AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Assignment Cover Page

Assignment Title:	DATA WAREHOUSING AND DATA MINING FINAL PROJECT		
Assignment No:	1	Date of Submission:	20 Aprill, 2022
Course Title:	DATA WAREHOUSING AND DATA MINING		
Course Code:	Click here to enter text.	Section:	A
Semester:	Spring 2019-20	Course Teacher:	AKINUL ISLAM JONY

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*

** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No. : 1

No	Name	ID	Program	Signature
1	MD SAKIB HOSSAIN SHOVON	19-39526-1	BSc [CSE]	SHOVON

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Introduction:

The main objective of this project is to analyzing data and to find out hidden pattern of the data, discovering new findings by applying Data Mining techniques. Here we have used several Data Mining techniques such as K-Nearest Neighbours, Decision Trees, Naïve Bayes, Random Forest, Support Vector Machine, and Neural Network. We have collected the “Drug200” dataset from kaggle. We are using weka tool to analyse the data.

Dataset:

The dataset we have taken is Drug200. We have 6 attributes in this dataset. These are Age, Sex, BP, Cholesterol, Na_to_K and Drug. Drug is the target attributes. There are 200 instances.There are no missing values in our dataset. All the values are in numeric form.

Age attributes: Age is numerical. Age of the patient is 15 to 74.

Sex attributes: Gender of the patients is male and female. Male percentage is 52% and Female percentage is 48%.

BP attributes: 39% patient blood pressure level is high, 32% patient blood pressure level is low and 30% patient blood pressure level is normal.

Cholesterol attributes: 52% patient cholesterol is high and 49% patient cholesterol is normal.

Na_to_K: Sodium to potassium Ration in Blood varies from 6.27 to 38.2

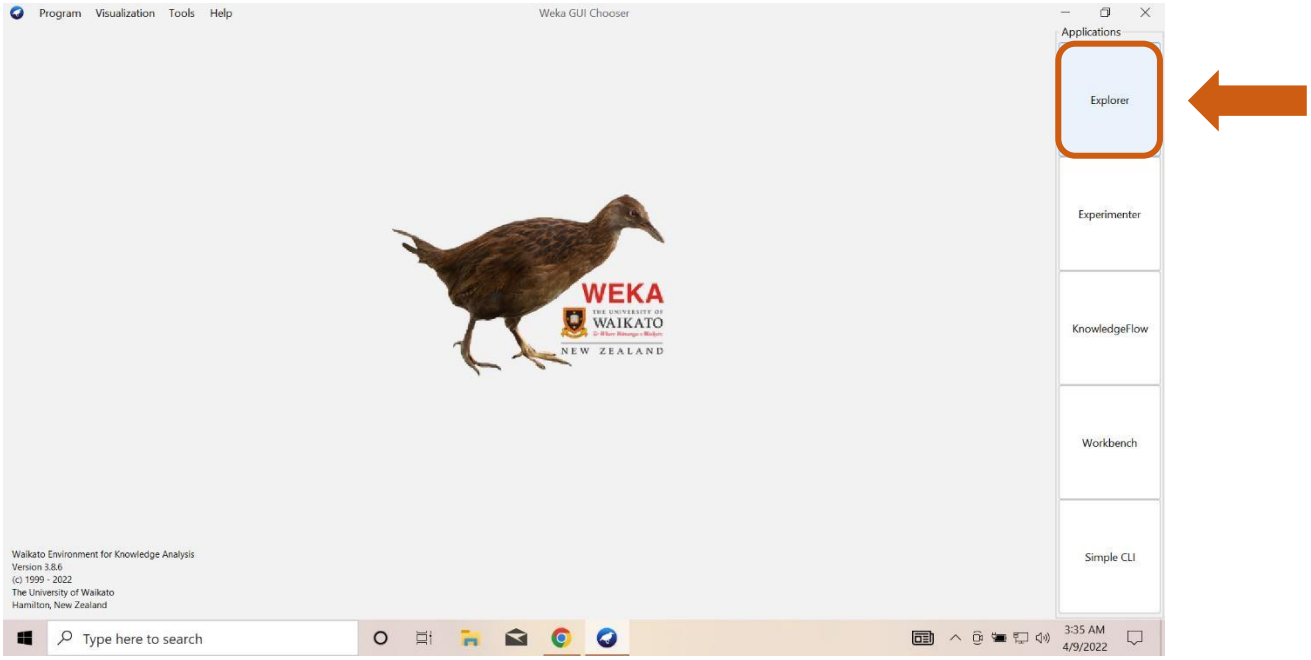
Drug: It's the target attribute that we want to classify. Here drag X is 27%, drag Y is 46% and other minor drags are 28% all together.

URL:

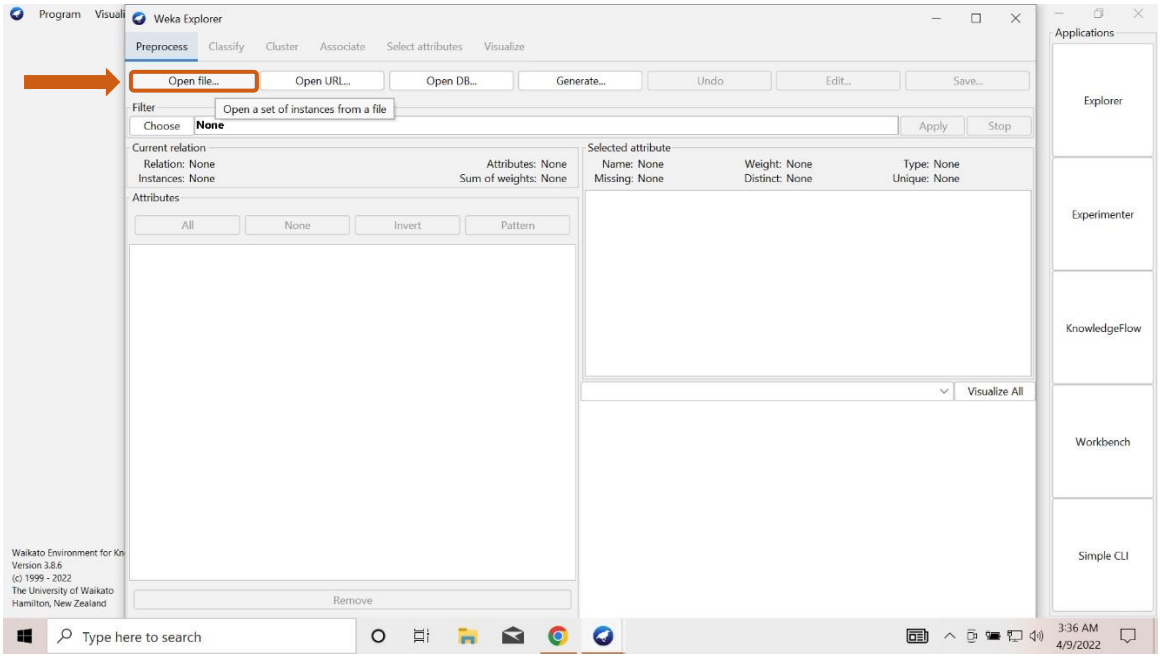
https://www.kaggle.com/code/ashwinshetgaonkar/drug-classification-random-forest-100-acc/data

Model Development:

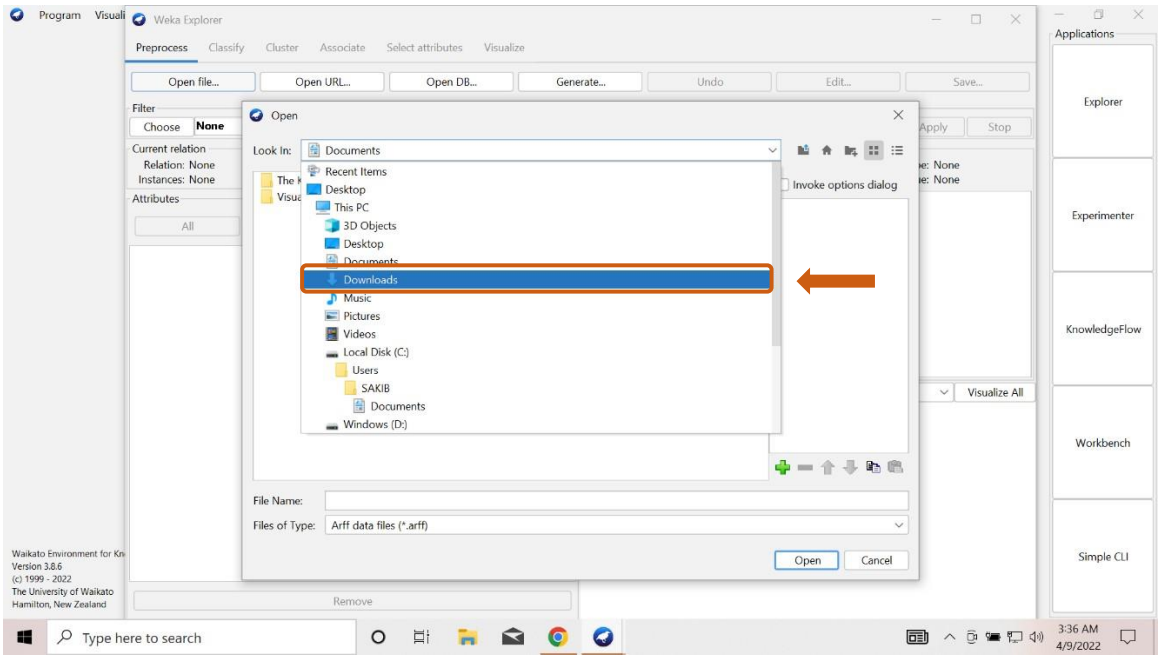
Weka was used in this project. After opening the software, the interface will look like this. We have to click on "Explorer".



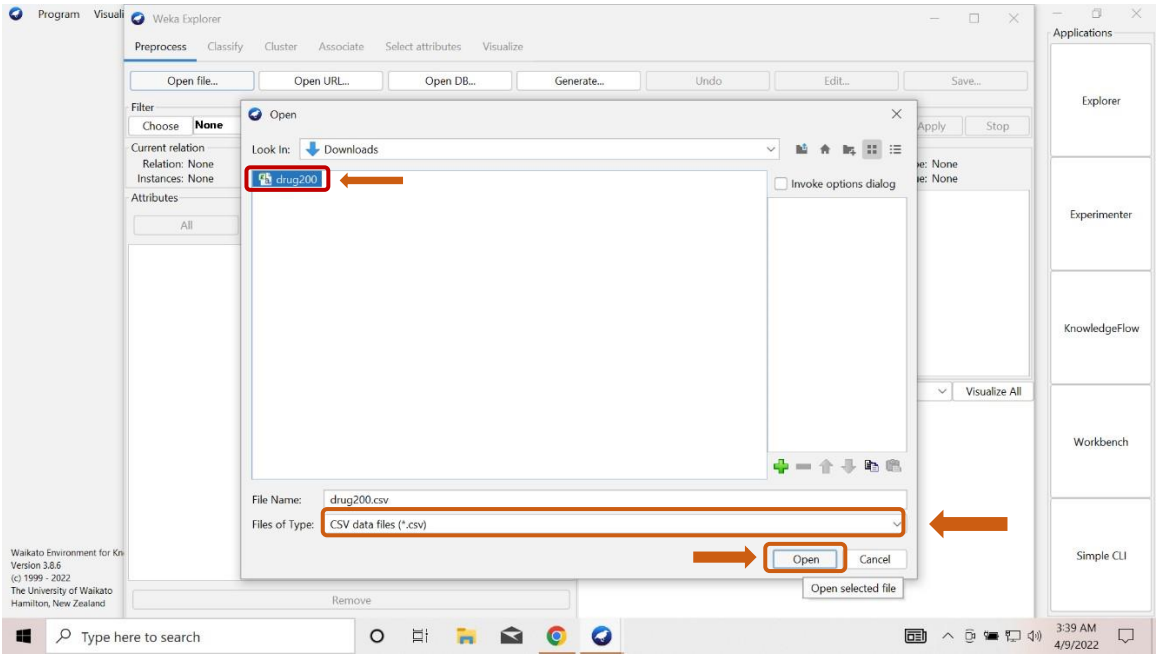
Then this interface will appear and we have to choose "Open file" option.



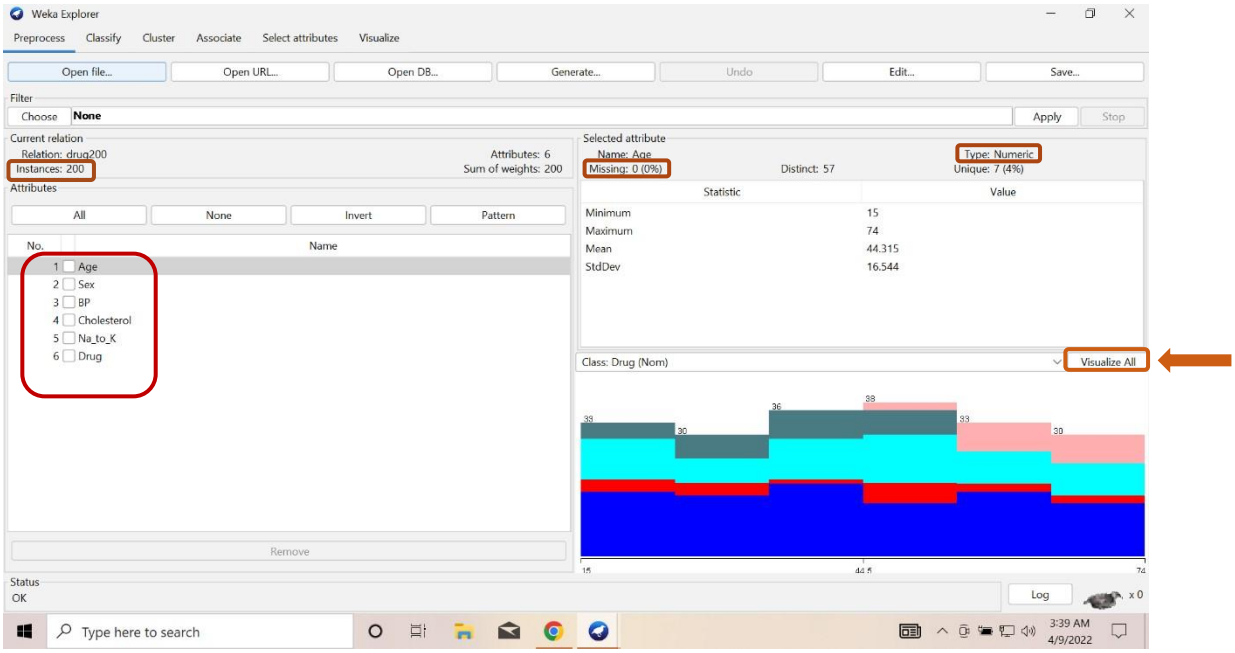
After clicking on “Open file” this interface will appear and we will be selecting the folder that have the dataset file.



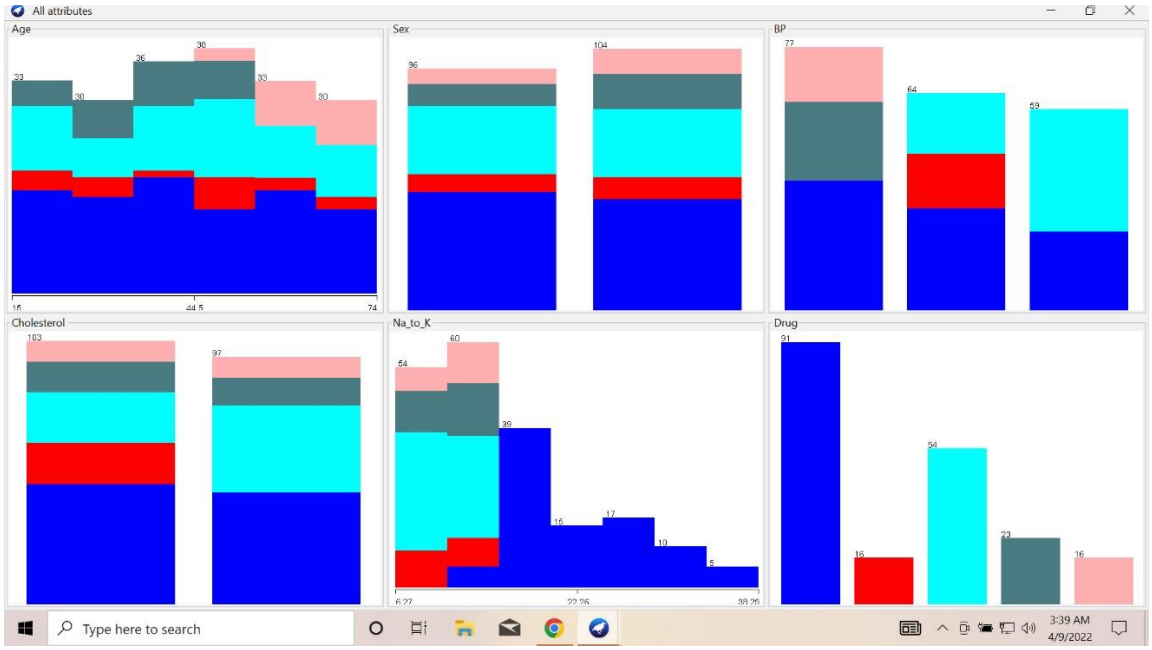
We have to make sure that we choose the correct files type. Our dataset was in .csv format so we have selected “CSV data files(.csv)”. Then our required dataset will be visible and we have to select it and by clicking on “Open” it will be open a new interface.



Here we can see that all the information is perfectly visible. All the six attributes are showing here. We can see the number of instances is 200. There is no missing value in this dataset and the type is numeric.



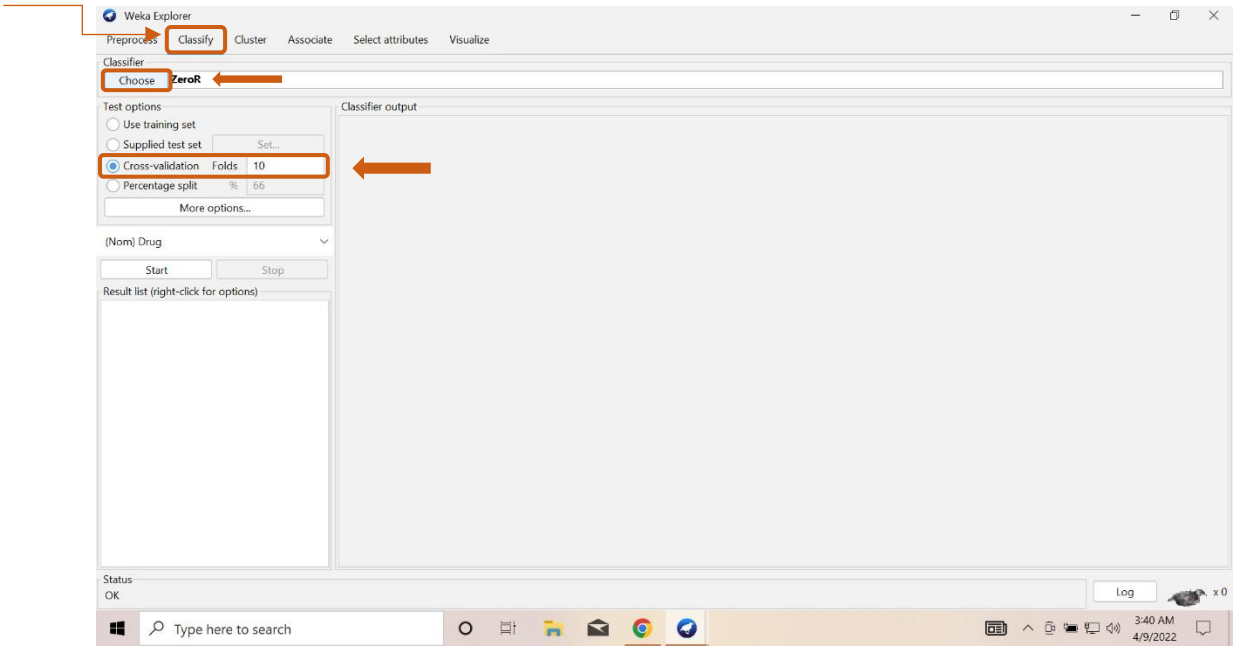
By clicking on “Visualize All” a new interface will appear. We can observe six types of graph for six types of attribute.



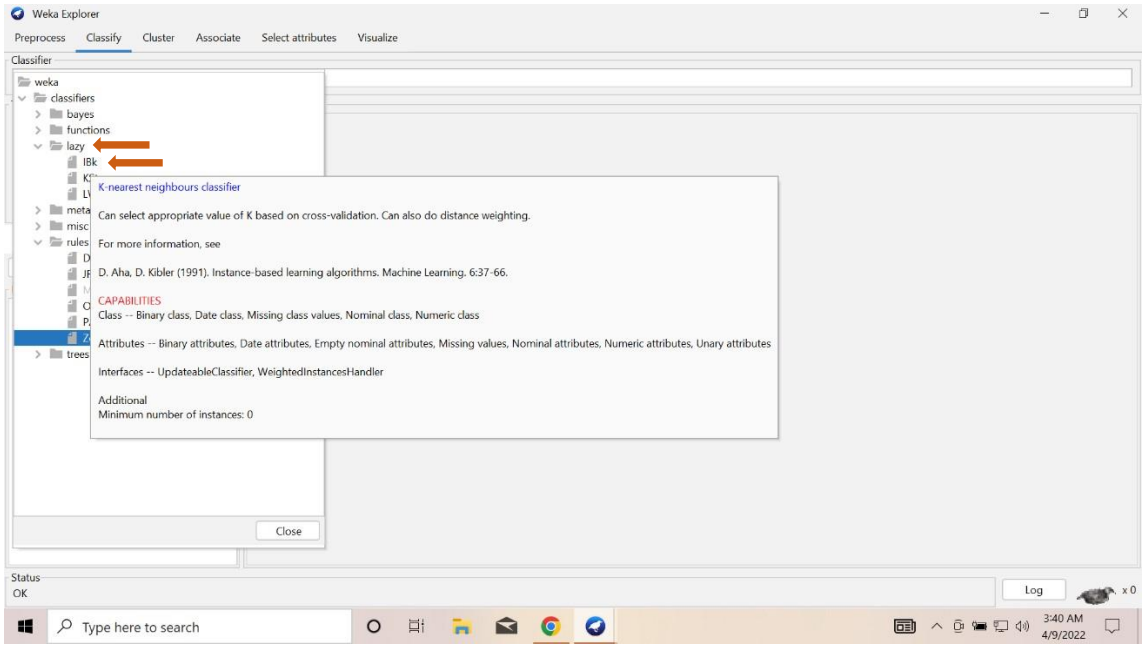
Till now we haven’t trained our model yet.For all kind of classification these steps will be same.Now are going to apply algorithm.

K-Nearest Neighbours (KNN):

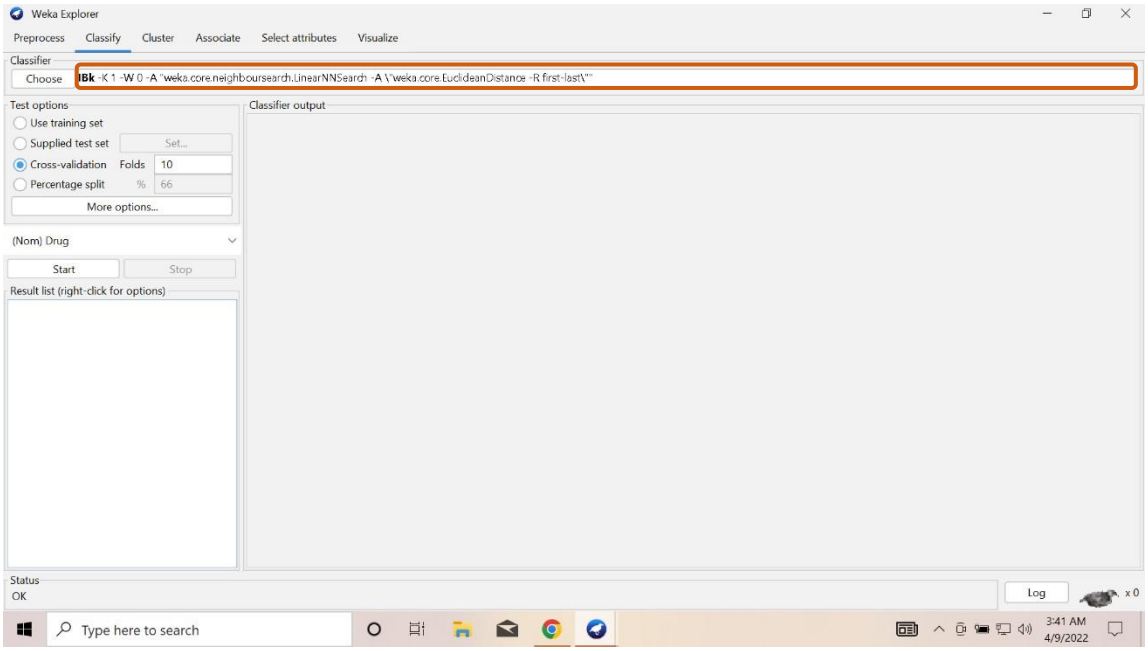
Now if we want to apply KNN the below steps will be followed. Firstly, choosing the option "Classify" and then choosing "Cross-validation, Fold 10". Then we will click the option "Choose".



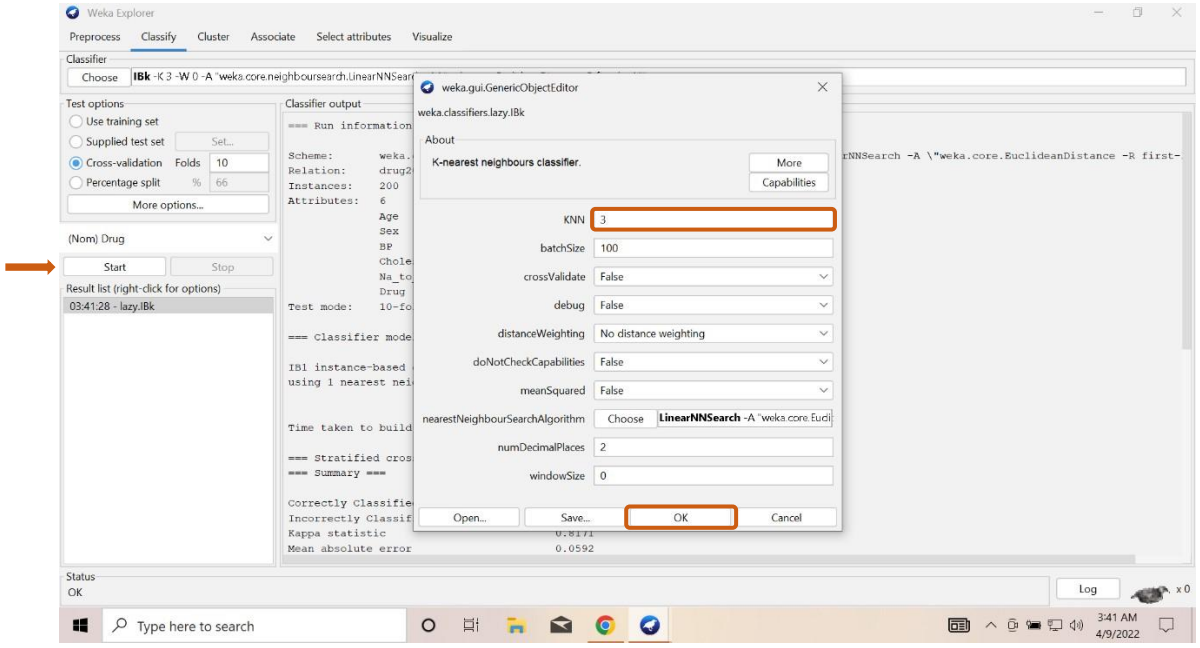
To work with K-Nearest Neighbour Classification we have choosen "lazy" then "IBk"



Then "IBk" will be loaded.

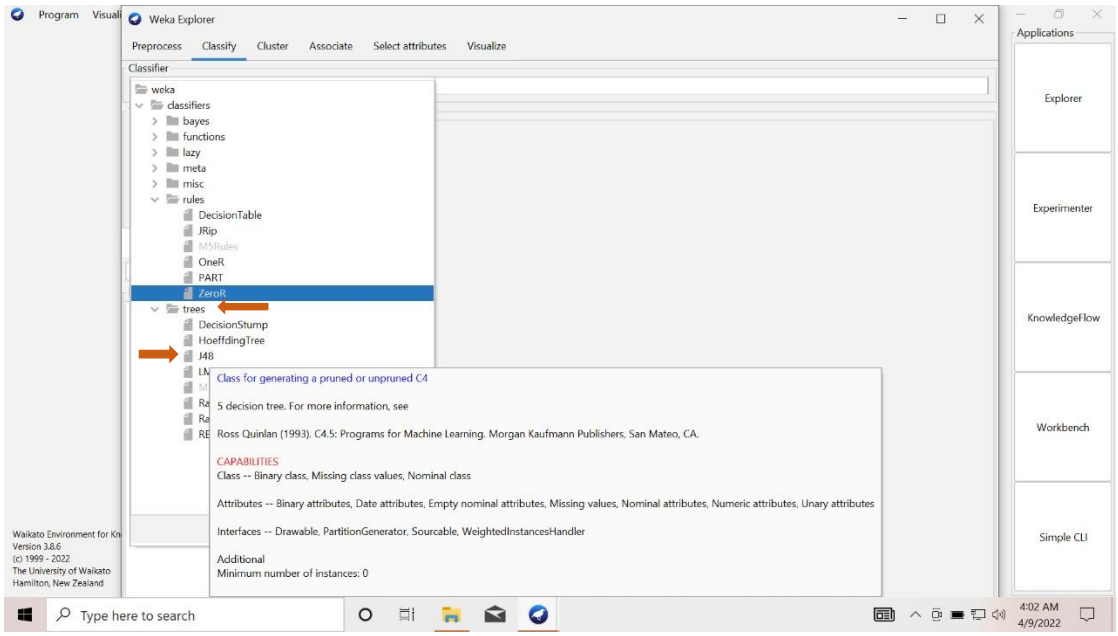


After left click in “IBk” this window will appear and we are choosing KNN value 3 then “OK” then “Start”.

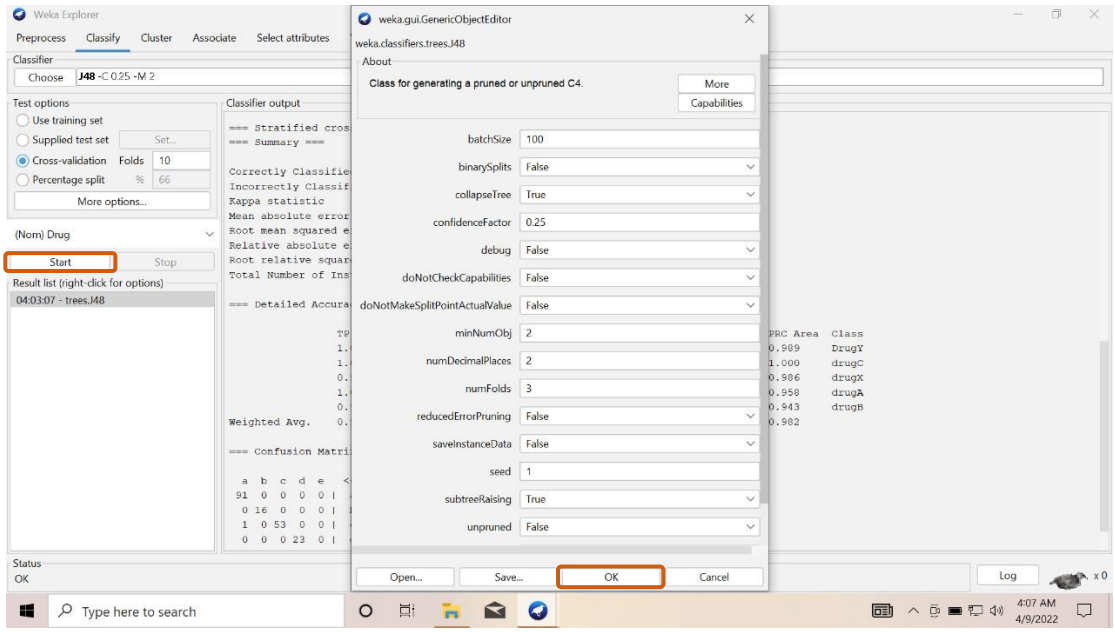


Decision Tree Classification Algorithm:

Now if we want to apply Decision Tree Classification Algorithm the below steps will be followed. Firstly, choosing the option “Classify” and then choosing “trees” then “J48”.

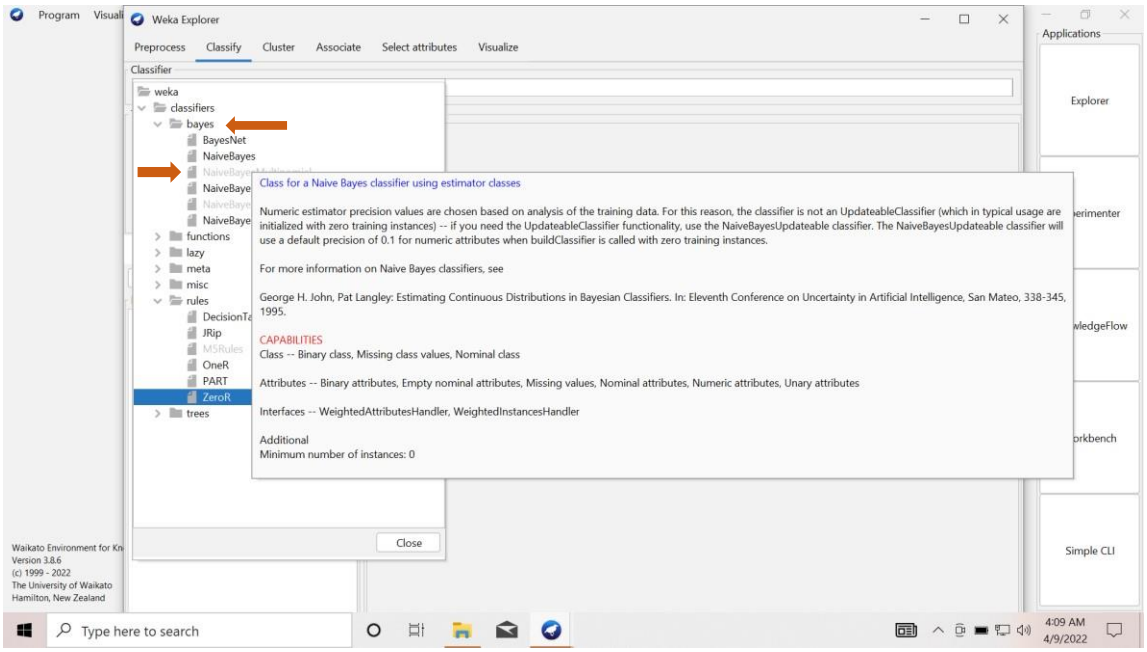


After left click in “J48” this window will appear and we are choosing “OK” then “Start”.

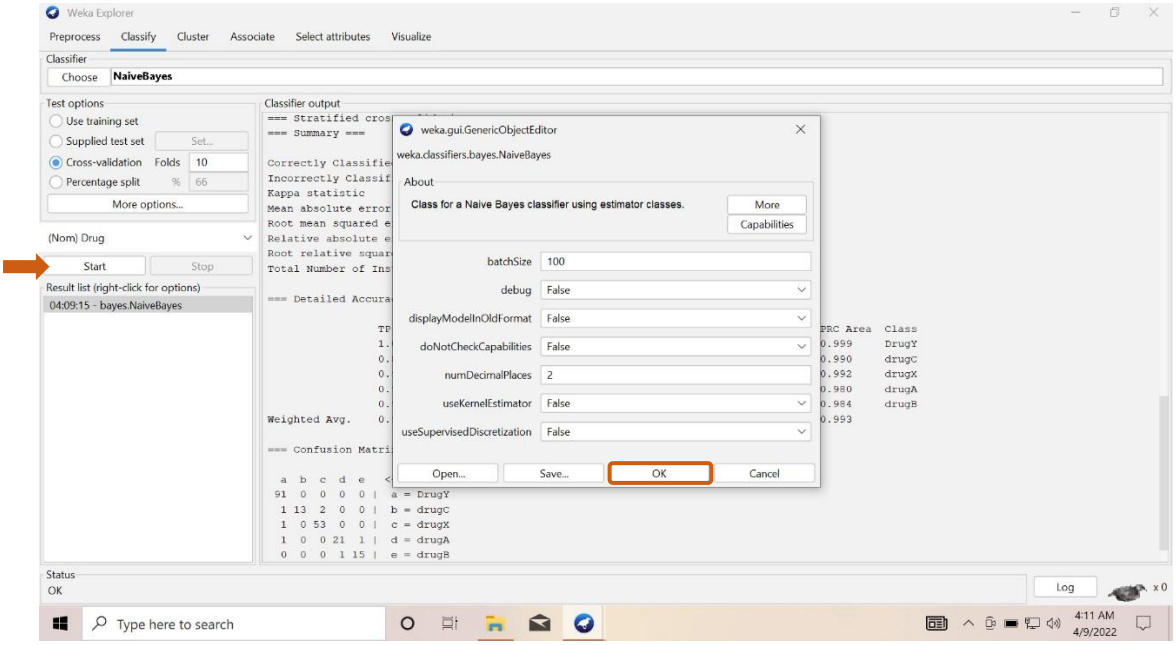


Naïve-Bayes Classification Algorithm:

Now if we want to apply Naïve-Bayes Classification Algorithm the below steps will be followed. Firstly, choosing the option “Classify” and then choosing “bayes”.

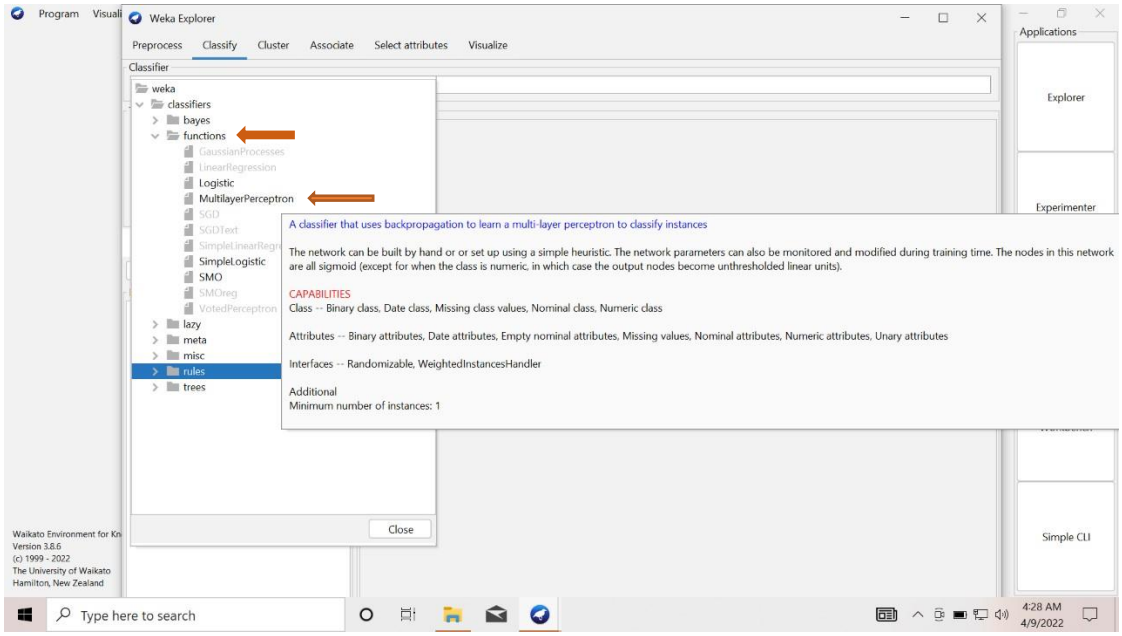


After left click in “NaiveBayes” this window will appear and we are choosing “OK” then“Start”.

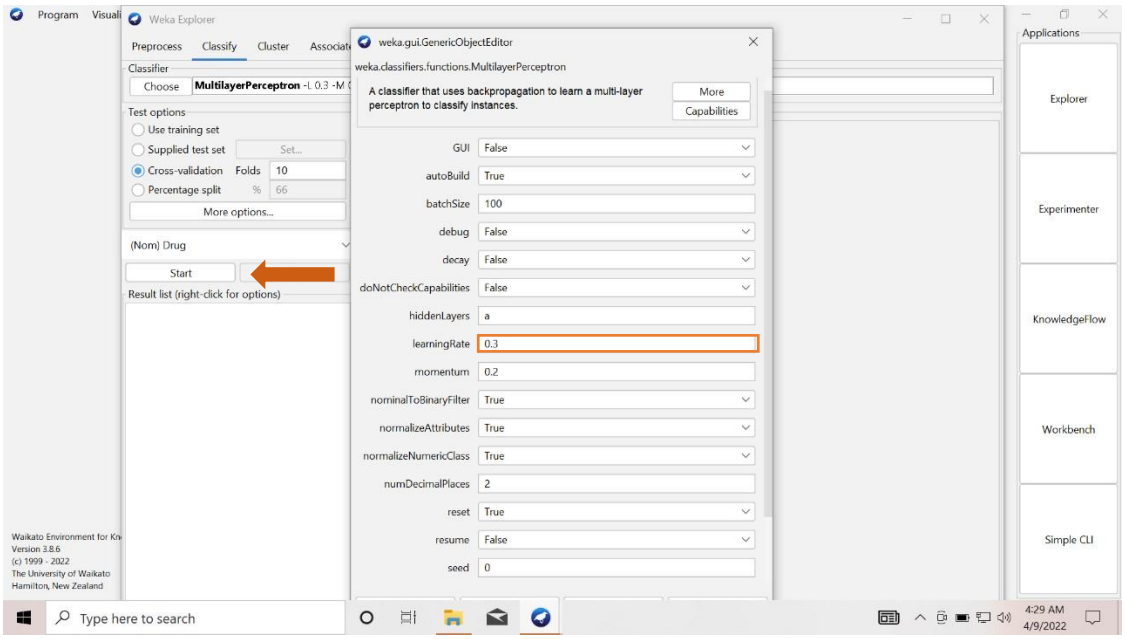


Neural Network Classification Algorithm:

Now if we want to apply Neural Network Classification Algorithm the below steps will be followed. Firstly, choosing the option “function” & then MultilayerPerceptron.

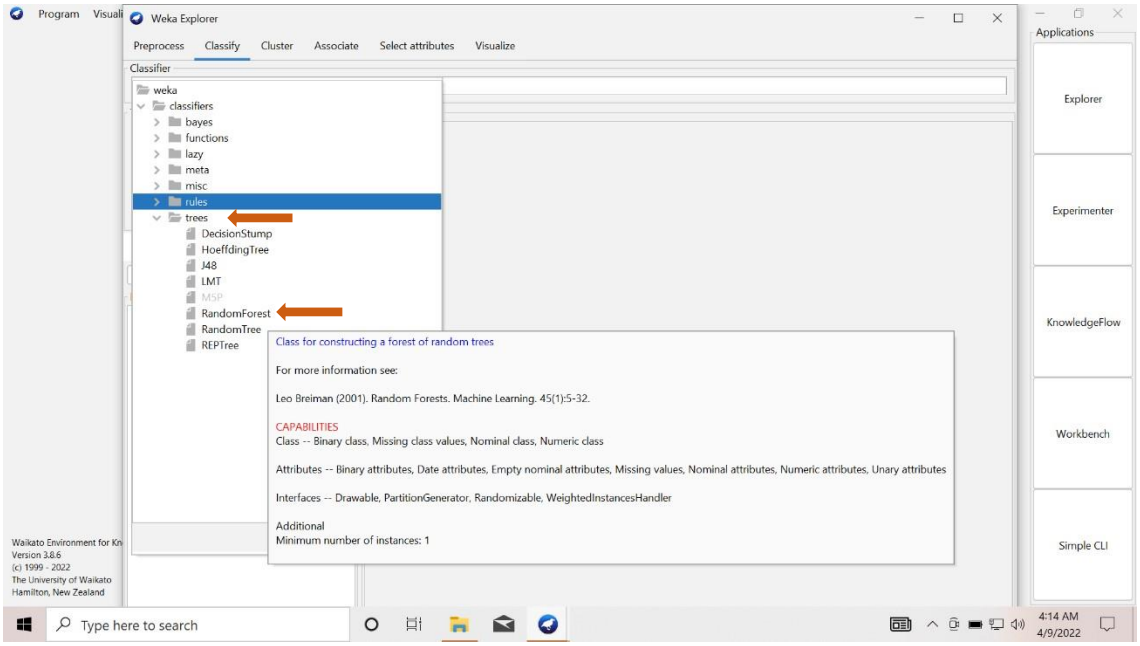


After left click in “MultilayerPerceptron” this window will appear and we are choosing “OK” then “Start”.Here learning rate is 0.3

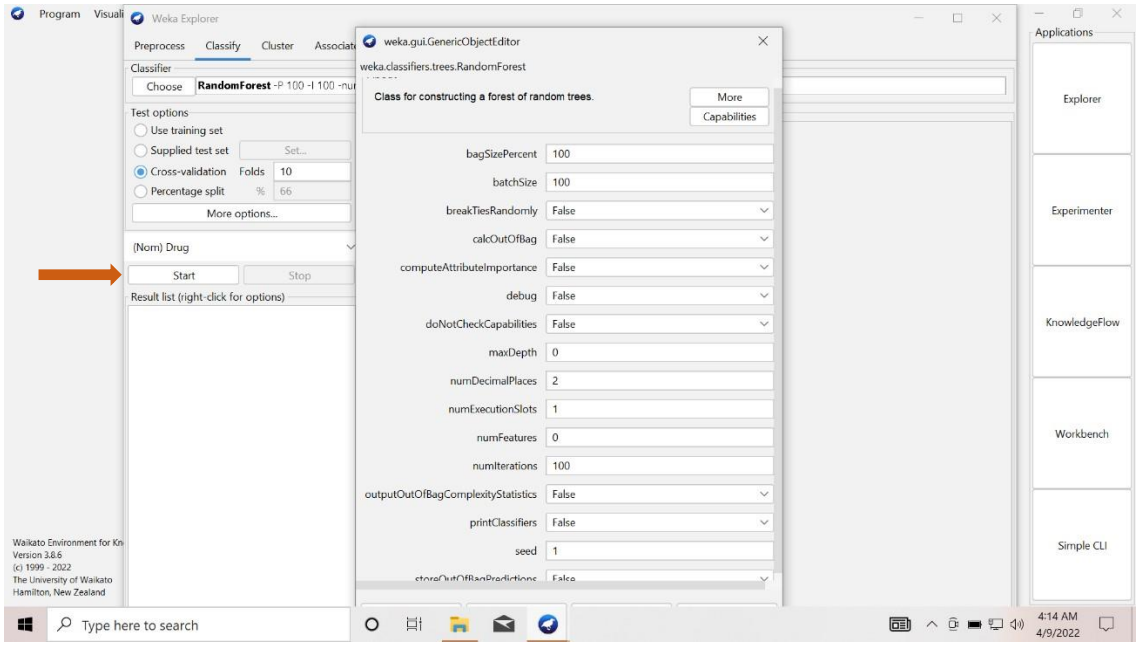


Random-Forest Classification Algorithm:

Now if we want to apply RandomForest Classification Algorithm the below steps will be followed. Firstly, choosing the option “rules” and then choosing “trees” then choosing “RandomForest”.

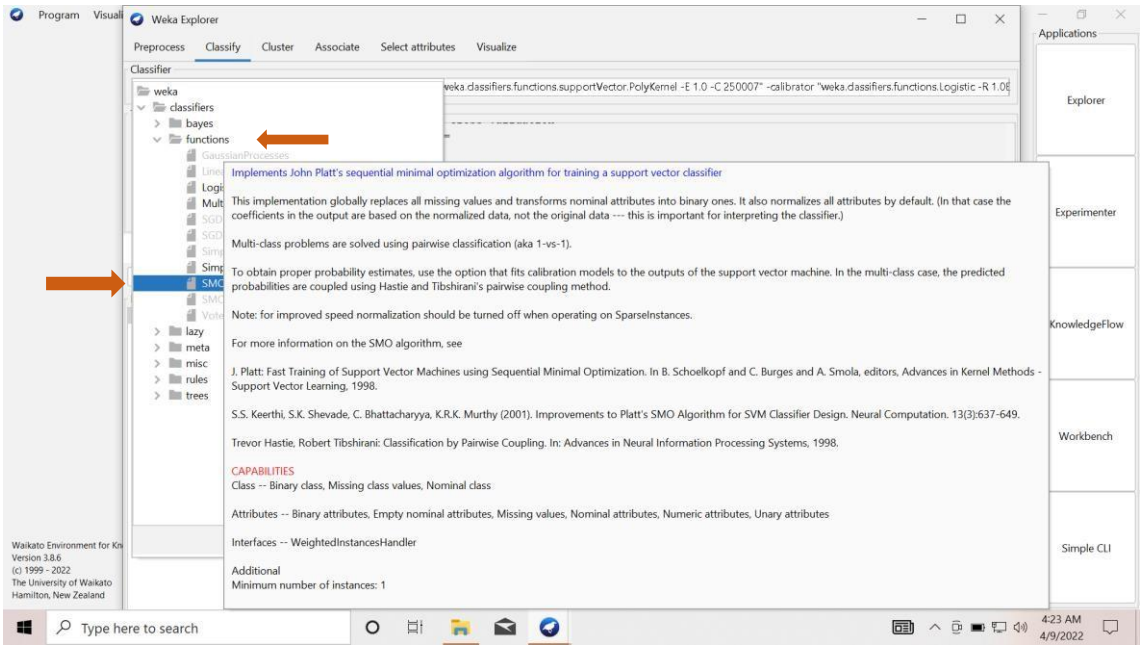


After left click in “RandomForest” this window will appear and we are choosing “OK” then “Start”.

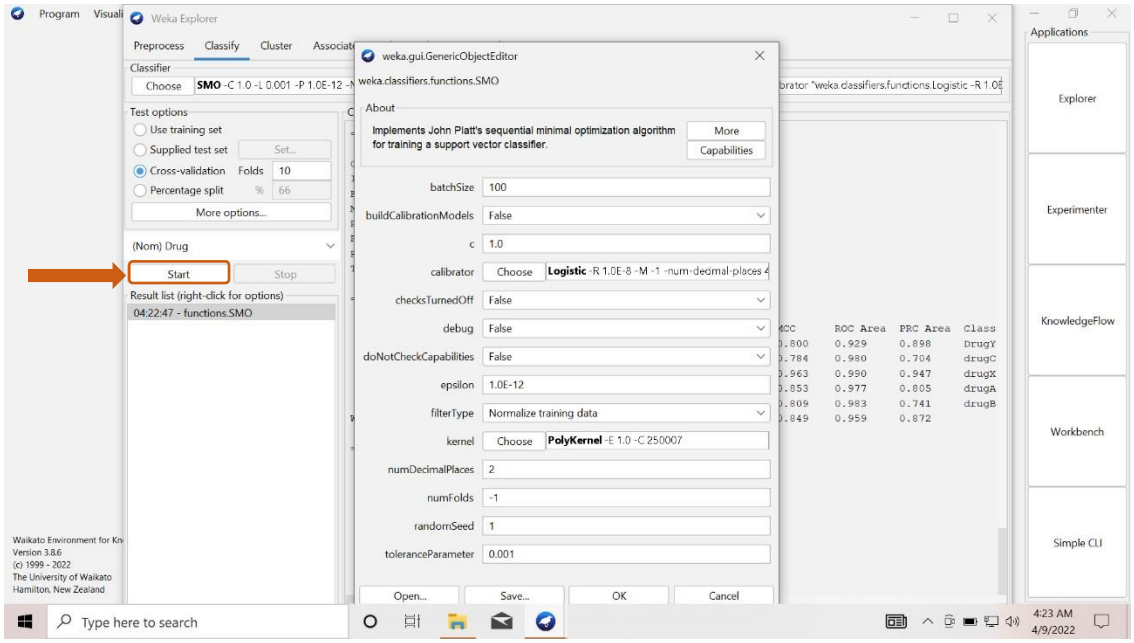


Support Vector Machine Classification Algorithm:

Now if we want to apply Support Vector Machine Classification Algorithm the below steps will be followed. Firstly, choosing the option “functions” then “SMC”.

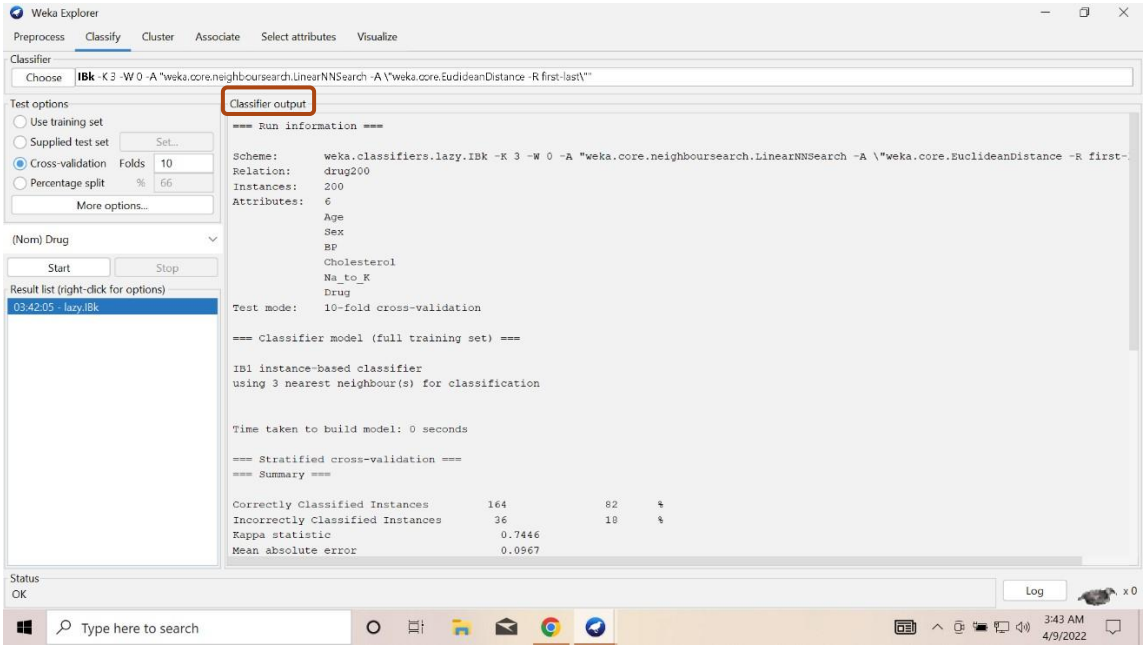


After left click in “SMO” this window will appear and we are choosing “OK” then “Start”.

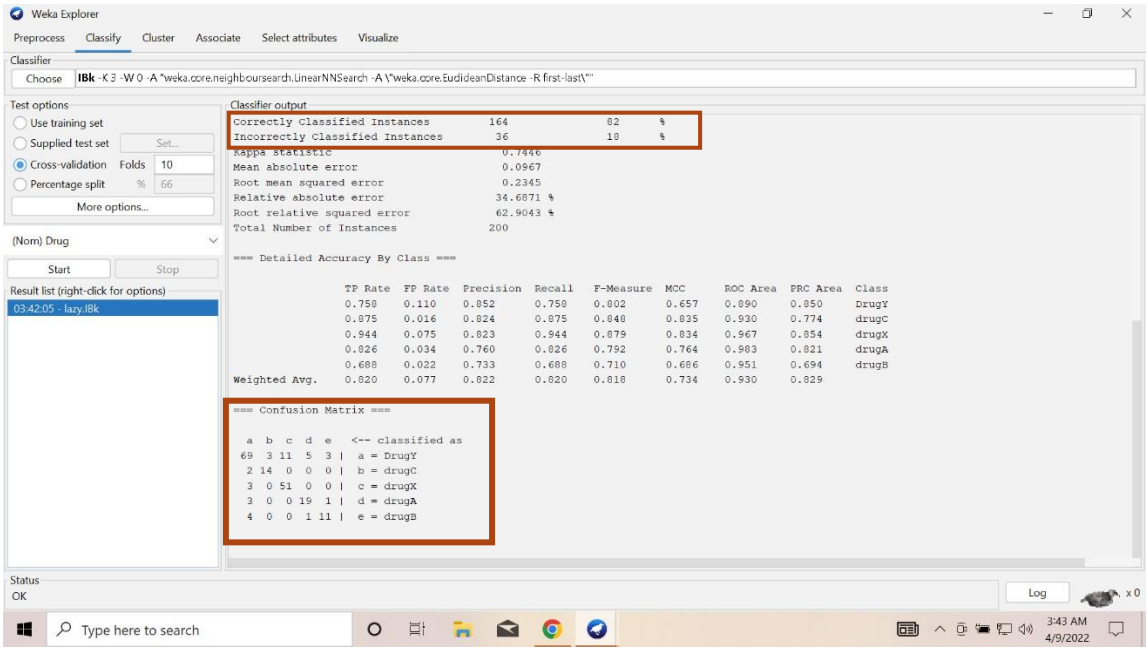


Discussion & Conclusion:

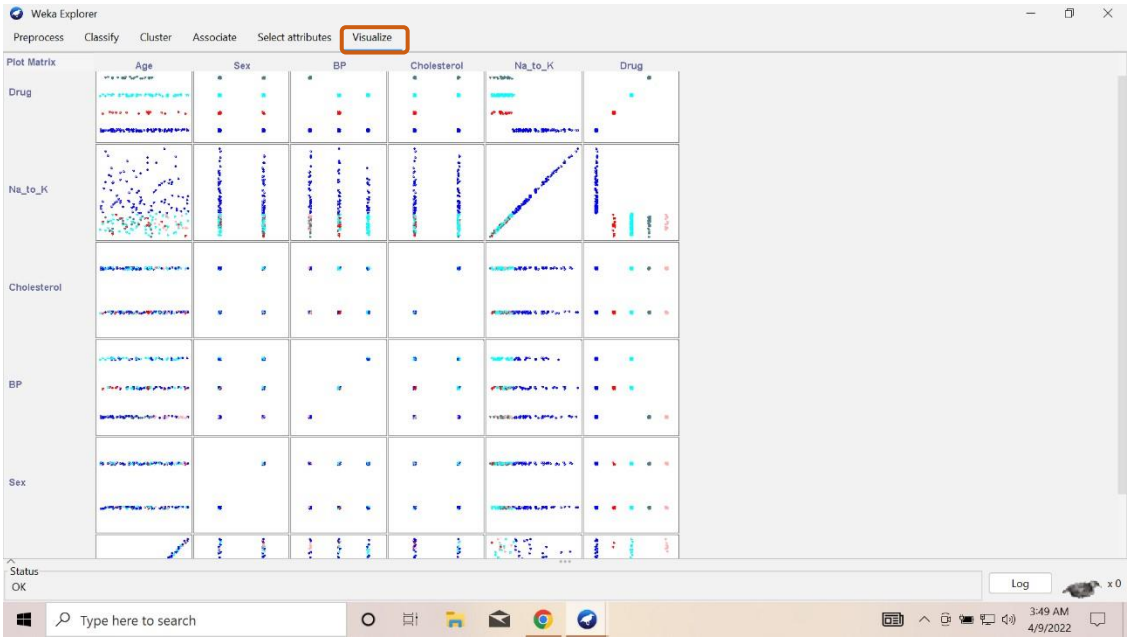
K-Nearest Neighbours:



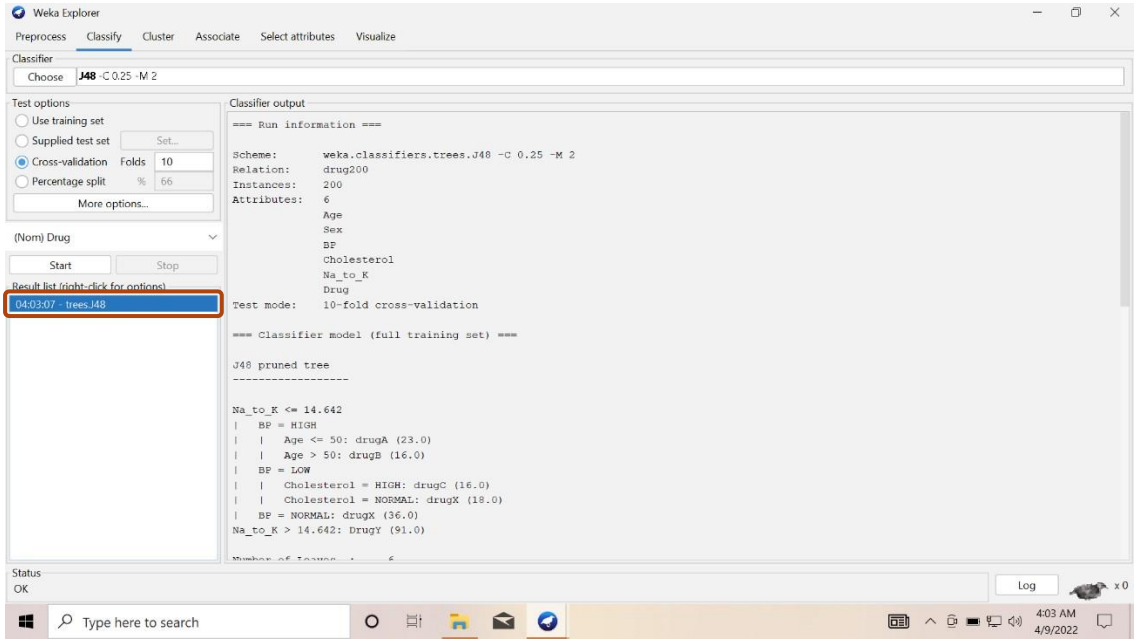
Correctly Classified Instances are 164 out of 200 instances. Incorrectly Classified Instances are 36. Accuracy correctly is 82% and incorrectly is 18%. We can also see the Confusion Matrix.



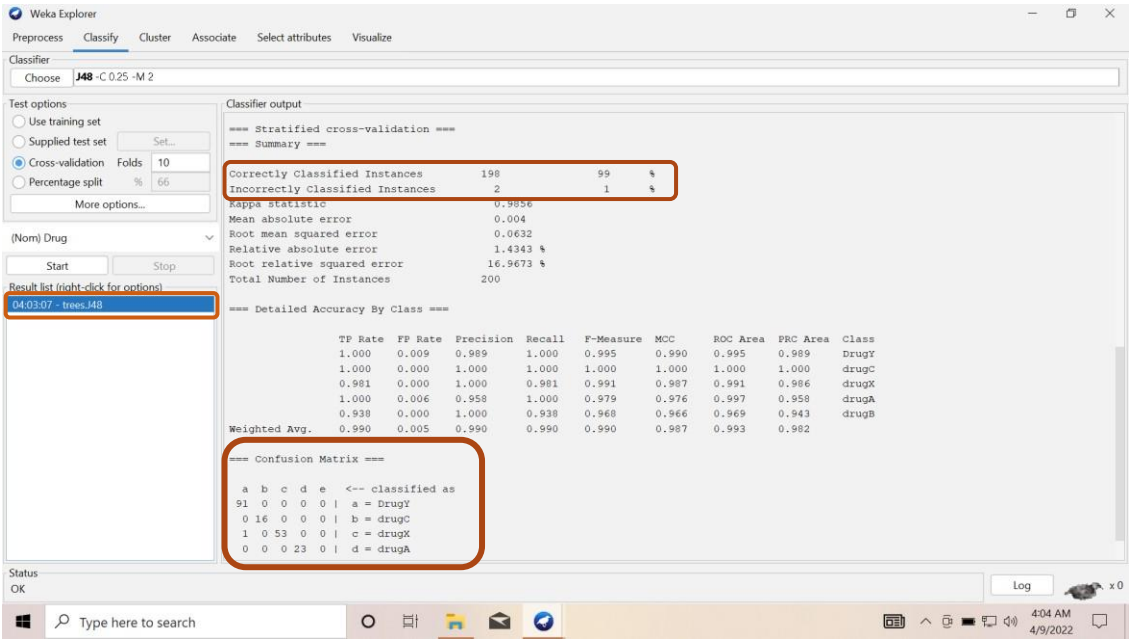
If we click on “Visualize” this interface will be shown.



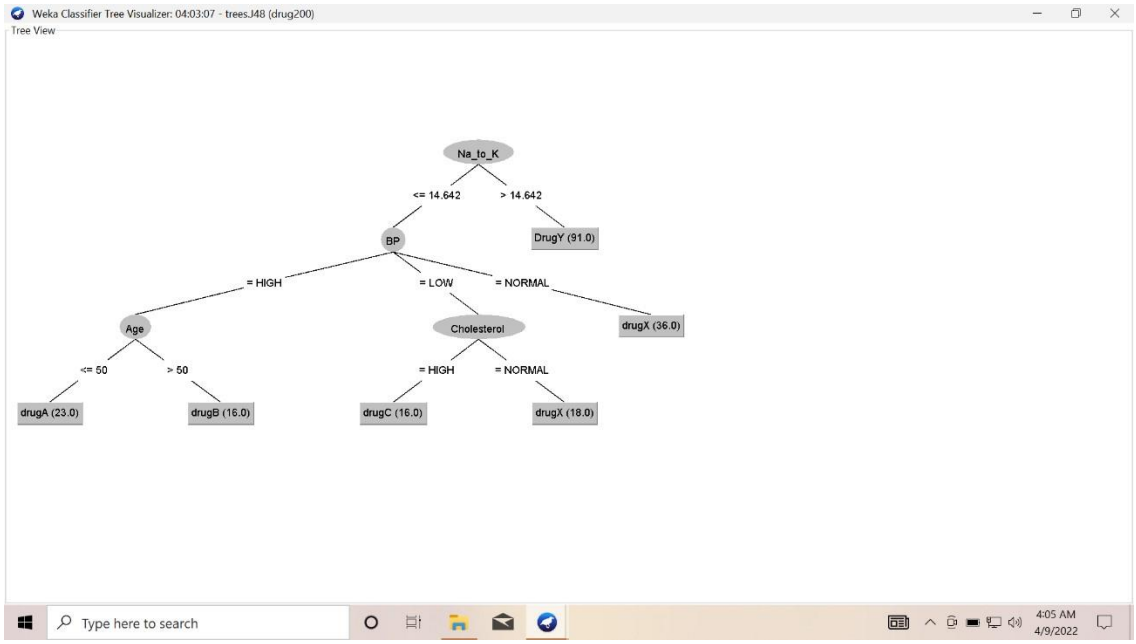
Decision Tree:



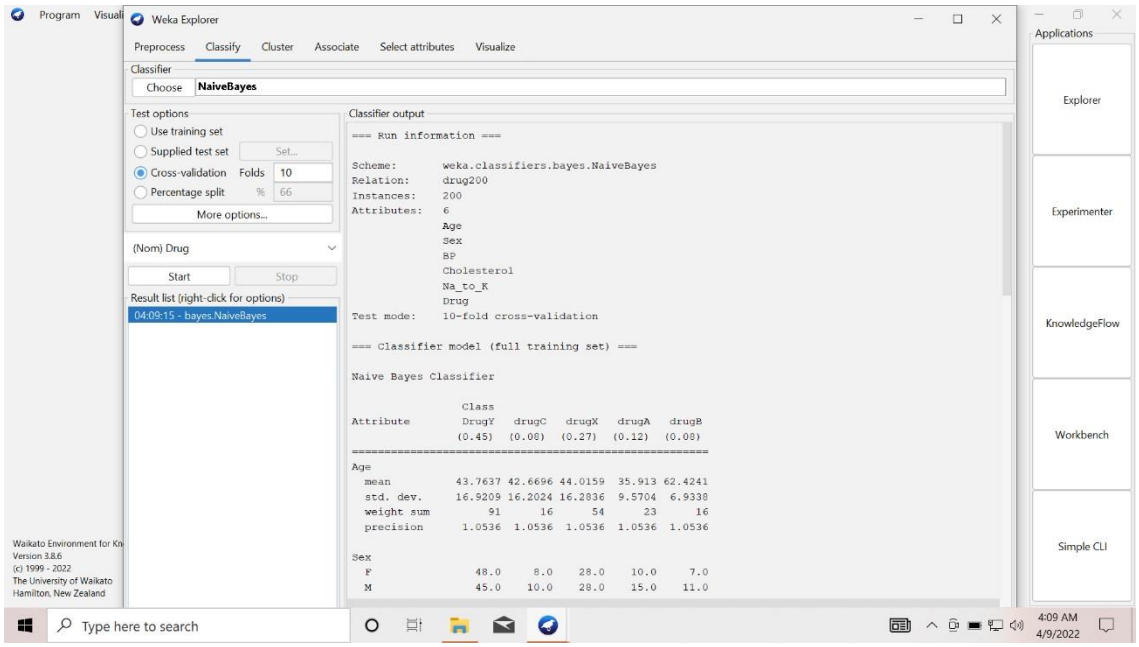
Here Correctly Classified Instances are 198 which is 99%.

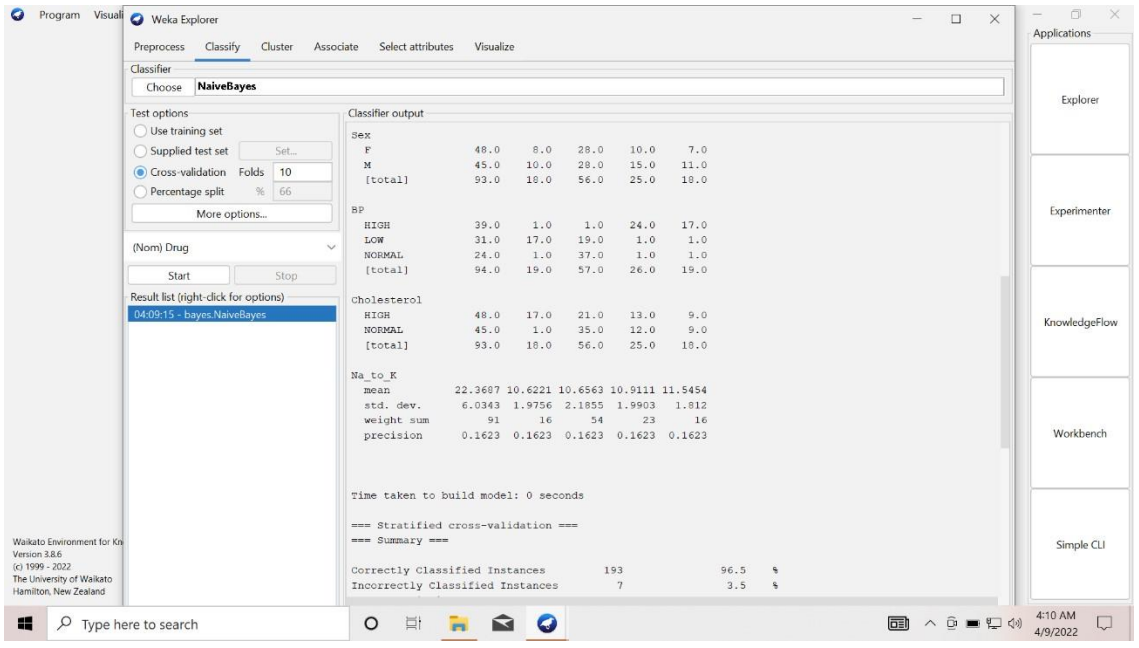


Here we can see the Tree.

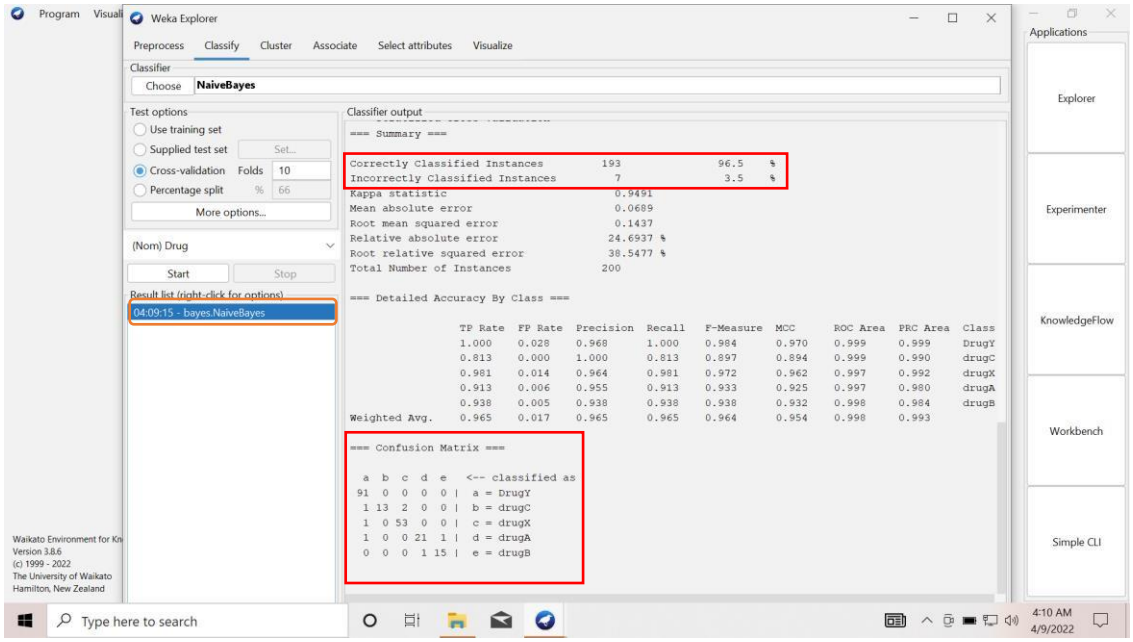


Naïve Bayes:

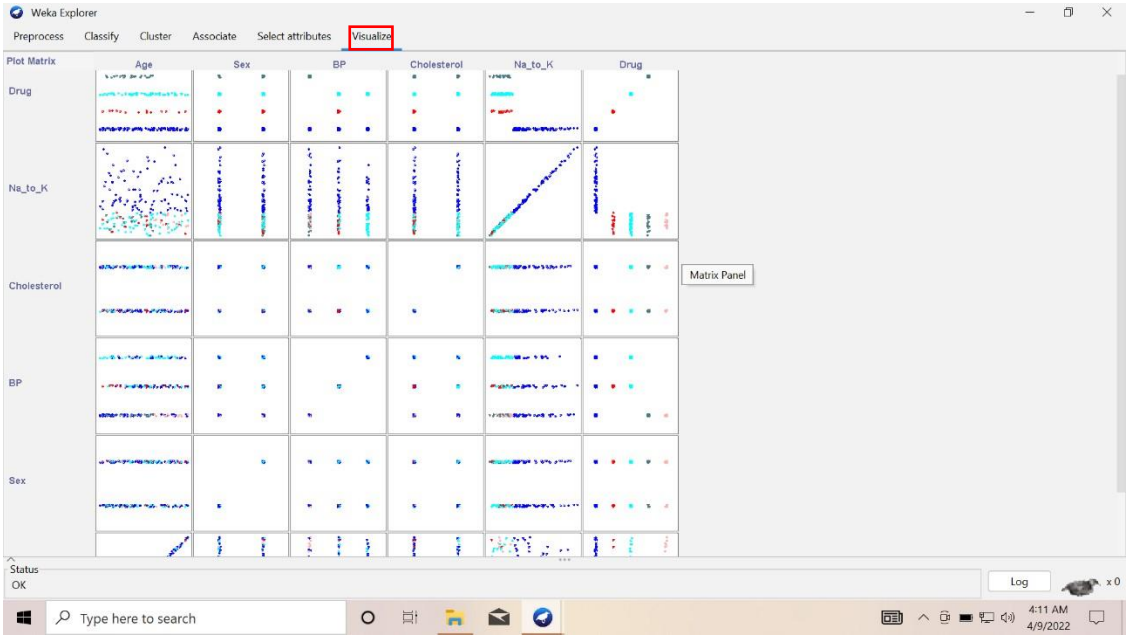

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:09:15 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: drug200
Instances: 200
Attributes: 6
Age
Sex
BP
Cholesterol
Na_to_K
Drug
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute DrugY drugC drugX drugA drugB
(0.45) (0.08) (0.27) (0.12) (0.08)
Age
mean 43.7637 42.6696 44.0159 35.913 62.4241
std. dev. 16.9209 16.2024 16.2836 9.5704 6.9338
weight sum 91 16 54 23 16
precision 1.0536 1.0536 1.0536 1.0536 1.0536
Sex
F 48.0 8.0 28.0 10.0 7.0
M 45.0 10.0 28.0 15.0 11.0
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Type here to search
4:09 AM
4/9/2022


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:09:15 - bayes.NaiveBayes
Classifier output
Sex
F 48.0 8.0 28.0 10.0 7.0
M 45.0 10.0 28.0 15.0 11.0
[total] 93.0 18.0 56.0 25.0 18.0
BP
HIGH 39.0 1.0 1.0 24.0 17.0
LOW 31.0 17.0 19.0 1.0 1.0
NORMAL 24.0 1.0 37.0 1.0 1.0
[total] 94.0 19.0 57.0 26.0 19.0
Cholesterol
HIGH 48.0 17.0 21.0 13.0 9.0
NORMAL 45.0 1.0 35.0 12.0 9.0
[total] 93.0 18.0 56.0 25.0 18.0
Na_to_K
mean 22.3687 10.6221 10.6563 10.9111 11.5454
std. dev. 6.0343 1.9756 2.1855 1.9903 1.812
weight sum 91 16 54 23 16
precision 0.1623 0.1623 0.1623 0.1623 0.1623
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 193 96.5 %
Incorrectly Classified Instances 7 3.5 %
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Type here to search
4:10 AM
4/9/2022

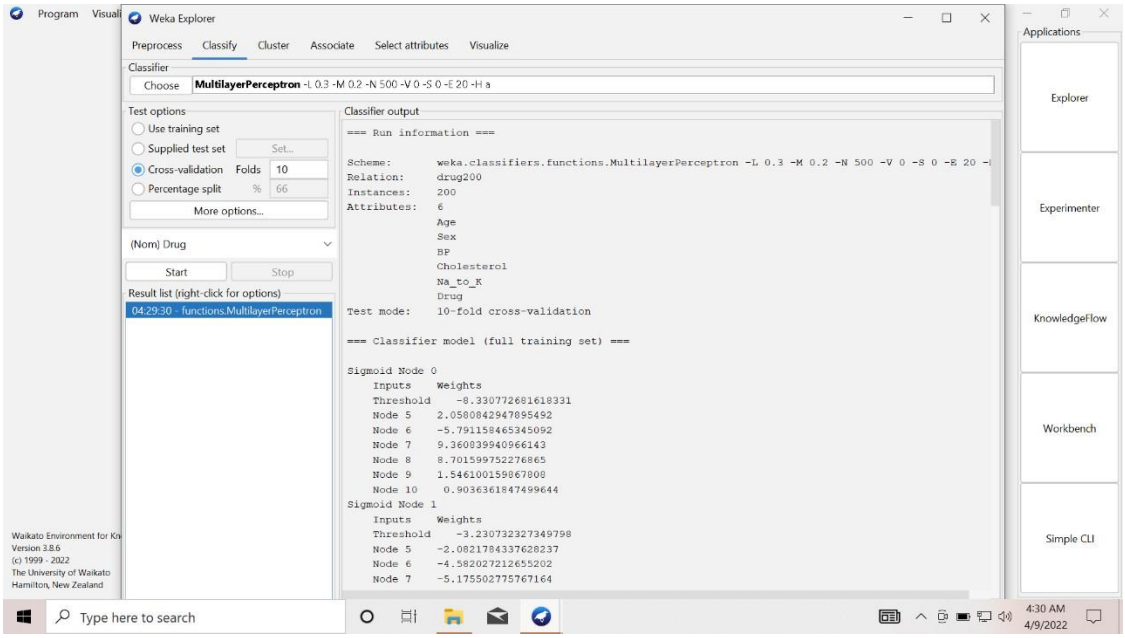
Correctly Classified Instances are 193 out of 200 instances. Incorrectly Classified Instances are 7. Accuracy is 96.5% and incorrectly is 3.5%. We can also see the Confusion Matrix.

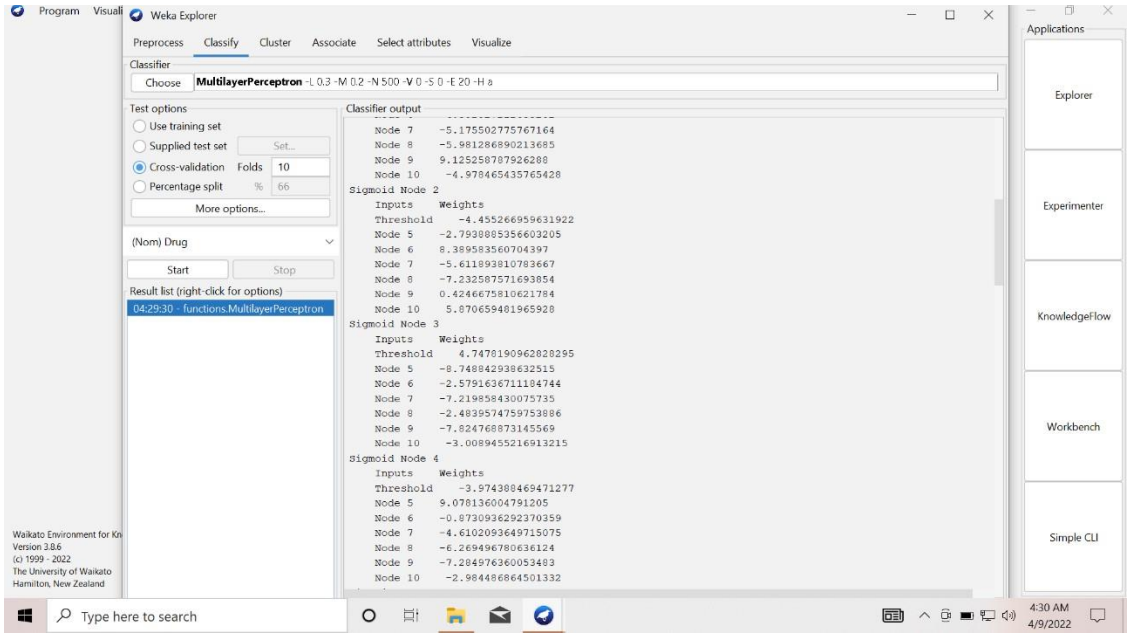


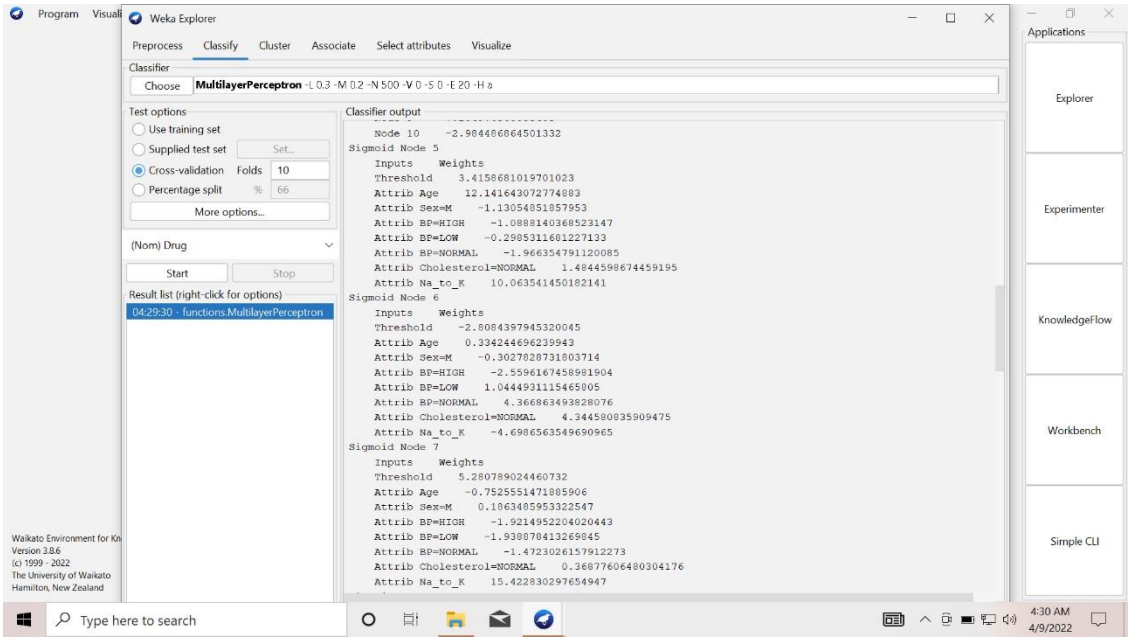
To Visualize click on the “Visualize” Option

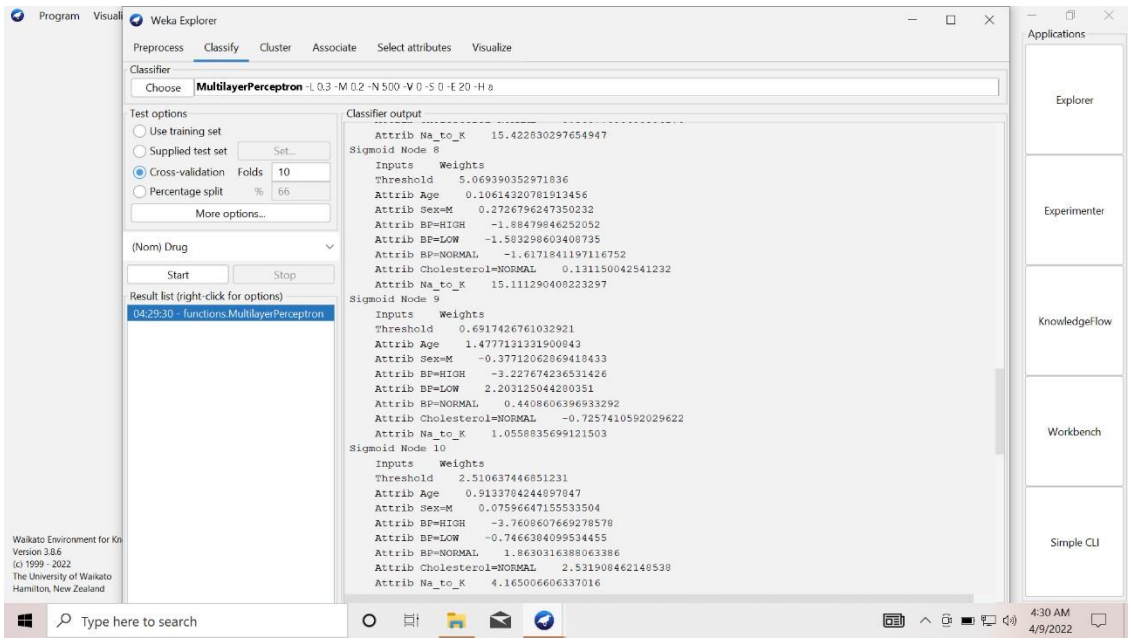


Neural Network:

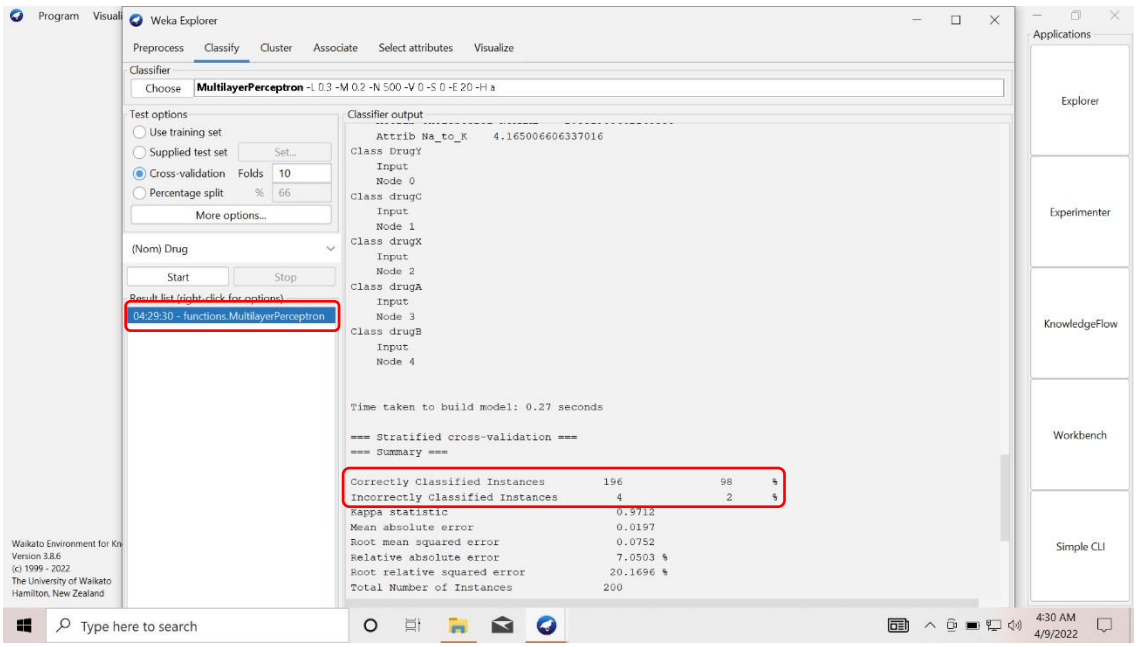


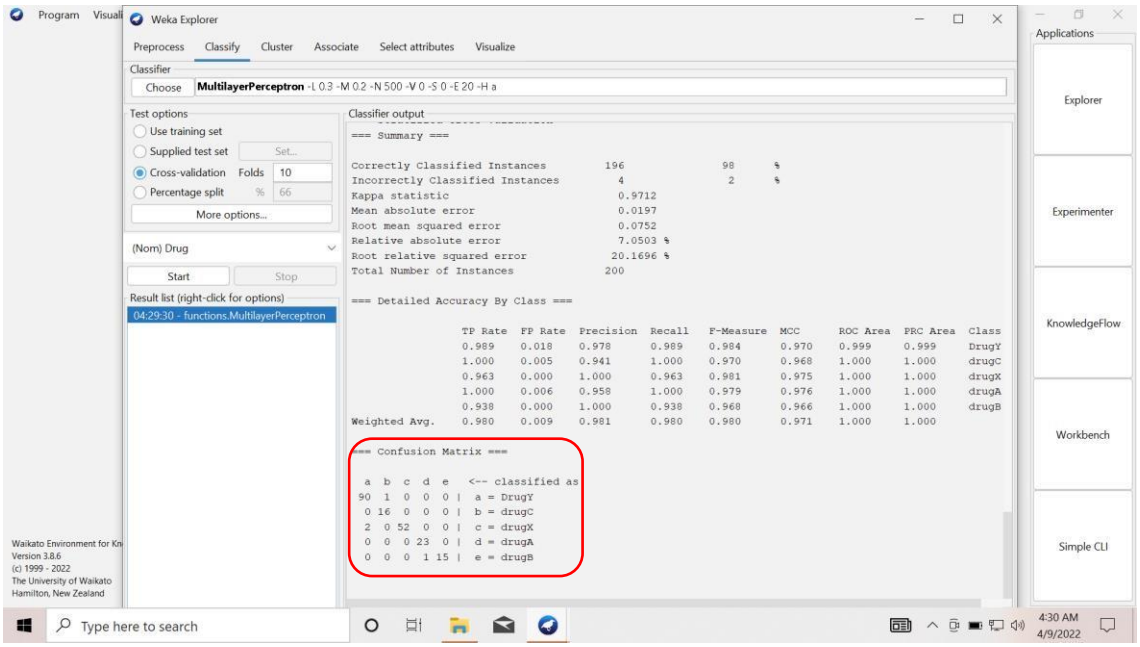


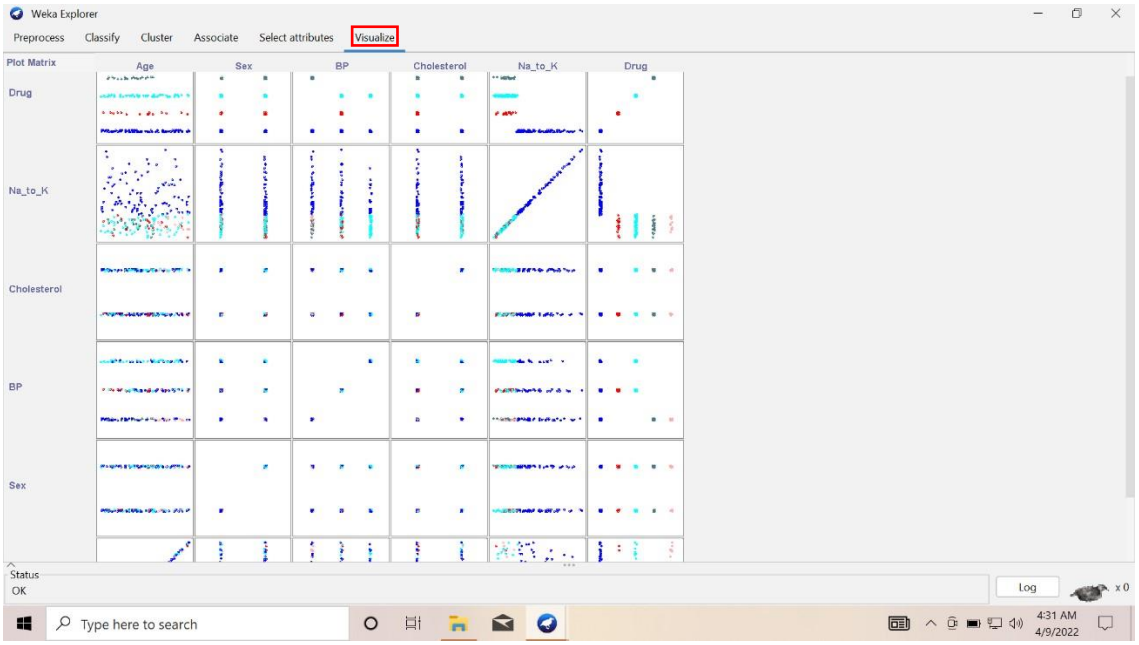

Program
Visuali
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:29:30 - functions.MultilayerPerceptron
Classifier output
Node 10 -2.984486864501332
Sigmoid Node 5
Inputs Weights
Threshold 3.4158681019701023
Attrib Age 12.141643072774883
Attrib Sex=M -1.13054851857953
Attrib BP=HIGH -1.0888140368523147
Attrib BP=LOW -0.2985311681227133
Attrib BP=NORMAL -1.966354791120085
Attrib Cholesterol=NORMAL 1.48445986744591955
Attrib Na_to_K 10.063541450182141
Sigmoid Node 6
Inputs Weights
Threshold -2.8084397945320045
Attrib Age 0.334244696239943
Attrib Sex=M -0.3027028731803714
Attrib BP=HIGH -2.5596167458981904
Attrib BP=LOW 1.0444931115465805
Attrib BP=NORMAL 4.366863493828076
Attrib Cholesterol=NORMAL 4.344580835909475
Attrib Na_to_K -4.6986563549690965
Sigmoid Node 7
Inputs Weights
Threshold 5.280789024460732
Attrib Age -0.7525551471085906
Attrib Sex=M 0.1863485953322547
Attrib BP=HIGH -1.9214952204020443
Attrib BP=LOW -1.938878413269845
Attrib BP=NORMAL -1.4723026157912273
Attrib Cholesterol=NORMAL 0.36877606480304176
Attrib Na_to_K 15.422830297654947
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Type here to search
4:30 AM
4/9/2022


Program
Visuali
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:29:30 - functions.MultilayerPerceptron
Classifier output
Attrib Na_to_K 15.422830297654947
Sigmoid Node 8
Inputs Weights
Threshold 5.069390352971836
Attrib Age 0.10614320781913456
Attrib Sex=M 0.27267962473502332
Attrib BP=HIGH -1.88479846252052
Attrib BP=LOW -1.583298603408735
Attrib BP=NORMAL -1.6171841197116752
Attrib Cholesterol=NORMAL 0.131150042541232
Attrib Na_to_K 15.111290408223297
Sigmoid Node 9
Inputs Weights
Threshold 0.6917426761032921
Attrib Age 1.477713133190043
Attrib Sex=M -0.37712062869418433
Attrib BP=HIGH -3.227674236531426
Attrib BP=LOW 2.203125044280351
Attrib BP=NORMAL 0.4408606396933292
Attrib Cholesterol=NORMAL -0.7257410592029622
Attrib Na_to_K 1.0558835699121503
Sigmoid Node 10
Inputs Weights
Threshold 2.5106374468512231
Attrib Age 0.9133784244897847
Attrib Sex=M 0.07596647155533504
Attrib BP=HIGH -3.7608607669278578
Attrib BP=LOW -0.7466384099534455
Attrib BP=NORMAL 1.8630316388063386
Attrib Cholesterol=NORMAL 2.531908462148538
Attrib Na_to_K 4.165006606337016
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Type here to search
4:30 AM
4/9/2022

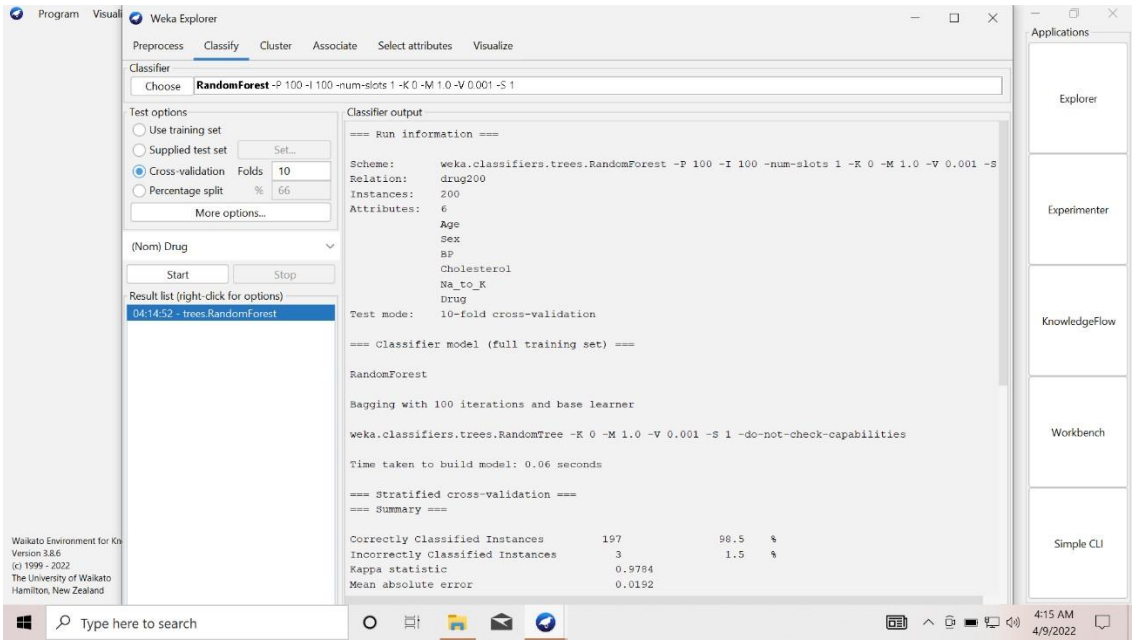
Correctly Classified Instances are 196 out of 200 instances. Incorrectly Classified Instances are 7. Accuracy is 98% and incorrectly is 2%. We can also see the Confusion Matrix.



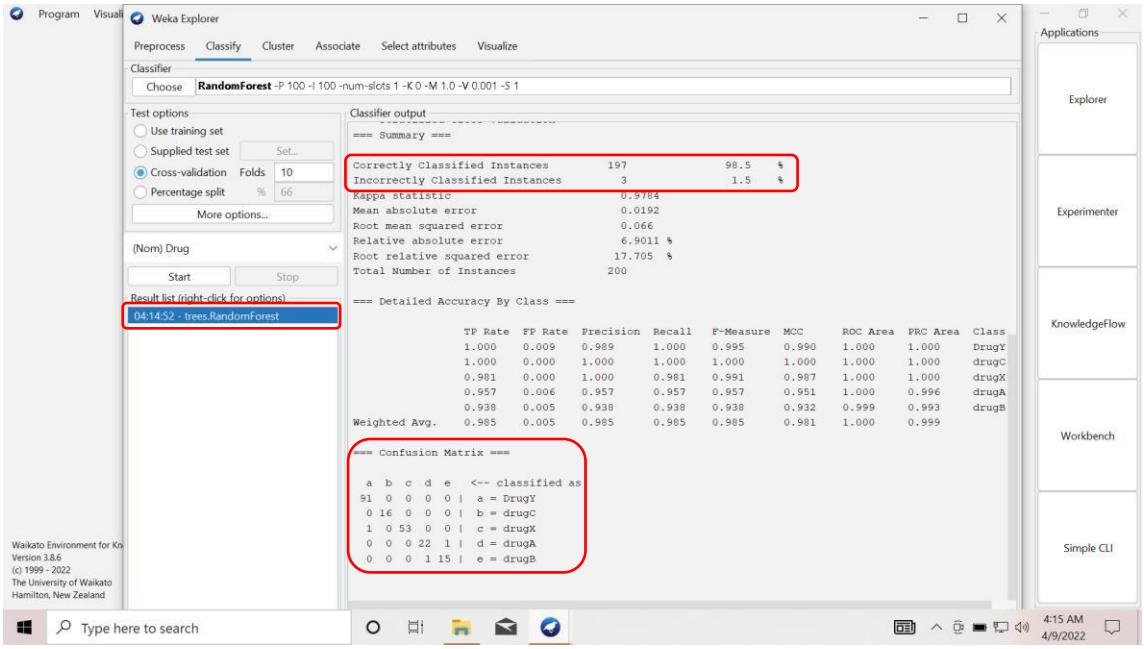




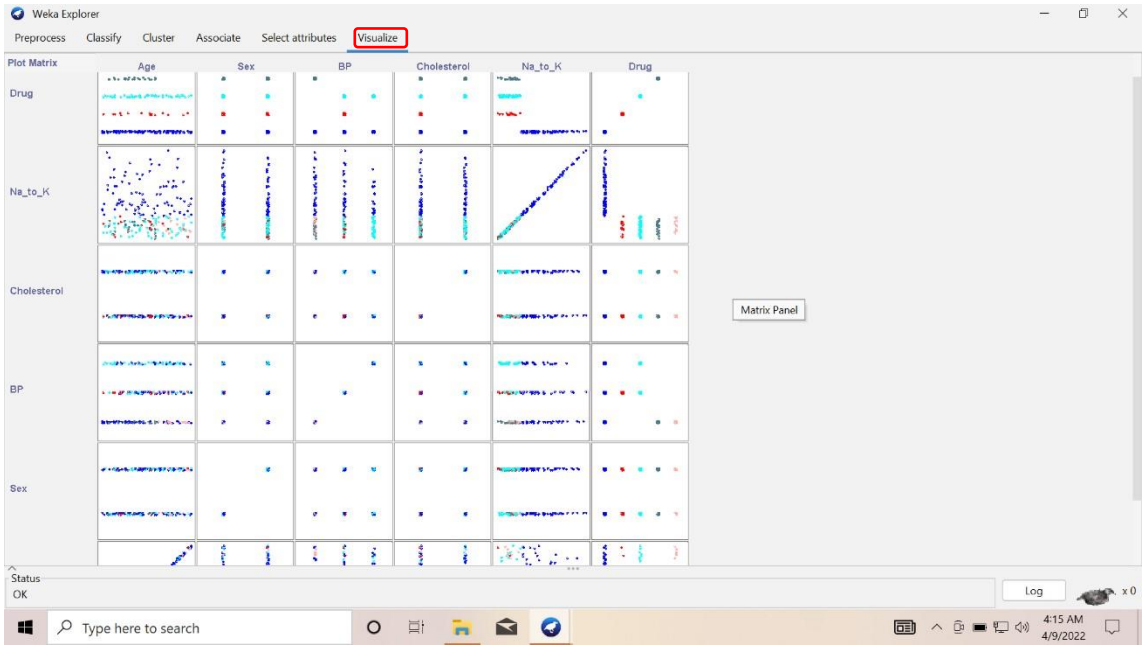
Random Forest:



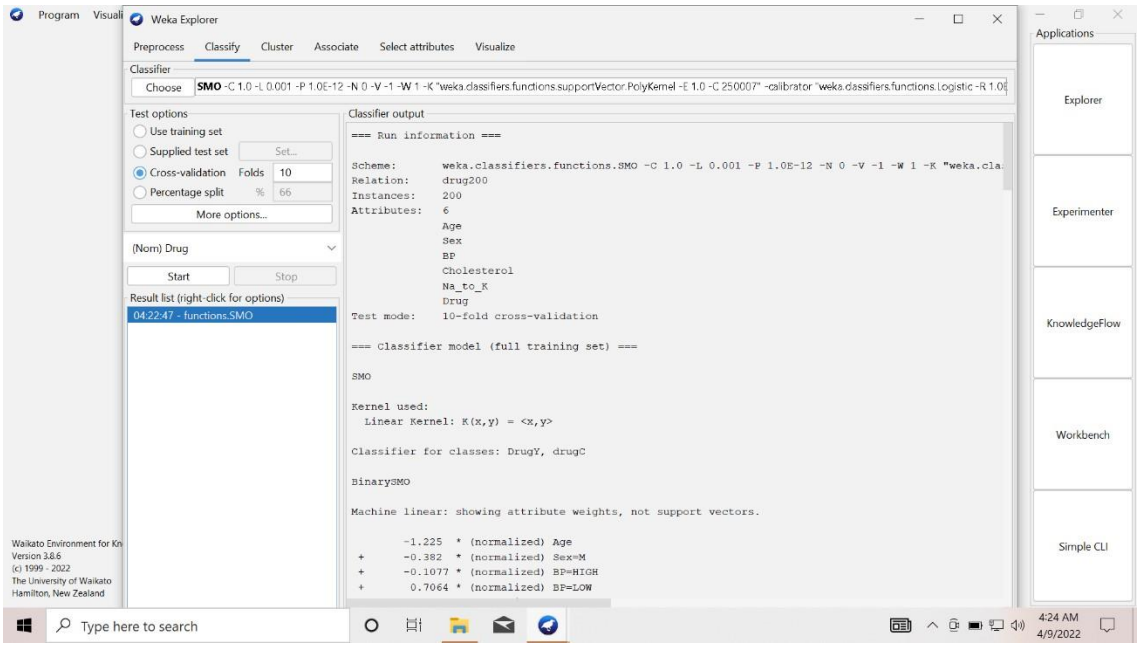
Correctly Classified Instances are 197 out of 200 instances. Incorrectly Classified Instances are 3.Accuracy is 98.5% and incorrectly is 1.5%. We can also see the Confusion Matrix

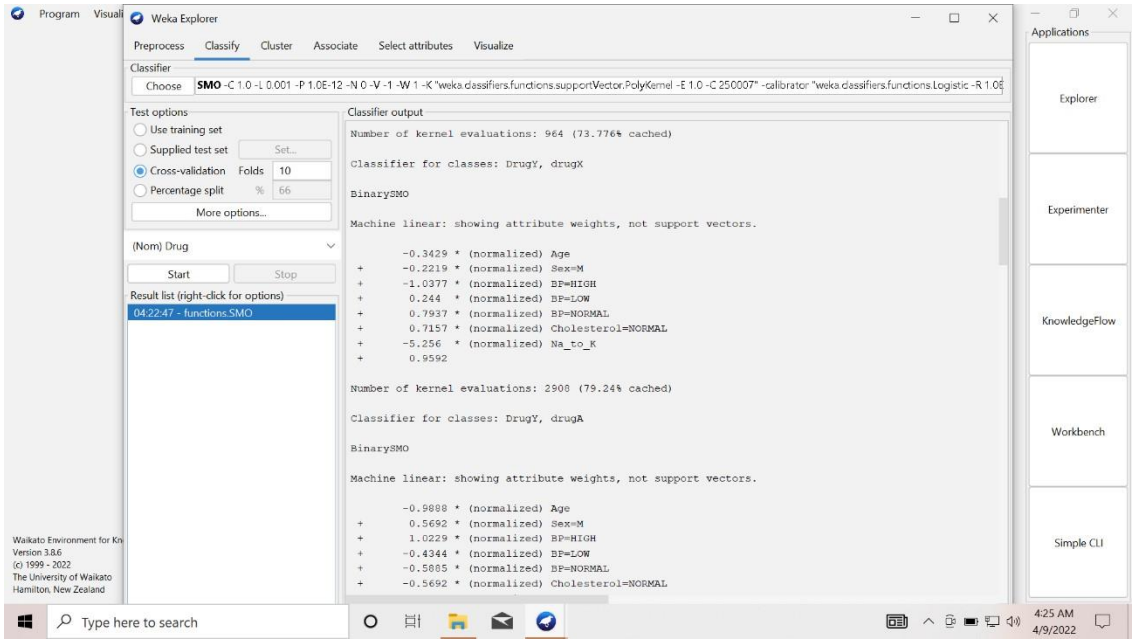


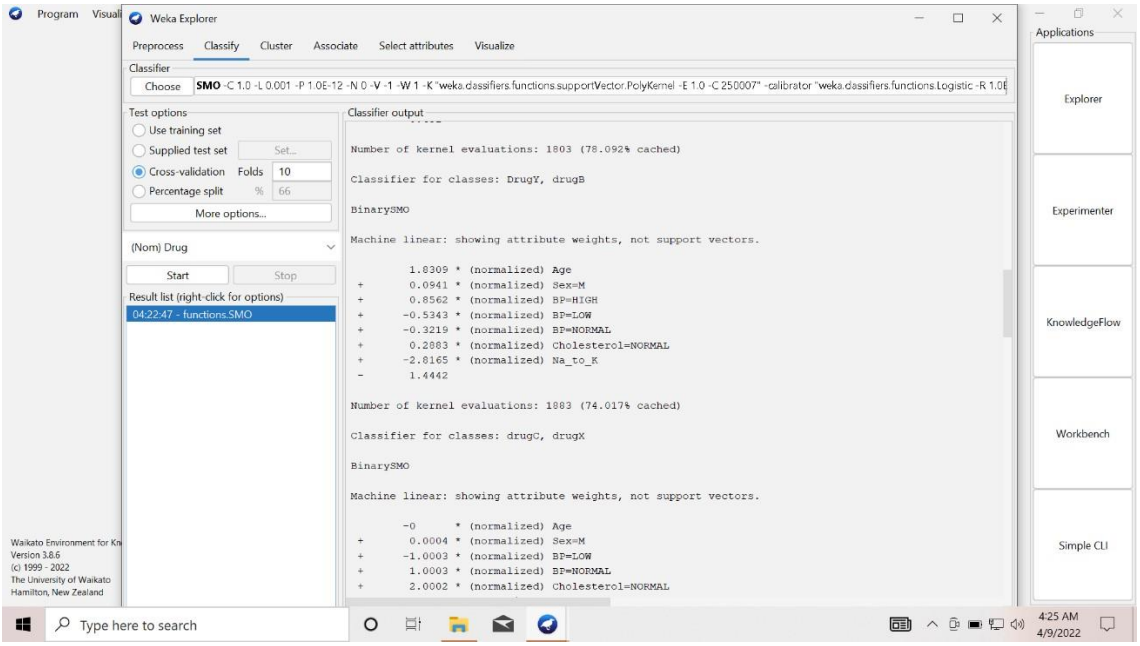
Clicking “Visualize” option to visualize data

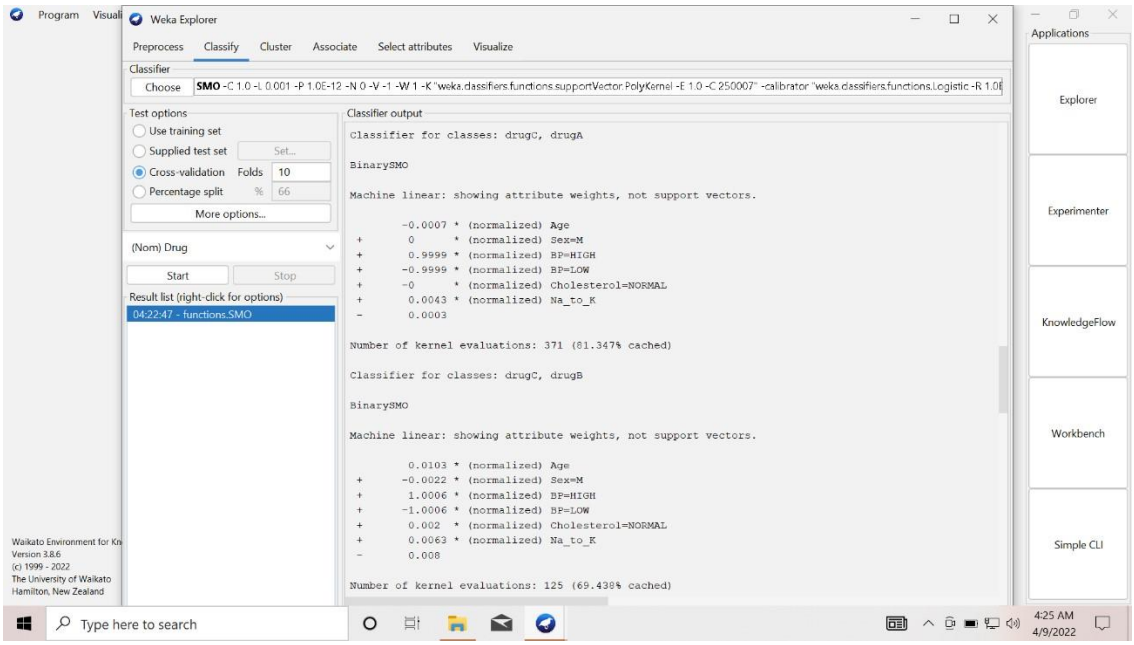


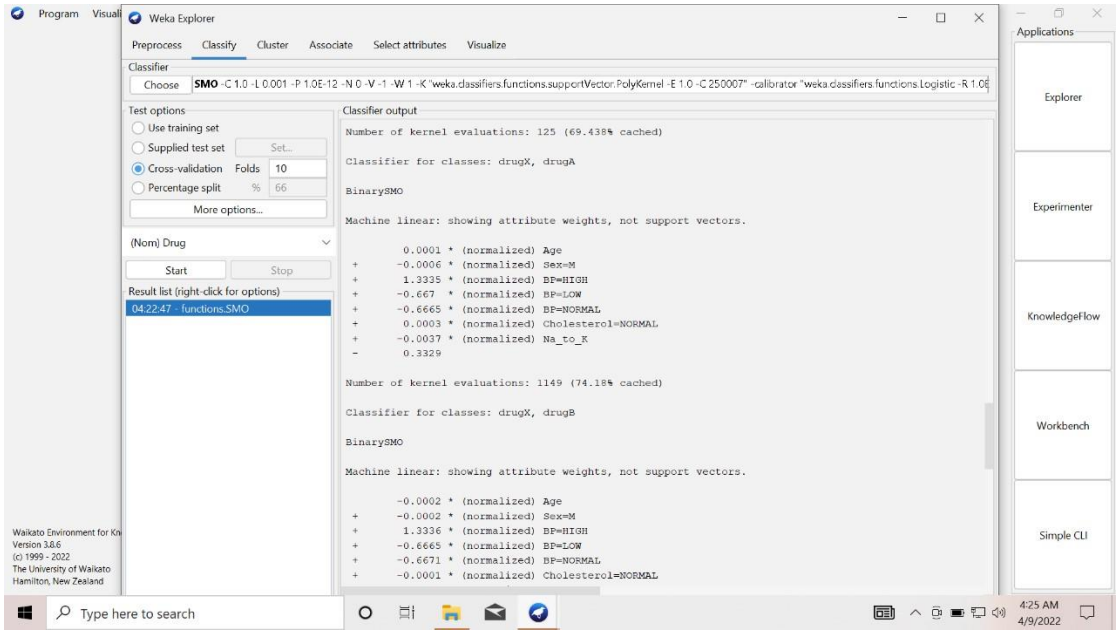
Support Vector Machine:

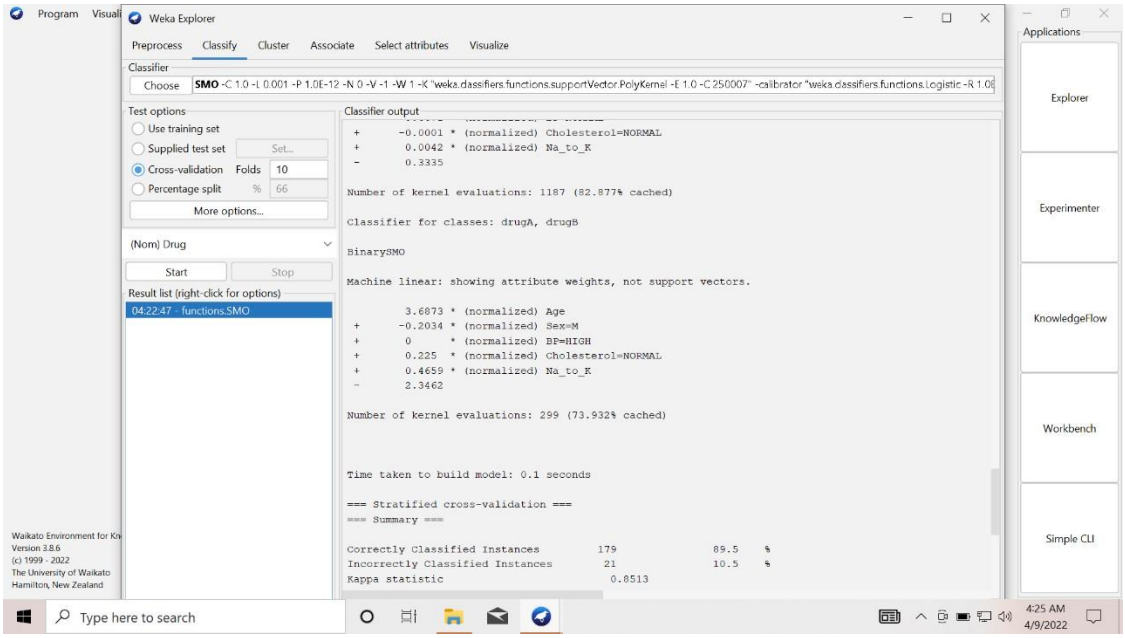




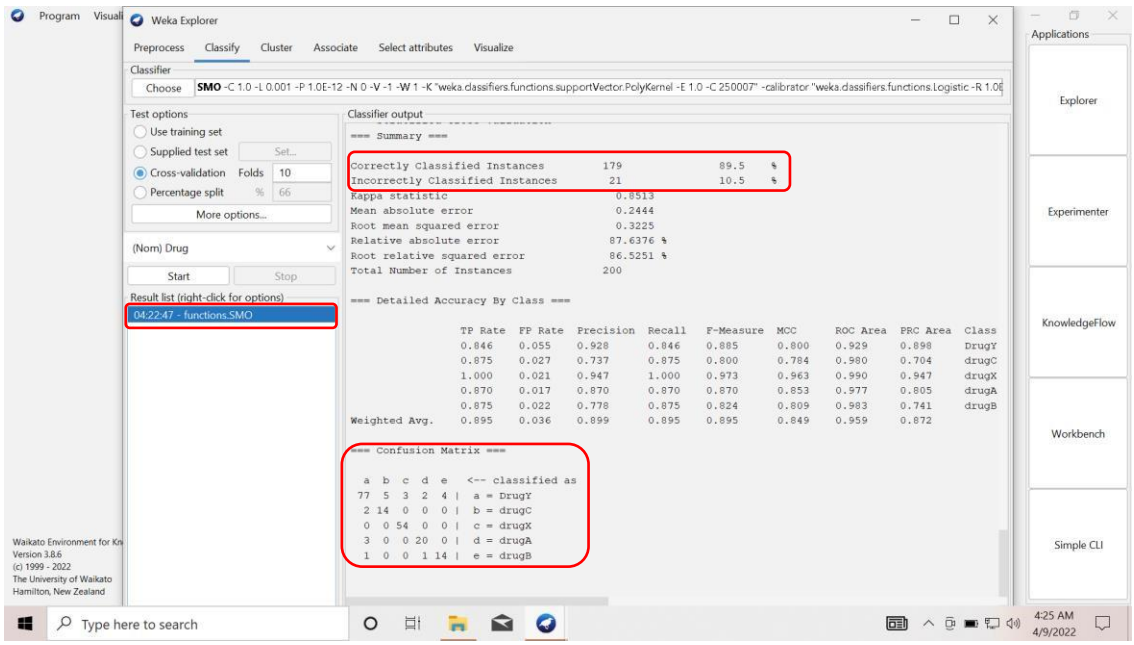

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:22:47 - functions.SMO
Classifier output
Number of kernel evaluations: 1803 (78.092% cached)
Classifier for classes: DrugY, drugB
BinarySMO
Machine linear: showing attribute weights, not support vectors.
1.8309 * (normalized) Age
+ 0.0941 * (normalized) Sex=M
+ 0.8562 * (normalized) BP=HIGH
+ -0.5343 * (normalized) BP=LOW
+ -0.3219 * (normalized) BP=NORMAL
+ 0.2883 * (normalized) Cholesterol=NORMAL
+ -2.8165 * (normalized) Na_to_K
- 1.4442
Number of kernel evaluations: 1883 (74.017% cached)
Classifier for classes: drugC, drugX
BinarySMO
Machine linear: showing attribute weights, not support vectors.
-0 * (normalized) Age
+ 0.0004 * (normalized) Sex=M
+ -1.0003 * (normalized) BP=LOW
+ 1.0003 * (normalized) BP=NORMAL
+ 2.0002 * (normalized) Cholesterol=NORMAL
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Type here to search
4:25 AM
4/9/2022


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:22:47 - functions.SMO
Classifier output
Classifier for classes: drugC, drugA
BinarySMO
Machine linear: showing attribute weights, not support vectors.
-0.0007 * (normalized) Age
+ 0 * (normalized) Sex=M
+ 0.9999 * (normalized) BP=HIGH
+ -0.9999 * (normalized) BP=LOW
+ -0 * (normalized) Cholesterol=NORMAL
+ 0.0043 * (normalized) Na_to_K
- 0.0003
Number of kernel evaluations: 371 (81.347% cached)
Classifier for classes: drugC, drugB
BinarySMO
Machine linear: showing attribute weights, not support vectors.
0.0103 * (normalized) Age
+ -0.0022 * (normalized) Sex=M
+ 1.0006 * (normalized) BP=HIGH
+ -1.0006 * (normalized) BP=LOW
+ 0.002 * (normalized) Cholesterol=NORMAL
+ 0.0063 * (normalized) Na_to_K
- 0.008
Number of kernel evaluations: 125 (69.438% cached)
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Type here to search
4:25 AM
4/9/2022

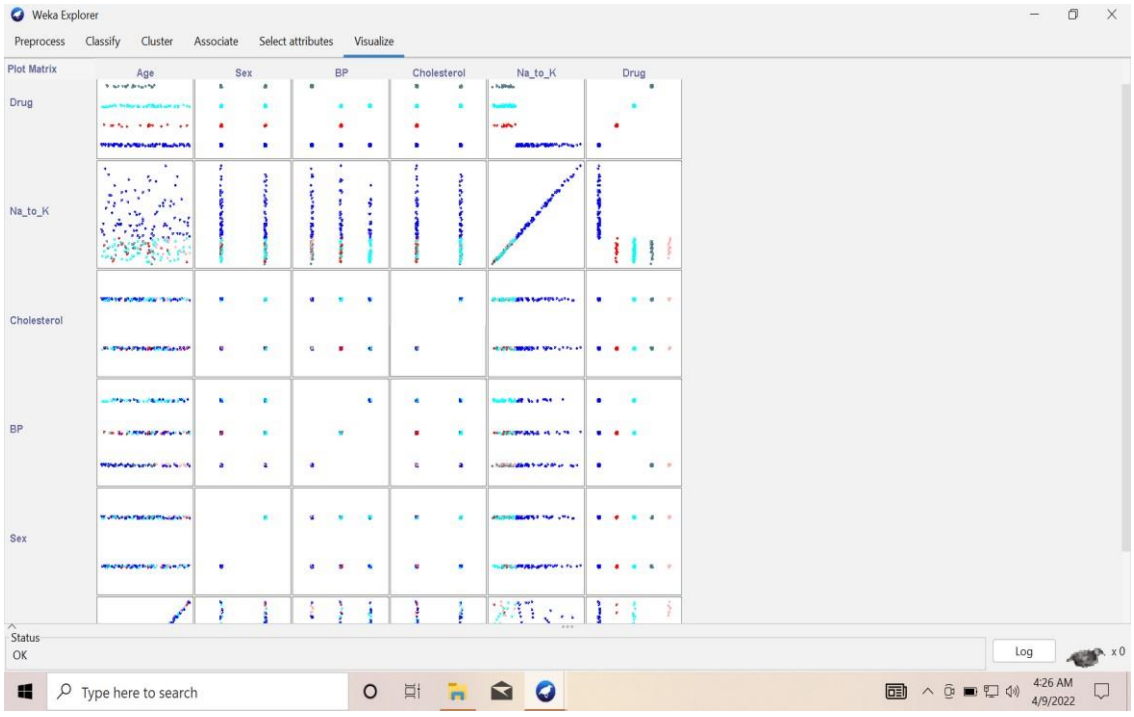

Program
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:22:47 - functions.SMO
Classifier output
Number of kernel evaluations: 125 (69.438% cached)
Classifier for classes: drugX, drugA
BinarySMO
Machine linear: showing attribute weights, not support vectors.
0.0001 * (normalized) Age
+ -0.0006 * (normalized) Sex=M
+ 1.3335 * (normalized) BP=HIGH
+ -0.667 * (normalized) BP=LOW
+ -0.6665 * (normalized) BP=NORMAL
+ 0.0003 * (normalized) Cholesterol=NORMAL
+ -0.0037 * (normalized) Na_to_K
- 0.3329
Number of kernel evaluations: 1149 (74.18% cached)
Classifier for classes: drugX, drugB
BinarySMO
Machine linear: showing attribute weights, not support vectors.
-0.0002 * (normalized) Age
+ -0.0002 * (normalized) Sex=M
+ 1.3336 * (normalized) BP=HIGH
+ -0.6665 * (normalized) BP=LOW
+ -0.6671 * (normalized) BP=NORMAL
+ -0.0001 * (normalized) Cholesterol=NORMAL
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Type here to search
4:25 AM
4/9/2022


Program
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Drug
Start
Stop
Result list (right-click for options)
04:22:47 - functions.SMO
Classifier output
+ -0.0001 * (normalized) Cholesterol=NORMAL
+ 0.0042 * (normalized) Na_to_K
- 0.3335
Number of kernel evaluations: 1187 (82.877% cached)
Classifier for classes: drugA, drugB
BinarySMO
Machine linear: showing attribute weights, not support vectors.
3.6873 * (normalized) Age
+ -0.2034 * (normalized) Sex=M
+ 0 * (normalized) BP=HIGH
+ 0.225 * (normalized) Cholesterol=NORMAL
+ 0.4659 * (normalized) Na_to_K
- 2.3462
Number of kernel evaluations: 299 (73.932% cached)
Time taken to build model: 0.1 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 179 89.5 %
Incorrectly Classified Instances 21 10.5 %
Kappa statistic 0.8513
Waikato Environment for Kn
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Type here to search
4:25 AM
4/9/2022

Correctly Classified Instances are 179 out of 200 instances. Incorrectly Classified Instances are 21.Accuracy is 89.5% and incorrectly is 10.5%. We can also see the Confusion Matrix.



To visualize the data we can select " Visualize" Option.



So, finally we have got the results according to the Accuracy are given below-

K-Nearest Neighbours: 82%

Decision Tree: 99%

Naïve Bayes: 96.5%

Neural Network: 98%

Random Forest: 98.5%

Support Vector Machine: 89.5%

From the given result we see Decision Tree gives highest accuracy to fit our model.