CSE-5334-001
DATA MINING
ASSIGNMENT-1 REPORT

STUDENT NAMES:
Kunjal Ahir - 1001877263
Devanshi Sheth - 1001959019
Pankaj Gope - 1001990547

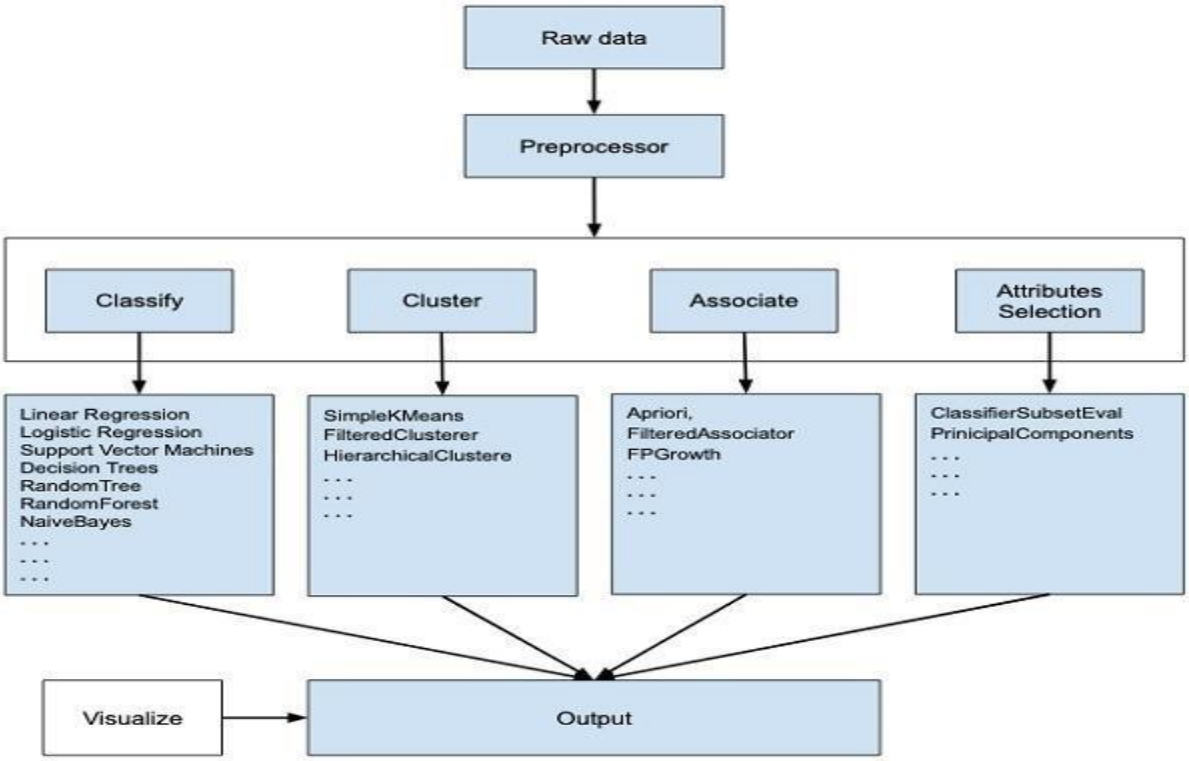
Professor:
Dr Elizabeth D Diaz

Tool:
Weka

Weka

Introduction

WEKA is an open source program that includes data preparation tools, implementation of numerous Machine Learning algorithms, and visualization tools to help you build machine learning approaches and apply them to real-world data mining situations.



Glimpse of dataset

In this assignment we are working on housing dataset.

Attributes:

- price
- lotsize
- bedrooms
- bathrms
- stories
- driveway
- recroom
- fullbase
- gashw
- airco
- garagepl
- prefarea

The following is a detailed definition of the attribute data type:

@attribute " numeric
@attribute price numeric
@attribute lotsize numeric
@attribute bedrooms numeric
@attribute bathrms numeric
@attribute stories numeric
@attribute driveway {yes,no}
@attribute recroom {no,yes}
@attribute fullbase {yes,no}
@attribute gashw {no,yes}
@attribute airco {no,yes}
@attribute garagepl numeric
@attribute prefarea {no,yes}

Viewer

Relation: Housing

No.	1: Numeric	2: price Numeric	3: lotsize Numeric	4: bedrooms Numeric	5: bathrms Numeric	6: stories Numeric	7: driveway Nominal	8: recroom Nominal	9: fullbase Nominal	10: gashw Nominal	11: airco Nominal	12: garagepl Numeric	13: **prefarea** Nominal
1	1.0	42000.0	5850.0	3.0	1.0	2.0	yes	no	yes	no	no	1.0	no
2	2.0	38500.0	4000.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no
3	3.0	49500.0	3060.0	3.0	1.0	1.0	yes	no	no	no	no	0.0	no
4	4.0	60500.0	6650.0	3.0	1.0	2.0	yes	yes	no	no	no	0.0	no
5	5.0	61000.0	6360.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no
6	6.0	66000.0	4160.0	3.0	1.0	1.0	yes	yes	yes	no	yes	0.0	no
7	7.0	66000.0	3880.0	3.0	2.0	2.0	yes	no	yes	no	no	2.0	no
8	8.0	69000.0	4160.0	3.0	1.0	3.0	yes	no	no	no	no	0.0	no
9	9.0	83800.0	4800.0	3.0	1.0	1.0	yes	yes	yes	no	no	0.0	no
10	10.0	88500.0	5500.0	3.0	2.0	4.0	yes	yes	no	no	yes	1.0	no
11	11.0	90000.0	7200.0	3.0	2.0	1.0	yes	no	yes	no	yes	3.0	no
12	12.0	30500.0	3000.0	2.0	1.0	1.0	no	no	no	no	no	0.0	no
13	13.0	27000.0	1700.0	3.0	1.0	2.0	yes	no	no	no	no	0.0	no
14	14.0	36000.0	2880.0	3.0	1.0	1.0	no	no	no	no	no	0.0	no
15	15.0	37000.0	3600.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no
16	16.0	37900.0	3185.0	2.0	1.0	1.0	yes	no	no	no	yes	0.0	no
17	17.0	40500.0	3300.0	3.0	1.0	2.0	no	no	no	no	no	1.0	no
18	18.0	40750.0	5200.0	4.0	1.0	3.0	yes	no	no	no	no	0.0	no
19	19.0	45000.0	3450.0	1.0	1.0	1.0	yes	no	no	no	no	0.0	no
20	20.0	45000.0	3986.0	2.0	2.0	1.0	no	yes	yes	no	no	1.0	no
21	21.0	48500.0	4785.0	3.0	1.0	2.0	yes	yes	yes	no	yes	1.0	no
22	22.0	65900.0	4510.0	4.0	2.0	2.0	yes	no	yes	no	no	0.0	no
23	23.0	37900.0	4000.0	3.0	1.0	2.0	yes	no	no	no	yes	0.0	no
24	24.0	38000.0	3934.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no

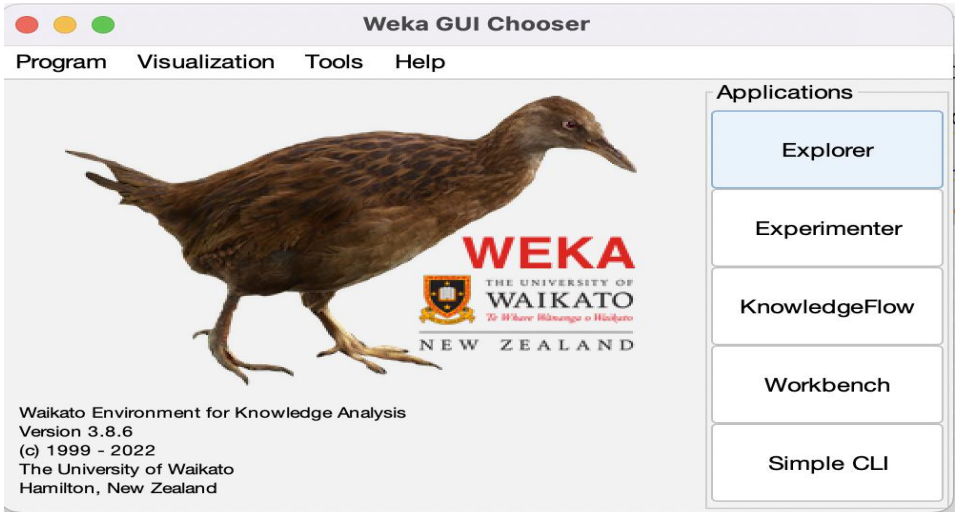
Add instance | Undo | OK | Cancel

Statistical Exploratory Data Analysis

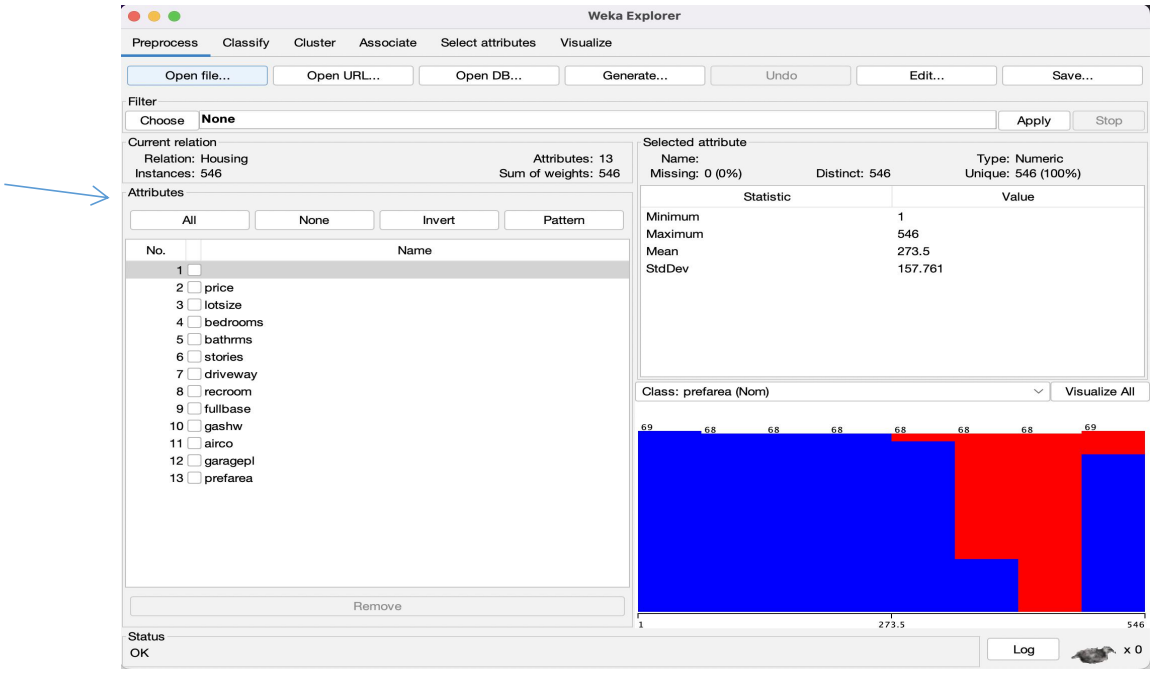
Question 1: How to load the dataset in Weka?

1. Open the Weka GUI Chooser
2. Click the “Explorer” button to open the Weka Explorer.
3. Click the “Open file…” button, navigate to the numeric/ directory and select housing.arff.
4. 4. Click the “Open button”.

The dataset is now loaded into Weka.



Question 2: Find the number of instances and attributes are present in the given dataset.



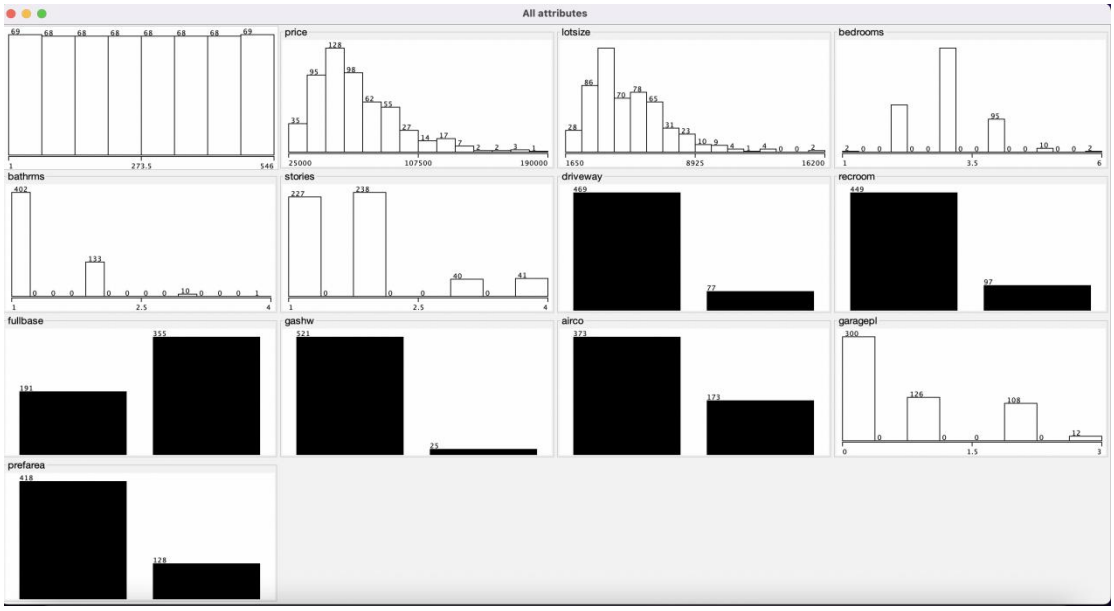
We can see that there are total 546 instances and 13 attributes. The names of attributes are listed in the attributes section above. The first four attributes are of numeric type while the class is a nominal type with 3 distinct values.

Analyse the dataset

Attribute distributions:

Question 3: How will you visualize all the attributes? Give a brief description of the data.

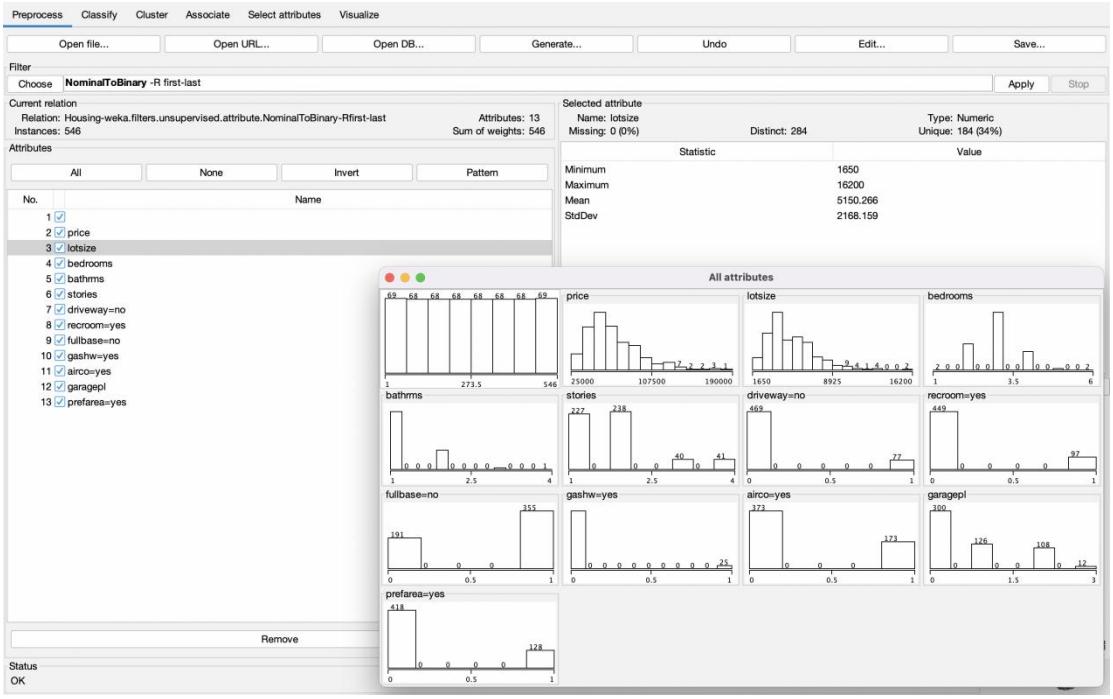
Click Preprocesses and then click on “Visualize All” button from the selected attributes, you’ll see the graphical distribution of each attribute.
Below is the Graphical distribution of each attribute:-



We can notice a few facts about our data:
1. There are no missing values for any of the attributes.
2. 7 inputs are numeric and 6 are binary attribute, and have values in differing ranges.
3. The price, lotsize, bedrooms, bathrms, stories and garagepl attributes are numeric.
4. The drivway recrom, fullbase, gashw, arico, prearea attributes look like a binary-like distribution (two values).

Question 4: How to convert binary data to numeric data? Give a brief description.

1. Click on Choose filter on Preprocess tab.
2. Then click on unsupervised filter -> attributes -> "nominal to binary".
This will visualize the binary data into the numeric form.

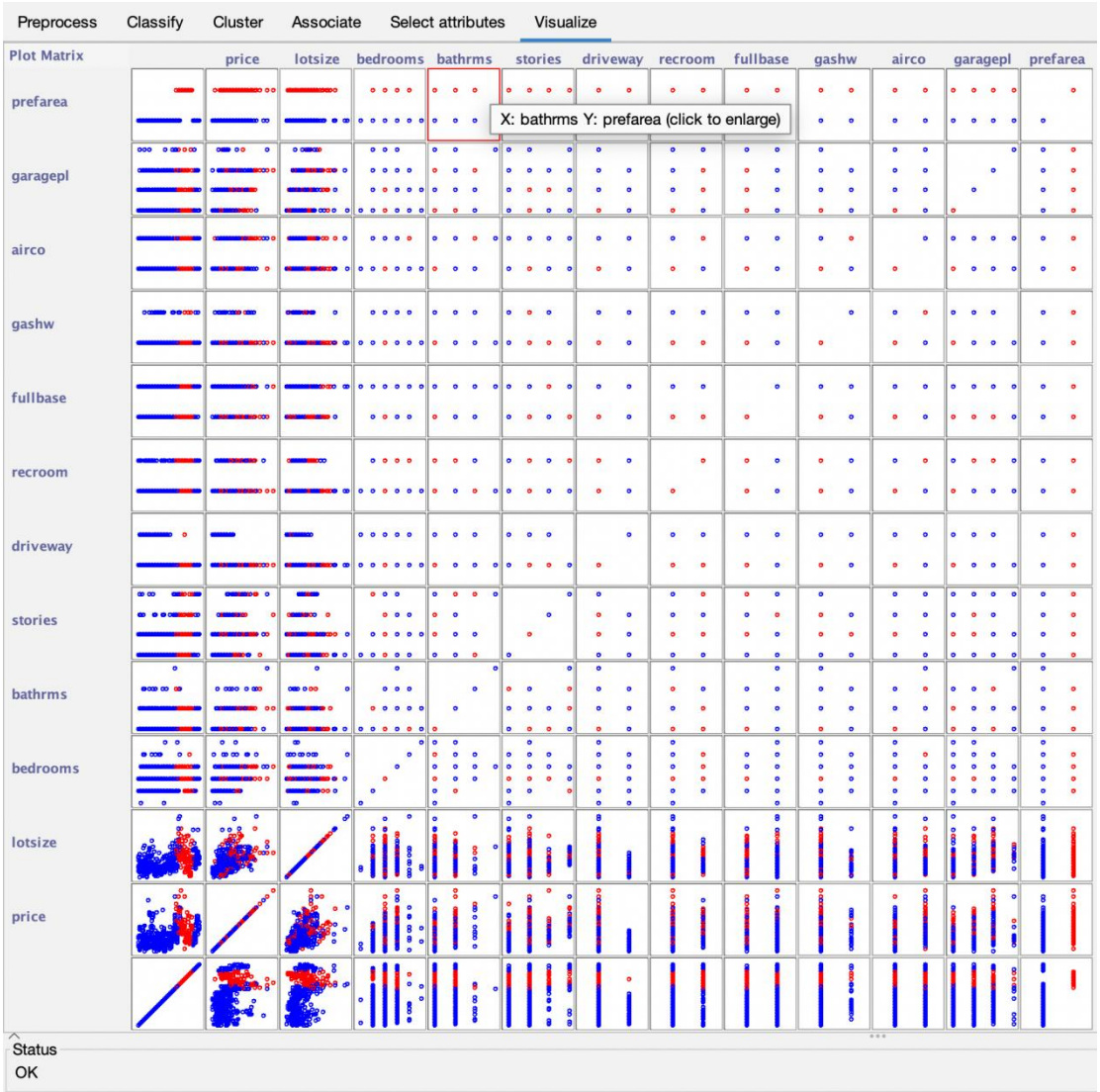


We can see that all the binary data is now represented in theform of numeric values.

Question 5: How will you visualize the interaction between attributes? Give a brief description of the data.

1. Click the "Visualize" option to see how the properties interact with one another.
2. Reduce the "PlotSize" to 50 and modify the window size to ensure that all plots are displayed.
3. Increase "PointSize" to 3 to make the dots more visible.

To apply the changes, click the "Update" button.

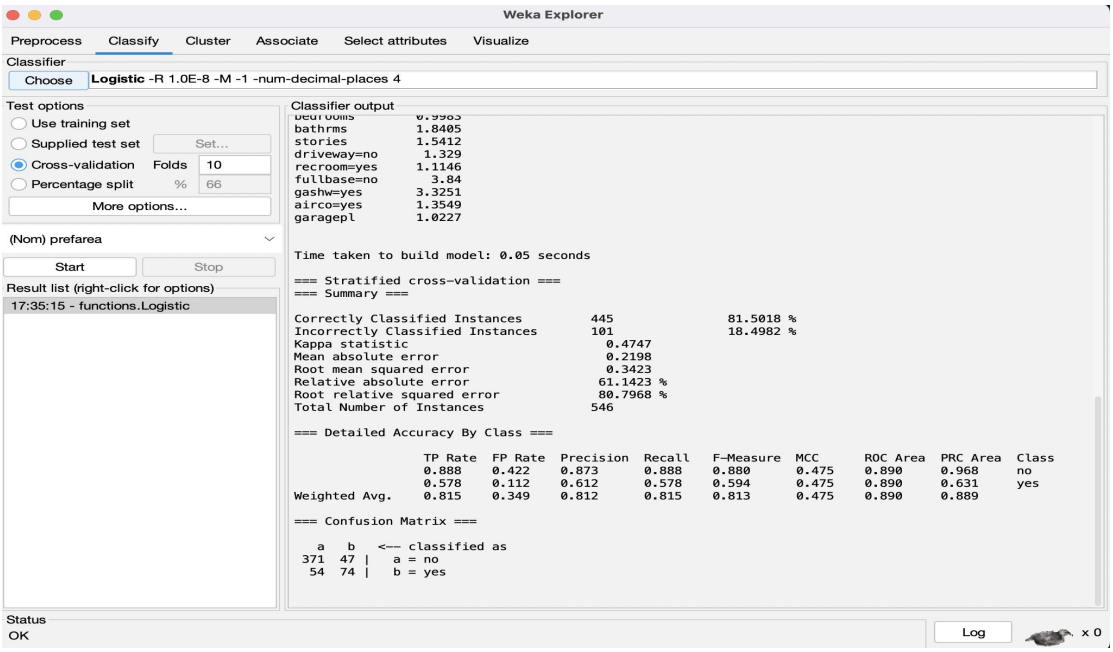


Looking across the graphs we can see some structured relationships that may aid in modeling such as PRICE vs LOTSIZE and PRICE vs BEDROOMS.

Classifiers

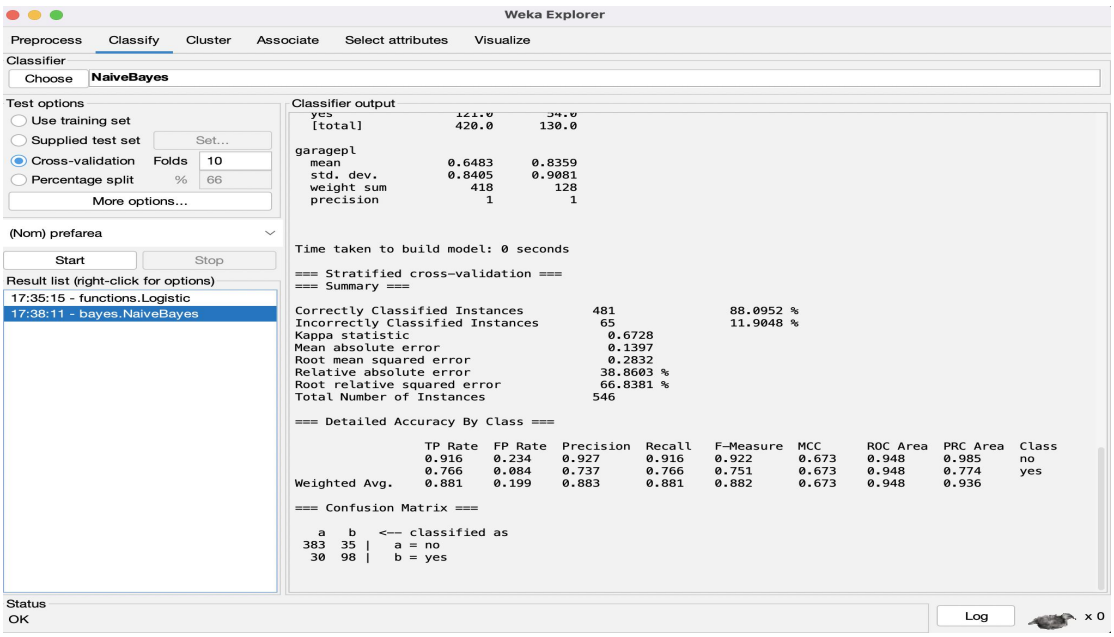
Question 6: Implement top five classifiers and state the accuracy of each.

1. Logistic:



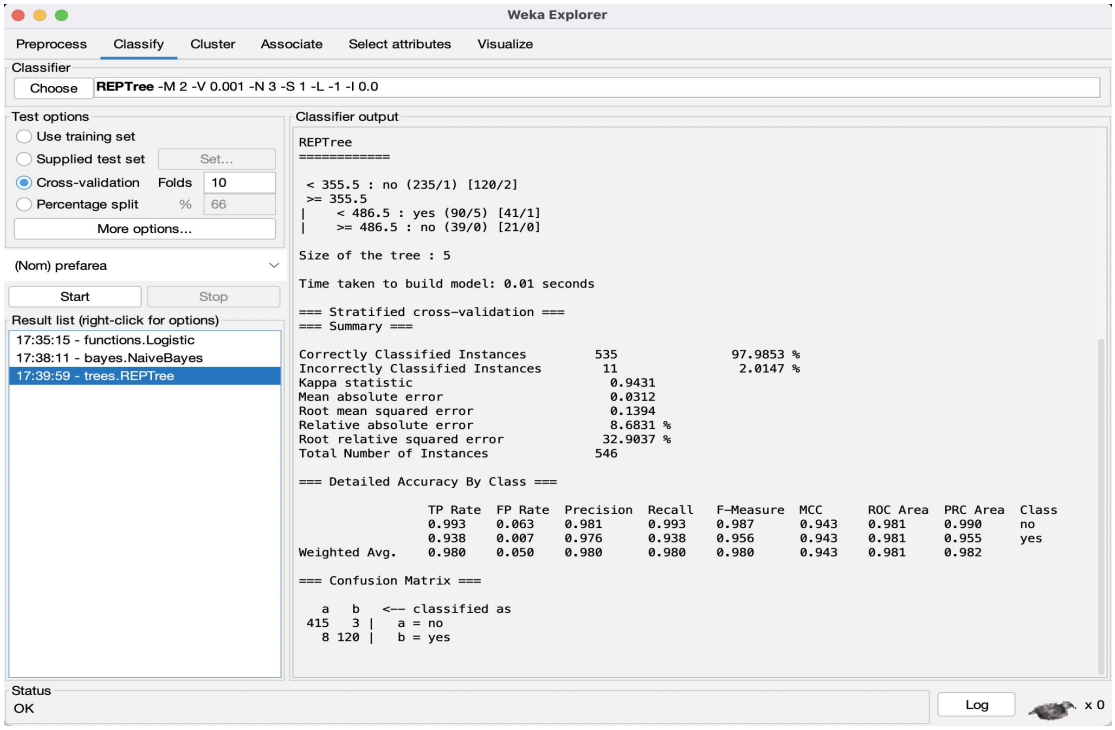
You can see that with the default configuration that the Logistic algorithm achieves an accuracy of 81.50%

2. Naive Bayes



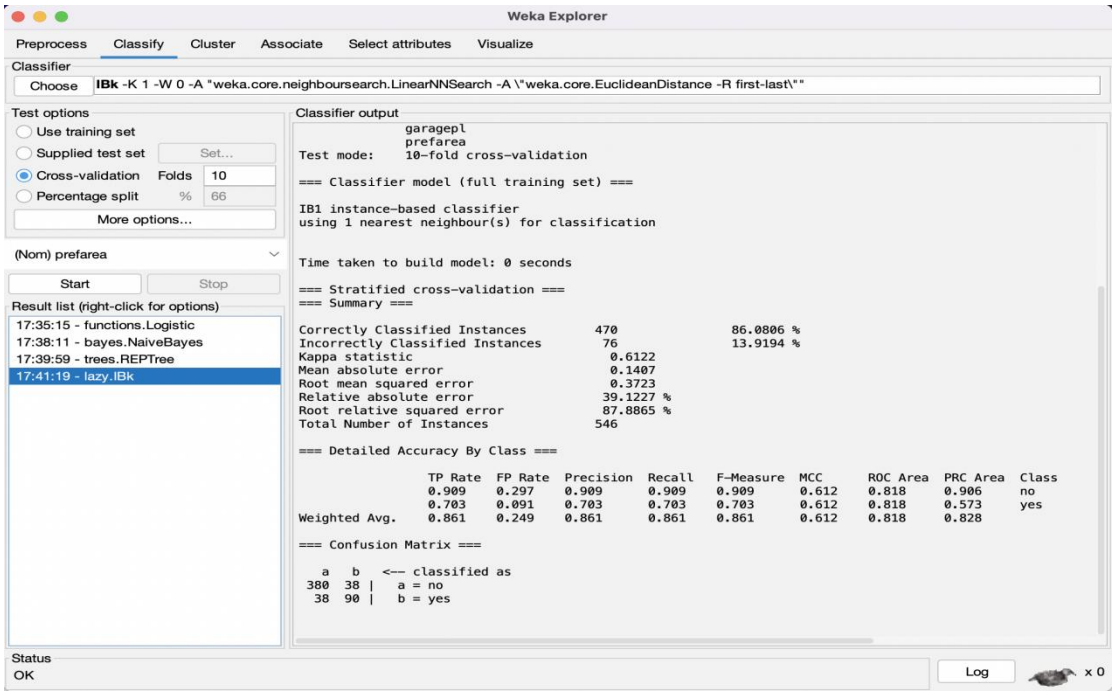
You can see that with the default configuration that the Naive Bayes algorithm achieves an accuracy of 88.09%

3. Decision Tree



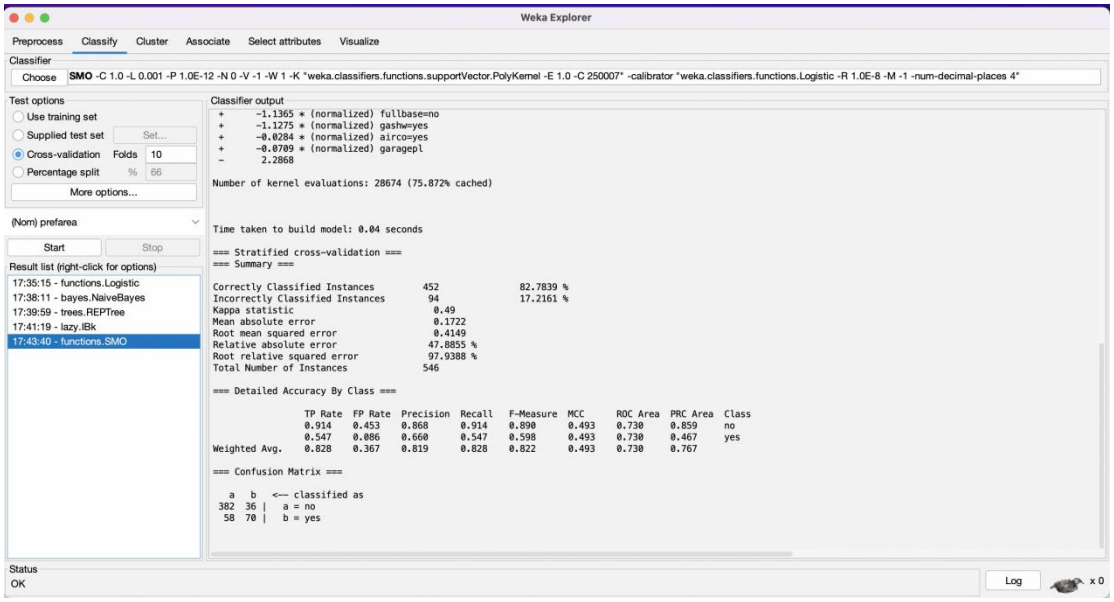
You can see that with the default configuration that the Decision Tree algorithm achieves an accuracy of 97.98%

4. k-Nearest Neighbors



You can see that with the default configuration that the k-nearest neighbors algorithm achieves an accuracy of 86.08%

5. Support Vector machine



You can see that with the default configuration that the SVM algorithm achieves an accuracy of 82.78%

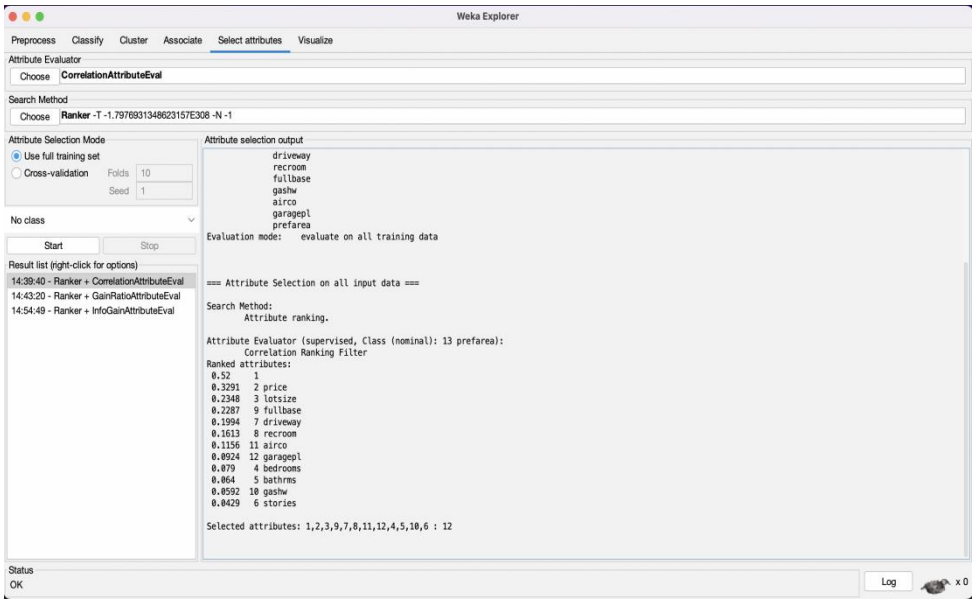
As a result, we may infer that Decision Tree methods provide the highest accuracy of the five algorithms.

Selection Attributes

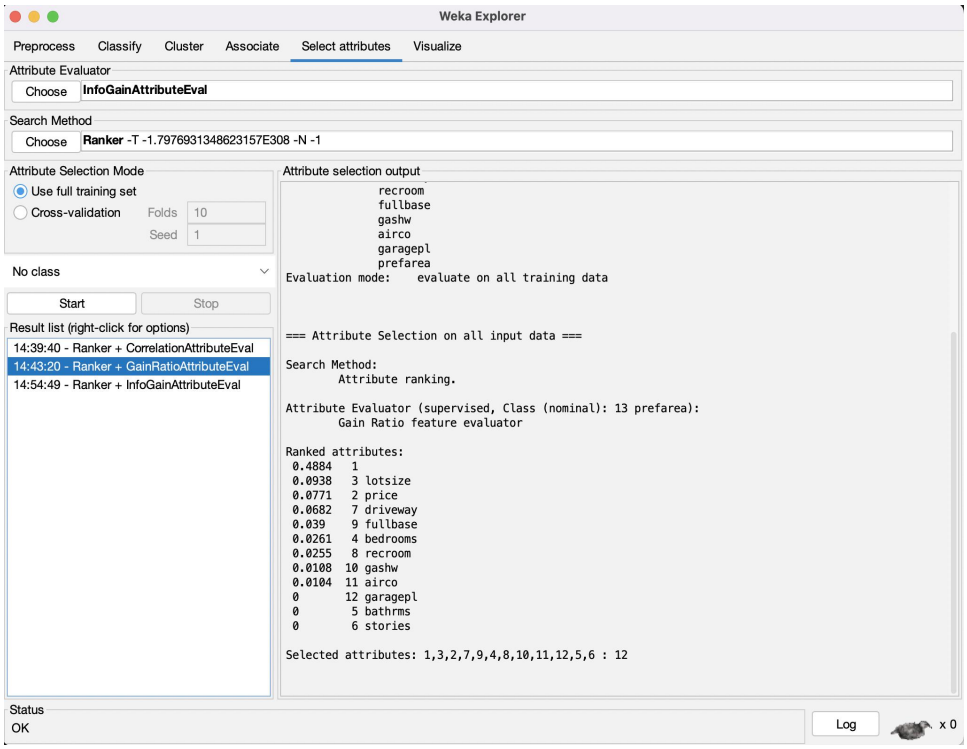
Attribute Selection (ranking)

Question 7: Evaluate any three atributes and find the difference in between them.

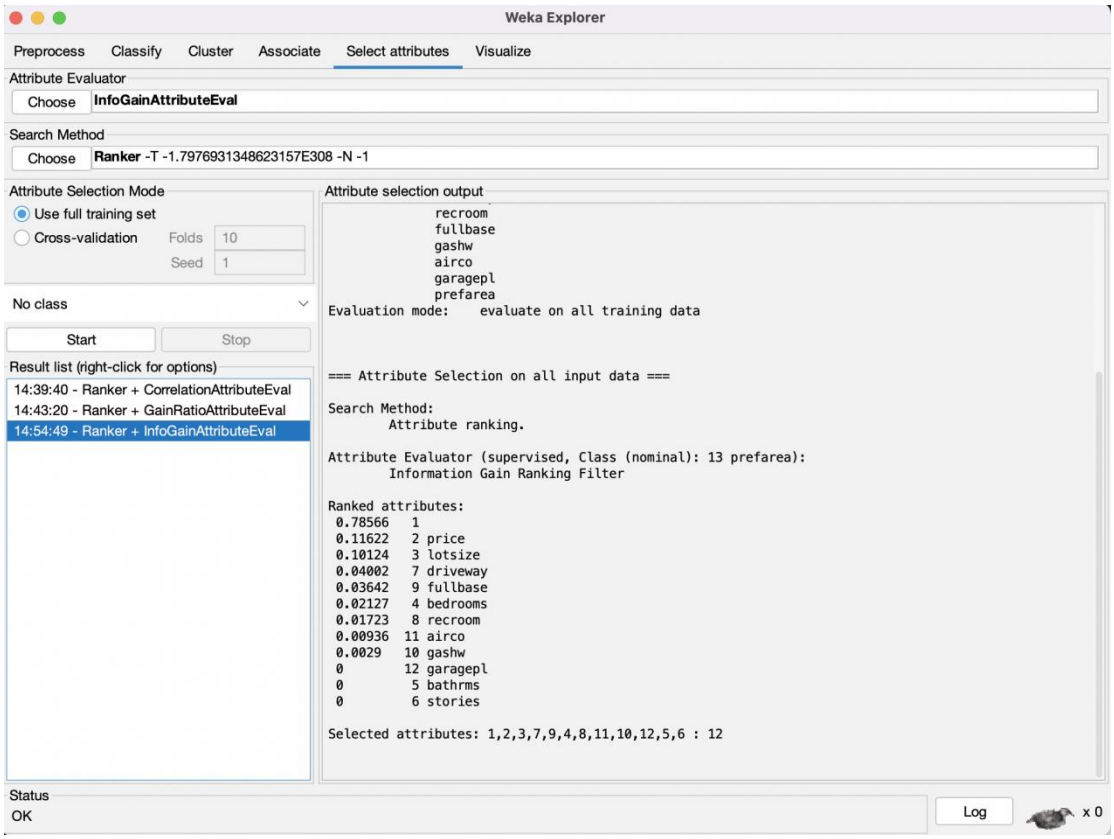
Correlation:



Gain Ratio:



InfoGain:



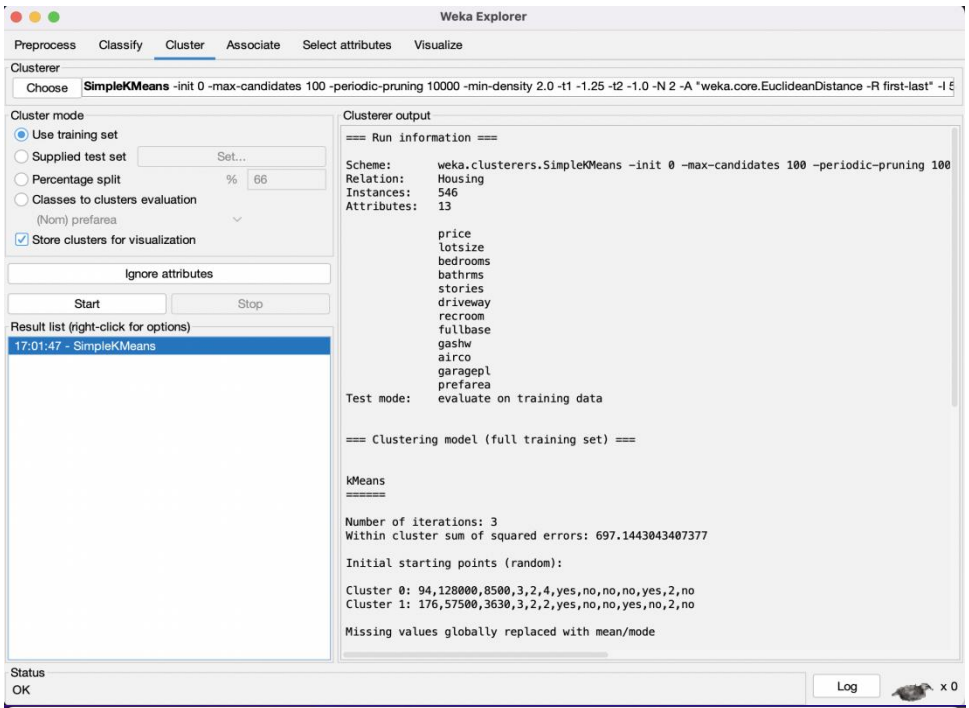
The sequence of ranking in all three selection characteristics differs.

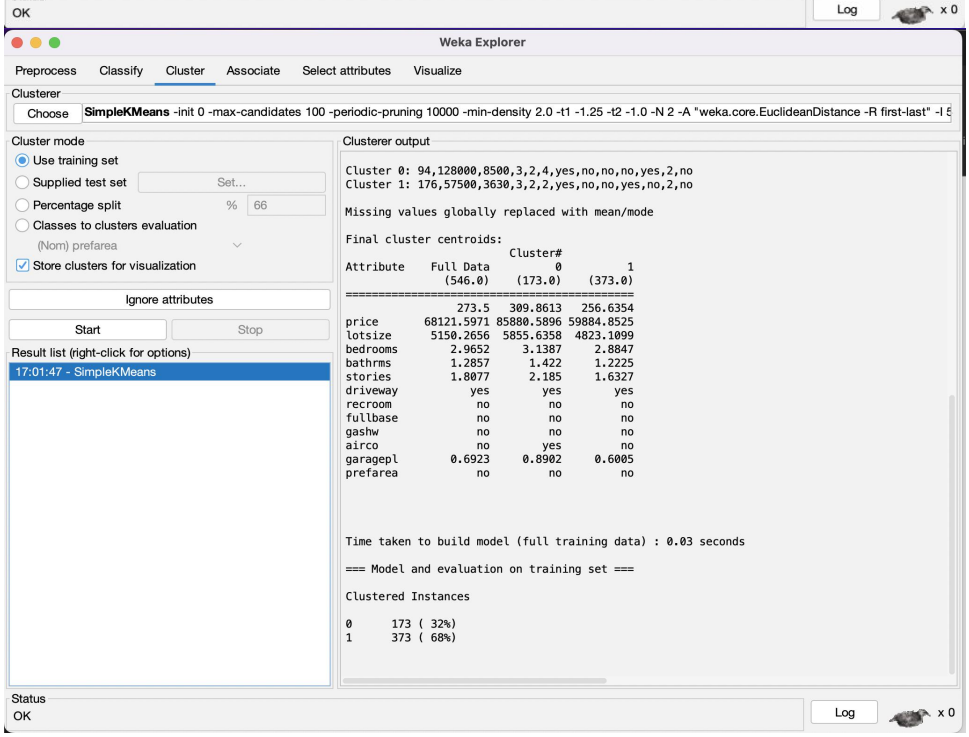
Clustering

Question 8: Describe any two of the clusturing methods and give the difference between them.

Simple KMeans clustering

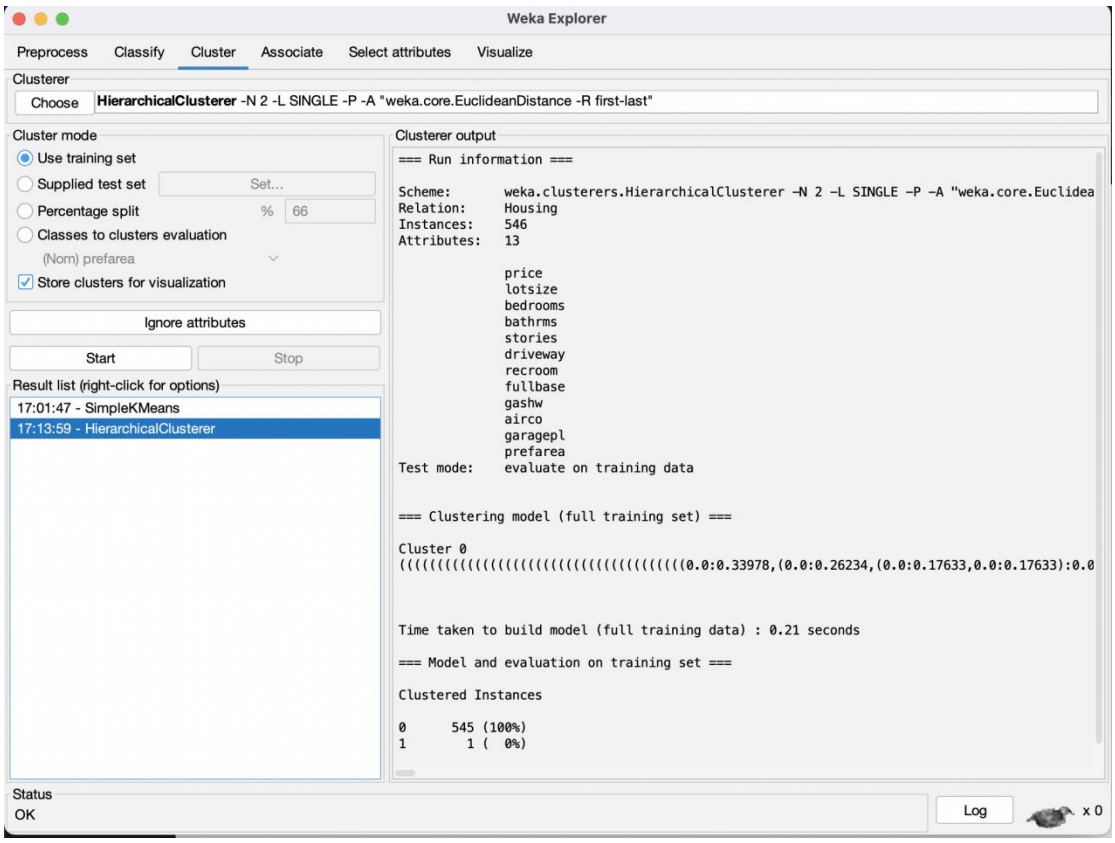
The result window displays the centroid of each cluster, as well as data on the number and percentage of instances assigned to each cluster. A mean vector represents each cluster centroid. A cluster may be described using this cluster.





Hierarchical Clustering:

The resultant window shows the centroid of each cluster as well as information on the number and proportion of instances assigned to each cluster. Each cluster centroid is represented by a mean vector. This cluster can be used to describe a cluster.



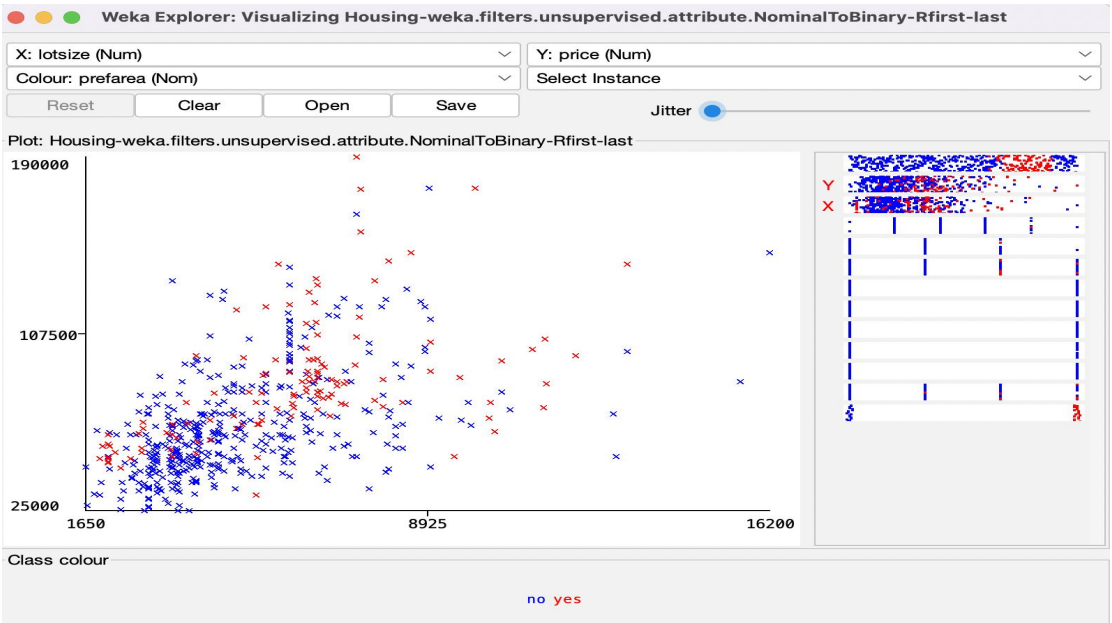
Difference:

1. The performance of K- mean algorithm is better than Hierarchical Clustering Algorithm.
2. The Time Take to build model (full training data) is faster and efficient in Kmeans clustering.
3. The Heirachical clustering is noisy than Kmeans Clustering.

Visualization of Scatter Plot

Question 9: Visualize the scatterplot of Lotsize vs price and Bedrooms vs Price.

lotsize vs price:



Bedrooms vs Price:

Conclusion

WEKA is a strong tool for creating machine learning models. It implements some of the most popular ML methods. It also allows you to preprocess your data before applying these algorithms to it. The supporting algorithms are classed as Classify, Cluster, Associate, and Select characteristics. A beautiful and strong visual representation may be used to visualize the outcome at various stages of processing. This allows a Data Scientist to easily apply several machine learning approaches to his dataset, evaluate the results, and develop the optimal model for ultimate application.

References

https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/
https://www.geeksforgeeks.org/k-means-clustering-using-weka/
https://cs.ccsu.edu/~markov/weka-tutorial.pdf
https://www.tutorialspoint.com/weka/what_is_weka.htm
https://www.tutorialspoint.com/weka/weka_clustering.htm
http://people.sabanciuniv.edu/berrin/cs512/hws/hw1/WEKA%20Explorer%20Tutorial-REFERENCE.pdf