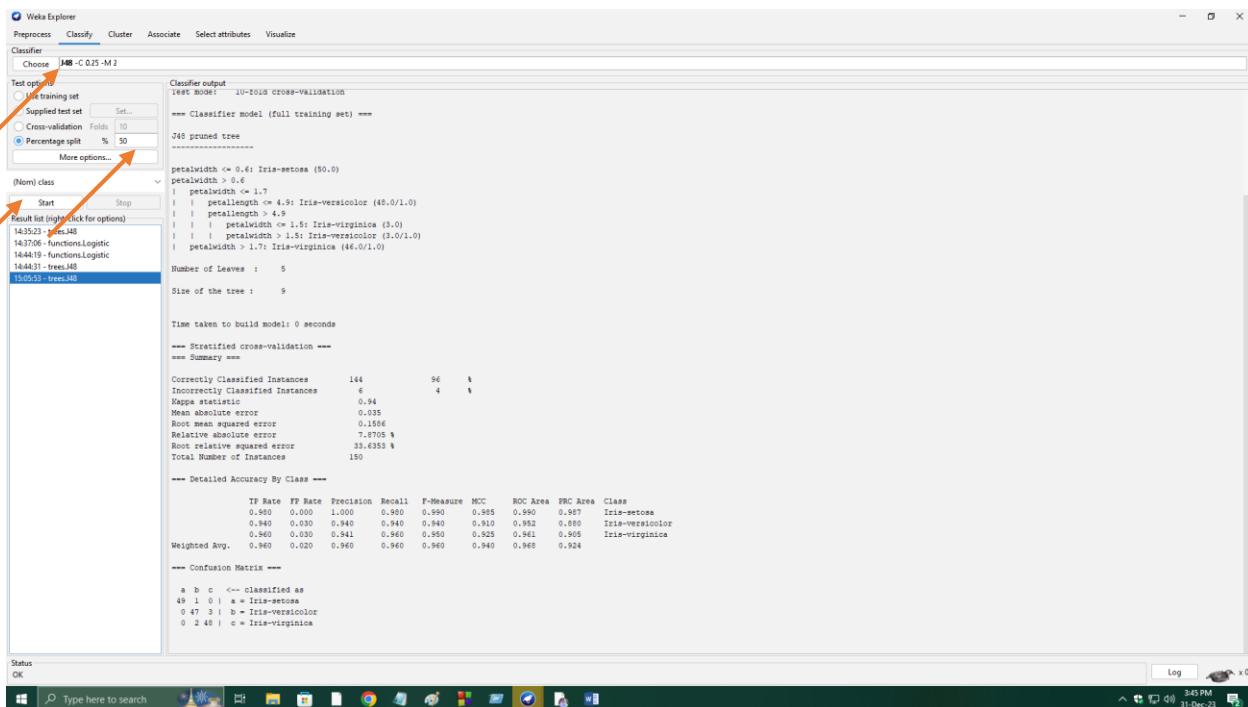
- v) Following the same way you can also remove the extreme values.
- vi) Tutorial video link: <https://www.youtube.com/watch?v=WrjpO7CmUoQ>

Step 6. Classification using Weka

- Open a file. Here iris.arff is opened.



- Follow: Classify > Classifier > Choose > tree > J48.
- Choose split percentage of your choice. Then click Start.



This is the accuracy of the decision tree.

```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier Choose J48 -C 0.25 -M 2
Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds: 10
 Percentage split % 50
More options...
(Nom) class
Start Stop
Result list (right-click for options)
143521 - trees.J48
143706 - FunctionsLogistic
144419 - FunctionsLogistic
144621 - trees.J48
150551 - trees.J48
154530 - trees.J48
154530 - trees.J48
J48 pruned tree
-----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petalwidth <= 4.9: Iris-versicolor (48.0/1.0)
| | | petalwidth <= 4.9: Iris-versicolor (3.0)
| | | | petalwidth <= 1.5: Iris-versicolor (3.0)
| | | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| | petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Size of the tree : 9
Time taken to build model: 0 seconds
*** Evaluation on test split ***
Time taken to test model on test split: 0 seconds
*** Summary ***
Correctly Classified Instances 71 94.6667 %
Incorrectly Classified Instances 4 5.3333 %
Kappa statistic 0.9198
Mean absolute error 0.0519
Root mean squared error 0.1195
Relative absolute error 11.6302 %
Root relative squared error 40.0146 %
Total Number of Instances 75
*** Detailed Accuracy By Class ***
IP Rate FP Rate Precision Recall F-Measure MCC ROC Area FRC Area Class
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Iris-setosa
1.000 0.000 0.867 1.000 0.929 0.907 0.959 0.867 Iris-versicolor
0.862 0.000 1.000 0.852 0.820 0.887 0.926 0.865 Iris-virginica
Weighted Avg. 0.947 0.028 0.984 0.947 0.946 0.922 0.959 0.920
*** Confusion Matrix ***
a b c <- classified as
22 0 0 | a = Iris-setosa
0 26 0 | b = Iris-versicolor
0 4 23 | c = Iris-virginica

```

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

```

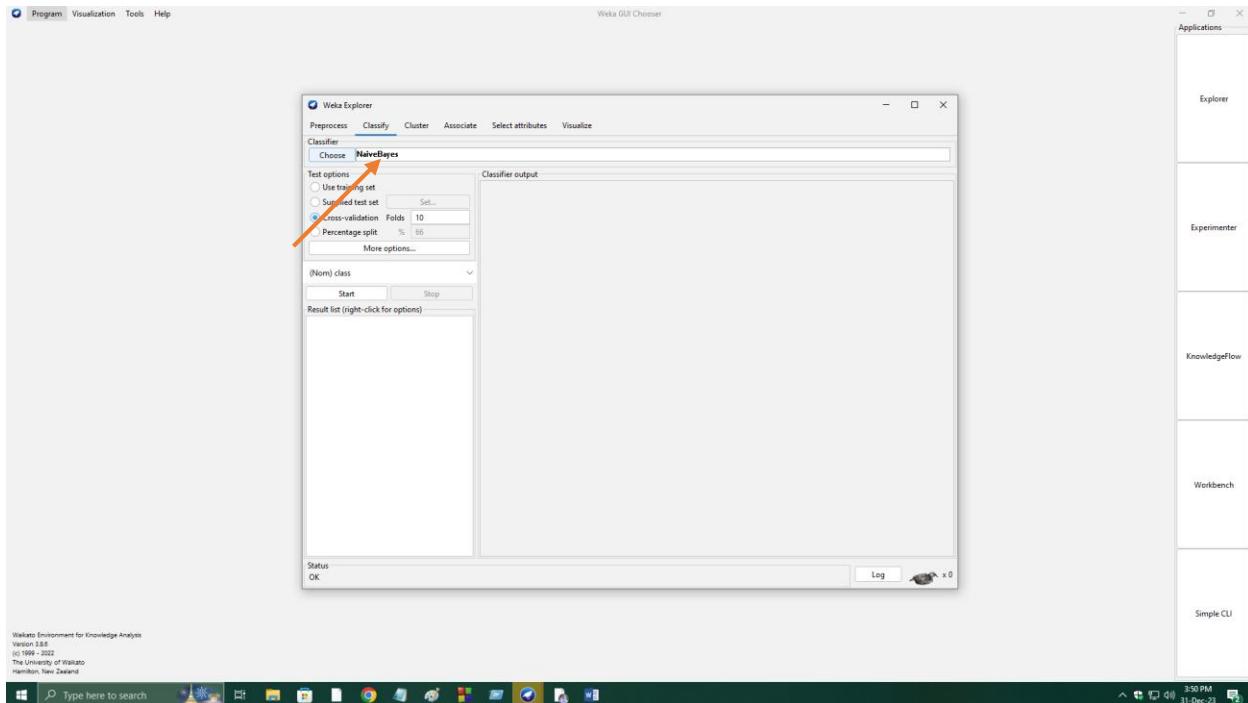
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier Choose J48 -C 0.25 -M 2
Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds: 10
 Percentage split % 70
More options...
(Nom) class
Start Stop
Result list (right-click for options)
143521 - trees.J48
143706 - FunctionsLogistic
144419 - FunctionsLogistic
144621 - trees.J48
150551 - trees.J48
154530 - trees.J48
154607 - trees.J48
J48 pruned tree
-----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petalwidth <= 4.9: Iris-versicolor (48.0/1.0)
| | | petalwidth <= 4.9: Iris-versicolor (3.0)
| | | | petalwidth <= 1.5: Iris-versicolor (3.0)
| | | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| | petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Size of the tree : 9
Time taken to build model: 0 seconds
*** Evaluation on test split ***
Time taken to test model on test split: 0 seconds
*** Summary ***
Correctly Classified Instances 43 95.1556 %
Incorrectly Classified Instances 2 4.4444 %
Kappa statistic 0.9331
Mean absolute error 0.0416
Root mean squared error 0.1146
Relative absolute error 9.1444 %
Root relative squared error 35.0559 %
Total Number of Instances 45
*** Detailed Accuracy By Class ***
IP Rate FP Rate Precision Recall F-Measure MCC ROC Area FRC Area Class
1.000 0.000 1.000 1.000 1.000 1.000 1.000 Iris-setosa
1.000 0.049 0.859 1.000 0.941 0.910 0.946 0.859 Iris-versicolor
0.859 0.000 1.000 0.847 0.829 0.901 0.944 0.921 Iris-virginica
Weighted Avg. 0.956 0.025 0.980 0.954 0.955 0.938 0.976 0.938
*** Confusion Matrix ***
a b c <- classified as
14 0 0 | a = Iris-setosa
0 16 0 | b = Iris-versicolor
0 2 13 | c = Iris-virginica

```

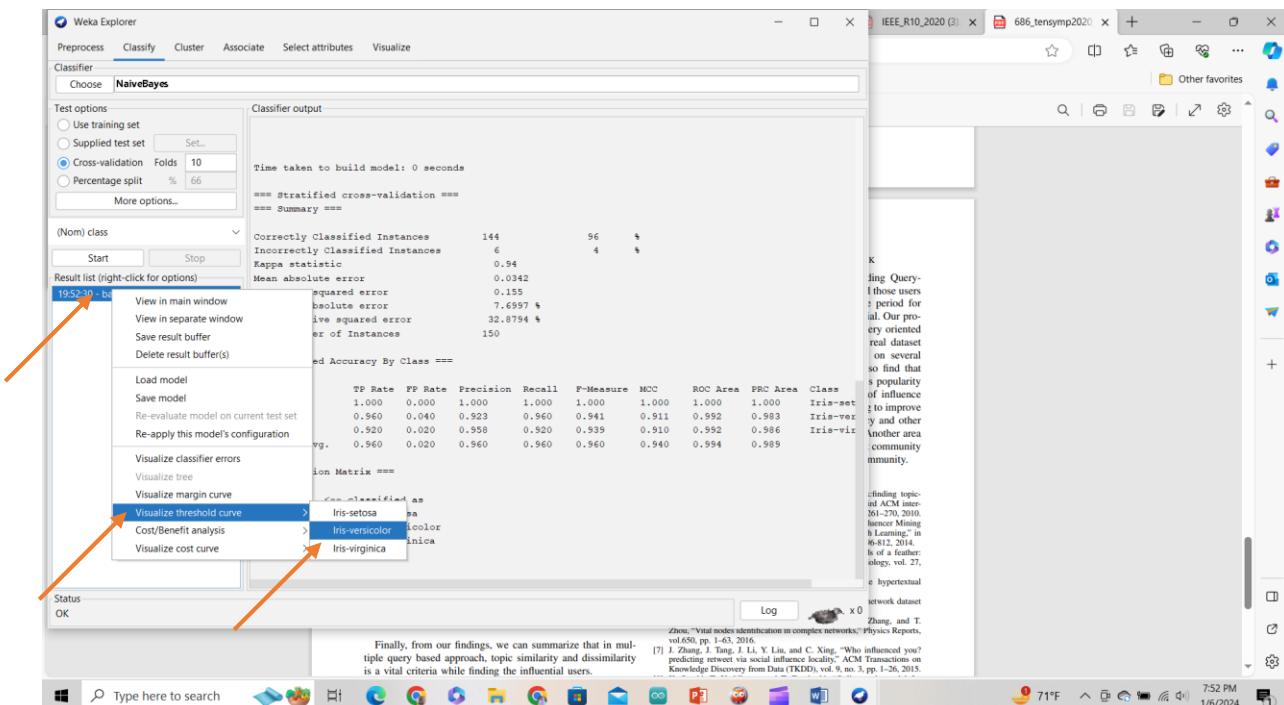
v) Tutorial video link: <https://www.youtube.com/watch?v=guf93Gq3wAI>

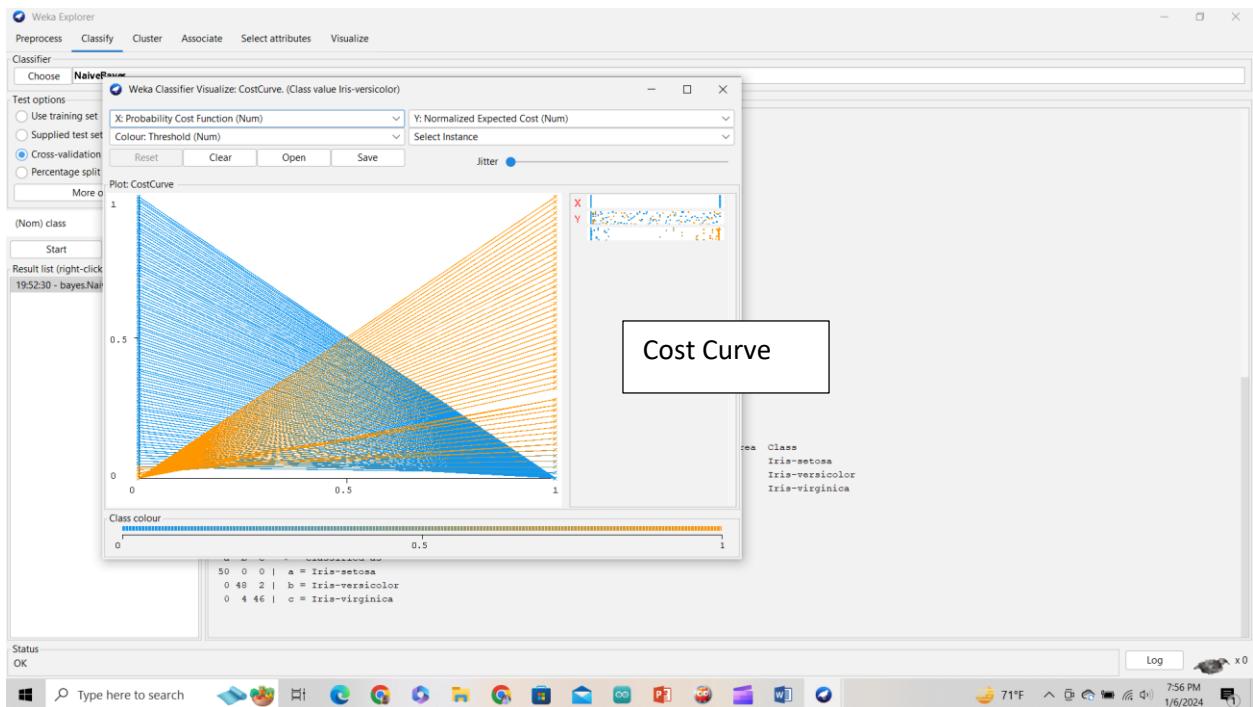
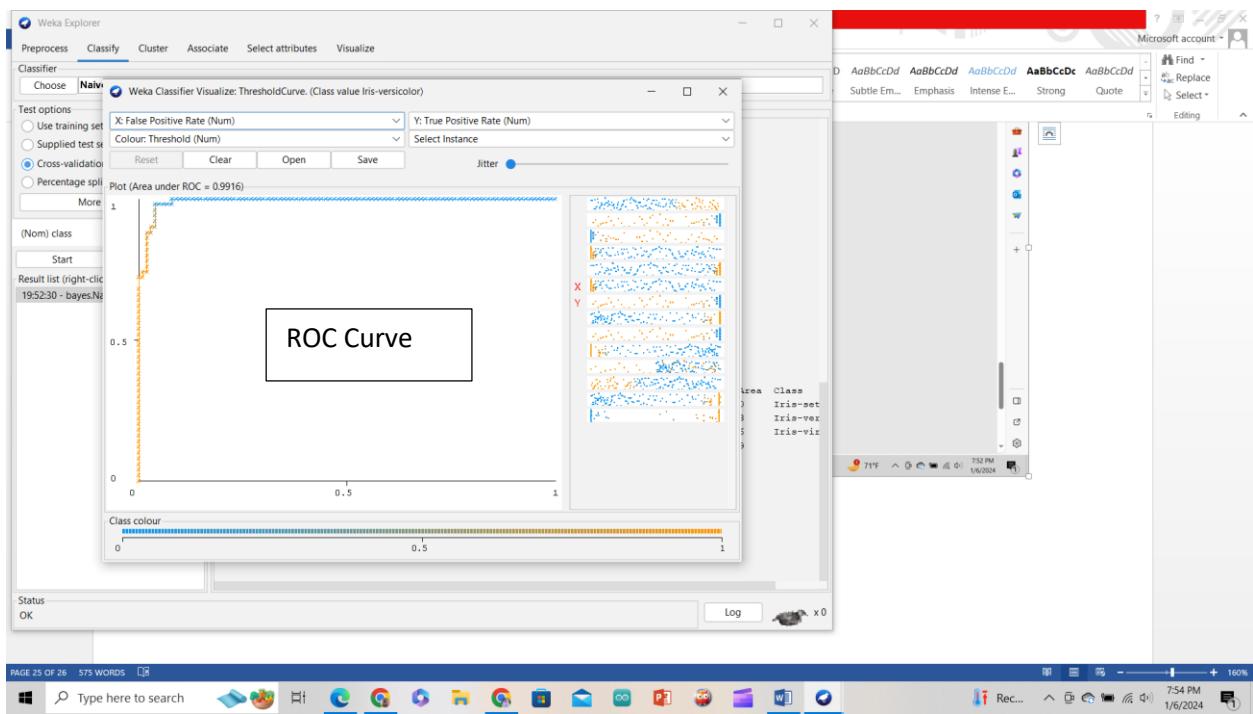
Step 7. Visualize result in Weka

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



- 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_-b0gwy