NAME: JOYADDAR MD JOBAYER

MATRIC NO:1731833

EXERCISE CLUSTERING

Answer of number 1

Clustering is a data mining technique for grouping data components into similar categories. That is the process of grouping data (or objects) into the same class. Those in one cluster are more similar to each other than data in another. Clustering is the process of dividing data items into subclasses. A cluster is made up of data objects that have a high level of inter similarity but a low level of intra similarity. Clustering has the advantage of being adaptive to changes and assisting in the identification of relevant qualities that separate various groupings.

Answer of number 2

Clustering is a technique for evaluating data that does not include pre-labeled groups. The notion of maximizing inters - class similarity while decreasing similarity across classes is used to group data instances together. This means that the clustering method will locate and group examples that are quite similar to one another, as opposed to ungrouped instances that are very dissimilar. Clustering is a type of unsupervised learning since it does not involve the pre-labeling of classes.

Answer of number 3

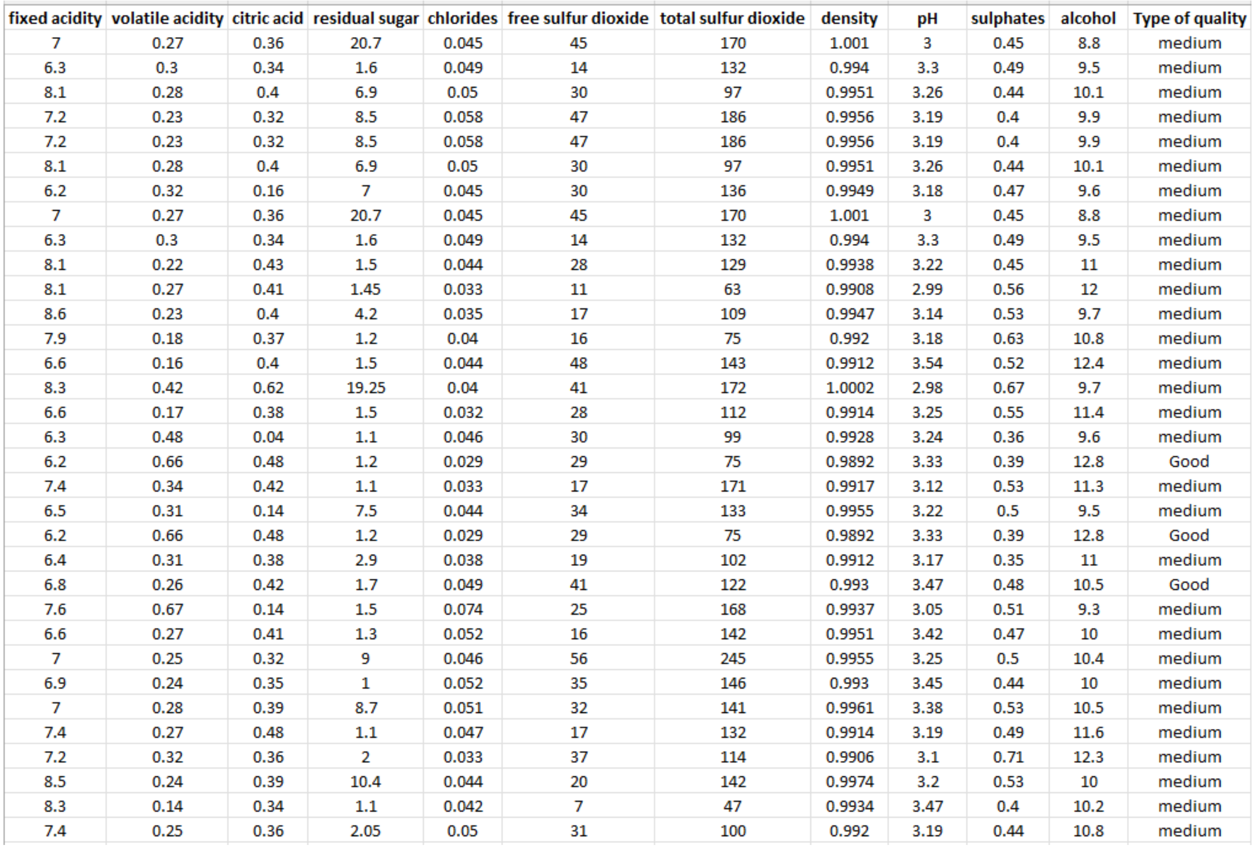
1. Cluster analysis is frequently utilized in market research, pattern identification, data analysis, and image processing, among other applications. Clustering can assist marketers identify unique groups in their consumer bases and describe them based on purchase behaviors in the business world.

2. Clustering can also aid in the identification of similar land use areas in an earth observation database, as well as the identification of groups of houses in a city based on house type, value, and geographic location, as well as the identification of groups of auto insurance policyholders with a high average claim **cost.**

3. In other applications, clustering is referred to as data segmentation since it divides big data sets into categories based on their similarities. Outlier identification may also be done via clustering, with outliers (values that are "far away" from any cluster) potentially being more interesting than usual examples. Outlier detection is used in the identification of credit card fraud and the monitoring of illegal activity in the electronic commerce industry.

Answer of number 4:

After download the dataset we have transformed the file to CSV, and we have created a new attribute from the quality attributes which is type of quality. The new attribute contains three types of data which are low (1-3), medium (4-7) and good (8-10). We have attached a screenshot as a reference bellow:



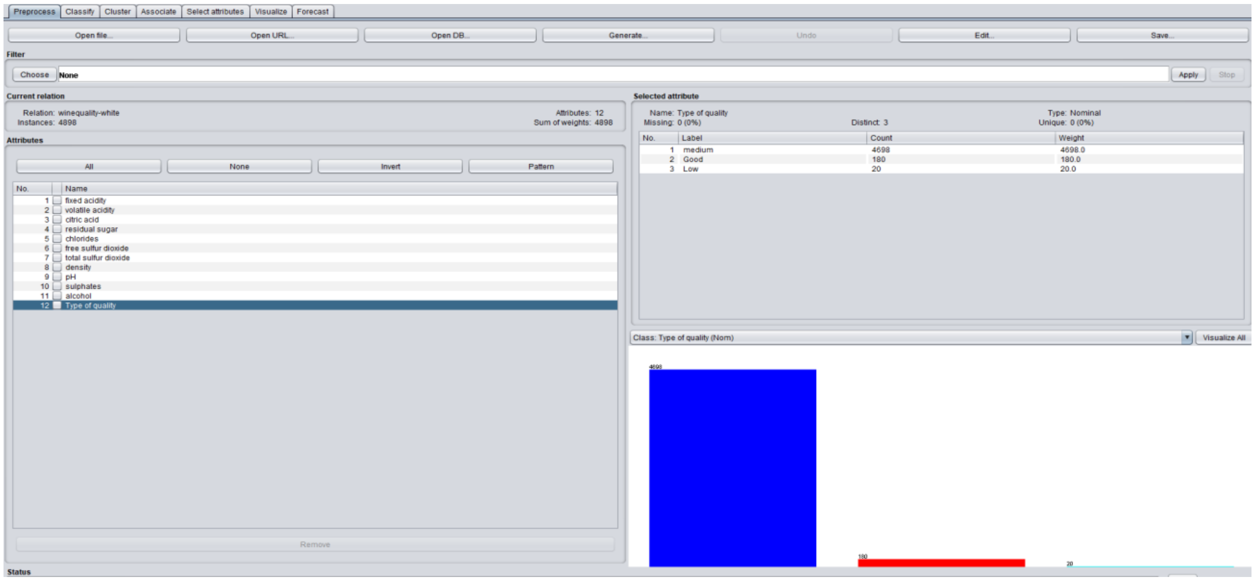
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	Type of quality
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	medium
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	medium
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	medium
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	medium
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	medium
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	medium
6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	medium
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	medium
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	medium
8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	medium
8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	medium
8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	medium
7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	medium
6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	medium
8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	medium
6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	medium
6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	medium
6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	Good
7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	medium
6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	medium
6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	Good
6.4	0.31	0.38	2.9	0.038	19	102	0.9912	3.17	0.35	11	medium
6.8	0.26	0.42	1.7	0.049	41	122	0.993	3.47	0.48	10.5	Good
7.6	0.67	0.14	1.5	0.074	25	168	0.9937	3.05	0.51	9.3	medium
6.6	0.27	0.41	1.3	0.052	16	142	0.9951	3.42	0.47	10	medium
7	0.25	0.32	9	0.046	56	245	0.9955	3.25	0.5	10.4	medium
6.9	0.24	0.35	1	0.052	35	146	0.993	3.45	0.44	10	medium
7	0.28	0.39	8.7	0.051	32	141	0.9961	3.38	0.53	10.5	medium
7.4	0.27	0.48	1.1	0.047	17	132	0.9914	3.19	0.49	11.6	medium
7.2	0.32	0.36	2	0.033	37	114	0.9906	3.1	0.71	12.3	medium
8.5	0.24	0.39	10.4	0.044	20	142	0.9974	3.2	0.53	10	medium
8.3	0.14	0.34	1.1	0.042	7	47	0.9934	3.47	0.4	10.2	medium
7.4	0.25	0.36	2.05	0.05	31	100	0.992	3.19	0.44	10.8	medium

Figure 1: After pre-process of white wine dataset

We have used Weka for providing a simple explanation about how good or bad the dataset is. In Weka, we have used three types of clustering algorithm to justify the quality of white wines. Clustering algorithms are as follows:

- Simple k-Means Cluster
- Density-Based Cluster
- Hierarchical Cluster

- This is the interface of Weka after opening the white wine dataset.



Filter: None

Current relation

Relation: winequality-white
Instances: 4898
Attributes: 12
Sum of weights: 4898

Attributes

No.		Name
1	□	fixed acidity
2	□	volatile acidity
3	□	citric acid
4	□	residual sugar
5	□	chlorides
6	□	free sulfur dioxide
7	□	total sulfur dioxide
8	□	density
9	□	pH
10	□	sulphates
11	□	alcohol
12	□	Type of quality

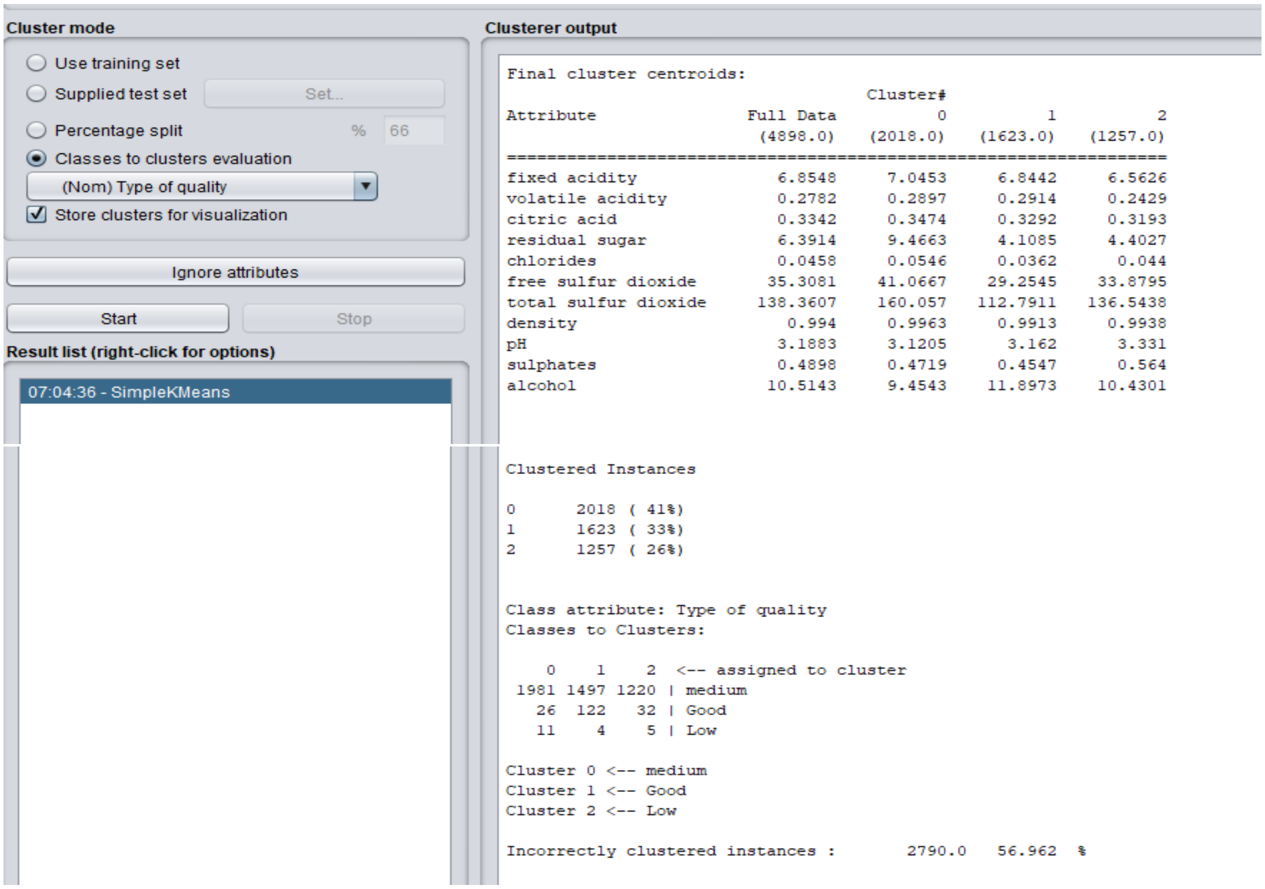
Selected attribute

Name: Type of quality
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	medium	4698	4698.0
2	Good	180	180.0
3	Low	20	20.0

Class: Type of quality (Nom)

- **Simple k-Means Cluster**



Cluster mode

○ Use training set
○ Supplied test set [Set...]
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) Type of quality
☑ Store clusters for visualization

Result list (right-click for options)

07:04:36 - SimpleKMeans

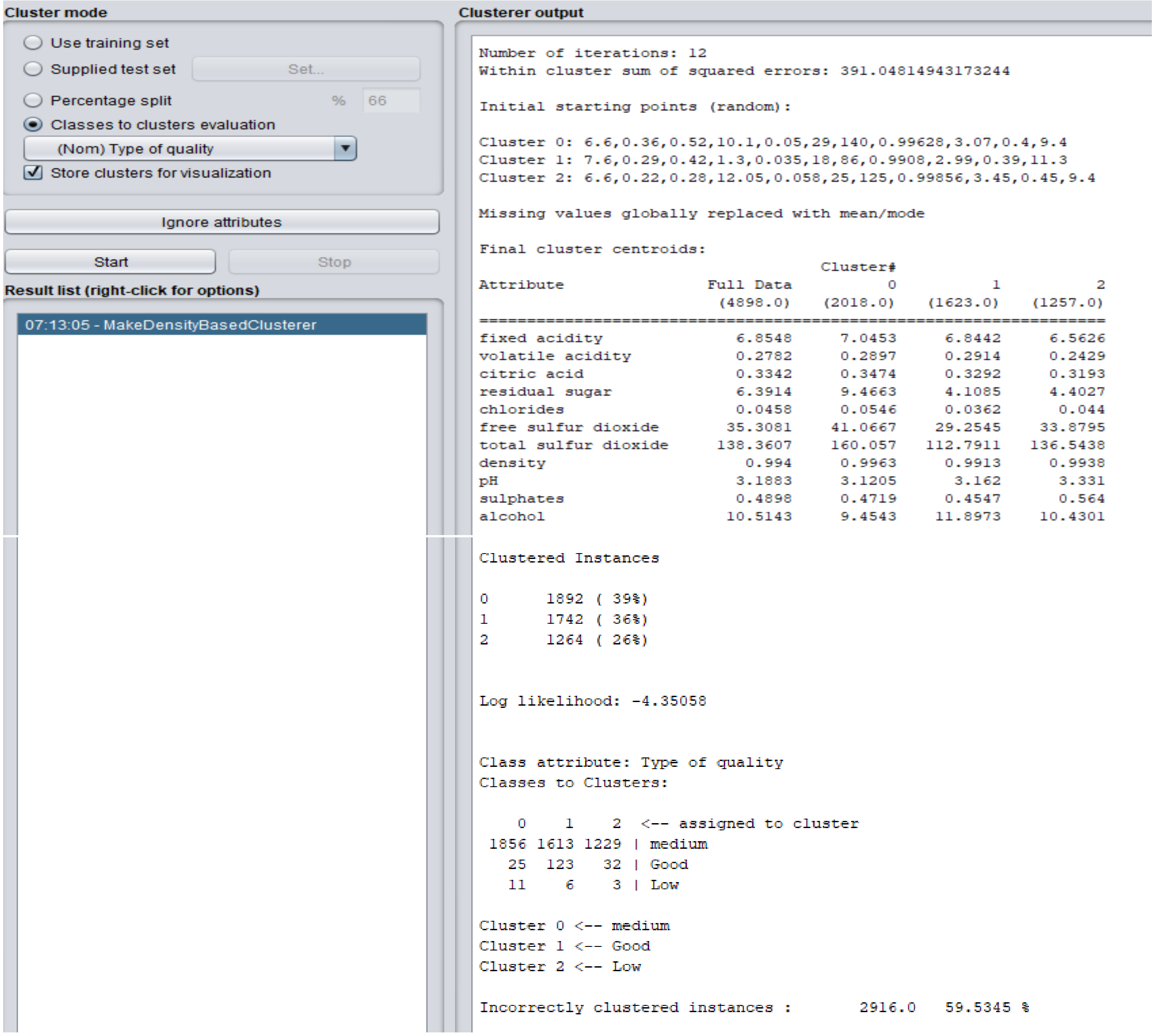
Clusterer output

```
Final cluster centroids:
                                          Cluster#
Attribute                 Full Data          0            1            2
                           (4898.0)     (2018.0)     (1623.0)     (1257.0)
==================================================================
fixed acidity                6.8548       7.0453       6.8442       6.5626
volatile acidity             0.2782       0.2897       0.2914       0.2429
citric acid                  0.3342       0.3474       0.3292       0.3193
residual sugar               6.3914       9.4663       4.1085       4.4027
chlorides                    0.0458       0.0546       0.0362        0.044
free sulfur dioxide         35.3081      41.0667      29.2545      33.8795
total sulfur dioxide       138.3607      160.057     112.7911     136.5438
density                       0.994       0.9963       0.9913       0.9938
pH                           3.1883       3.1205        3.162        3.331
sulphates                    0.4898       0.4719       0.4547        0.564
alcohol                     10.5143       9.4543      11.8973      10.4301



Clustered Instances

0      2018 ( 41%)
1      1623 ( 33%)
2      1257 ( 26%)


Class attribute: Type of quality
Classes to Clusters:

    0    1    2  <-- assigned to cluster
 1981 1497 1220 | medium
   26  122   32 | Good
   11    4    5 | Low

Cluster 0 <-- medium
Cluster 1 <-- Good
Cluster 2 <-- Low

Incorrectly clustered instances :       2790.0   56.962  %
```

Figure 1: Weka Explorer- classification using K-means cluster

For k-means has three cluster which are cluster 0 stand for medium, cluster 1 stand for good and cluster 2 stand for low. As we can see in figure 1, incorrectly clustered instances 2790 or 56.962% From the mean distance of the centroid, for fixed acidity of cluster 0(medium) value is 7.0453, cluster 1 (good) value is 6.8442 and cluster 2(low) value is 6. 5626.So, we can conclude that using k-means cluster algorithm incorrectly instances is 56.962%

- **Density-Based Cluster**



Cluster mode

○ Use training set
○ Supplied test set
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) Type of quality
☑ Store clusters for visualization

Result list (right-click for options)

07:13:05 - MakeDensityBasedClusterer

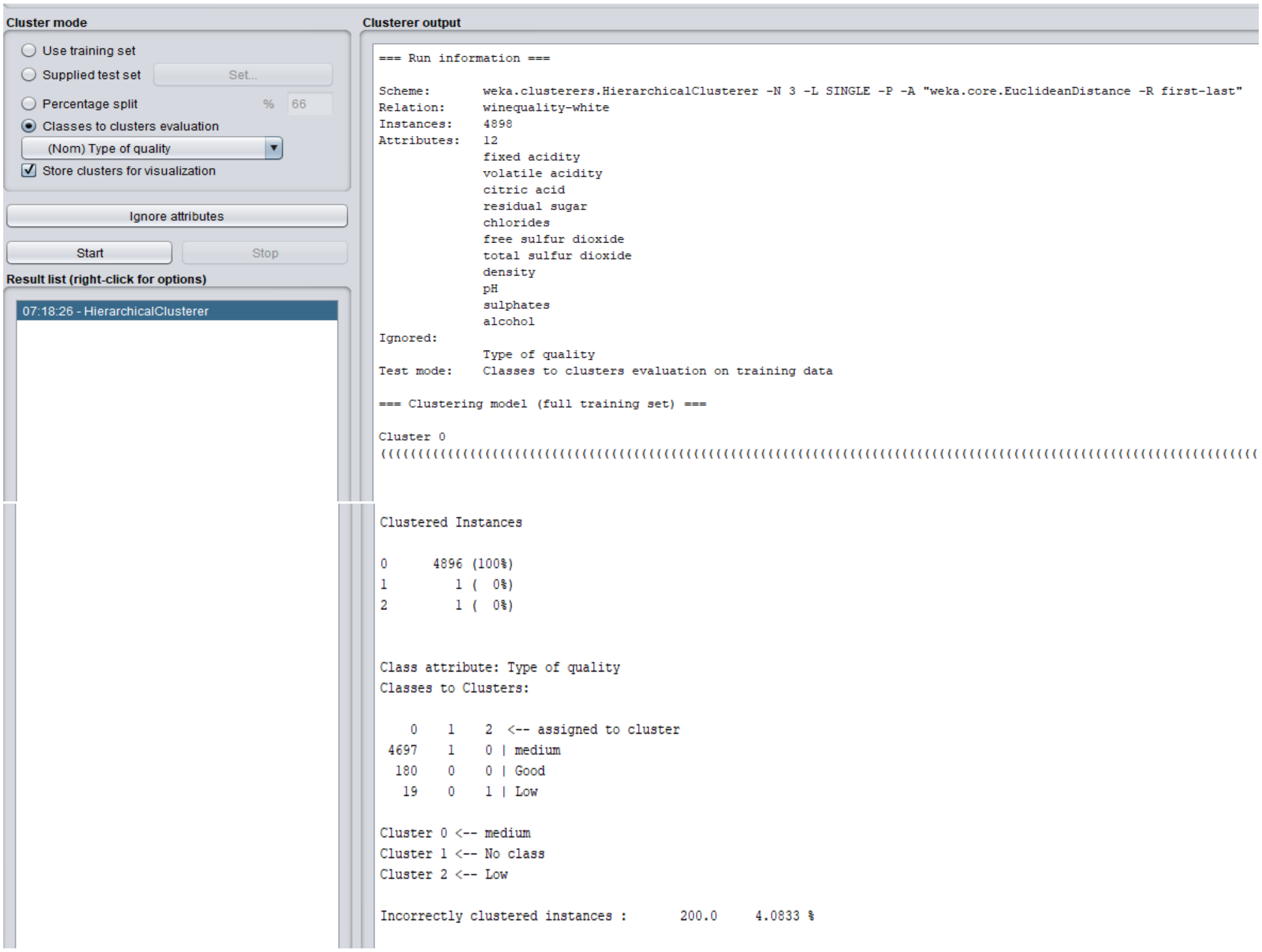
Clusterer output

```
Number of iterations: 12
Within cluster sum of squared errors: 391.04814943173244

Initial starting points (random):

Cluster 0: 6.6,0.36,0.52,10.1,0.05,29,140,0.99628,3.07,0.4,9.4
Cluster 1: 7.6,0.29,0.42,1.3,0.035,18,86,0.9908,2.99,0.39,11.3
Cluster 2: 6.6,0.22,0.28,12.05,0.058,25,125,0.99856,3.45,0.45,9.4

Missing values globally replaced with mean/mode

Final cluster centroids:
                                      Cluster#
Attribute               Full Data            0            1            2
                         (4898.0)     (2018.0)     (1623.0)     (1257.0)
==========================================================================
fixed acidity              6.8548       7.0453       6.8442       6.5626
volatile acidity           0.2782       0.2897       0.2914       0.2429
citric acid                0.3342       0.3474       0.3292       0.3193
residual sugar             6.3914       9.4663       4.1085       4.4027
chlorides                  0.0458       0.0546       0.0362        0.044
free sulfur dioxide       35.3081      41.0667      29.2545      33.8795
total sulfur dioxide     138.3607      160.057     112.7911     136.5438
density                     0.994       0.9963       0.9913       0.9938
pH                         3.1883       3.1205        3.162        3.331
sulphates                  0.4898       0.4719       0.4547        0.564
alcohol                   10.5143       9.4543      11.8973      10.4301


Clustered Instances

0       1892 ( 39%)
1       1742 ( 36%)
2       1264 ( 26%)


Log likelihood: -4.35058


Class attribute: Type of quality
Classes to Clusters:

    0    1    2  <-- assigned to cluster
 1856 1613 1229 | medium
   25  123   32 | Good
   11    6    3 | Low

Cluster 0 <-- medium
Cluster 1 <-- Good
Cluster 2 <-- Low

Incorrectly clustered instances :       2916.0    59.5345 %
```

Figure 2: Weka Explorer- classification using Density-based Cluster

For density has three cluster which are cluster 0 stand for medium, cluster 1 stand for good and cluster 2 stand for low. As we can see in figure 2, incorrectly clustered instances 2916 or 59.5345% From the mean distance of the centroid, for volatile acidity of cluster 0(medium) value is 0.2897, cluster 1 (good) value is 0.2914 and cluster 2(low) value is 0. 2429.So, we can conclude that using density-based cluster algorithm incorrectly instances is 59.5345%

➢ Hierarchical Cluster



○ Use training set
○ Supplied test set
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) Type of quality
☑ Store clusters for visualization

07:18:26 - HierarchicalClusterer

```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 3 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     winequality-white
Instances:    4898
Attributes:   12
              fixed acidity
              volatile acidity
              citric acid
              residual sugar
              chlorides
              free sulfur dioxide
              total sulfur dioxide
              density
              pH
              sulphates
              alcohol
Ignored:
              Type of quality
Test mode:    Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

Cluster 0
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((

Clustered Instances

0      4896 (100%)
1         1 (  0%)
2         1 (  0%)


Class attribute: Type of quality
Classes to Clusters:

    0    1    2  <-- assigned to cluster
 4697    1    0 | medium
  180    0    0 | Good
   19    0    1 | Low

Cluster 0 <-- medium
Cluster 1 <-- No class
Cluster 2 <-- Low

Incorrectly clustered instances :       200.0     4.0833 %
```

Figure 3: Weka Explorer- classification using hierarchical Cluster

For hierarchical cluster we can see, incorrectly clustered instances only are 200 or 4.0833%

DECISION

After using three types of cluster algorithm (**Simple k-Means Cluster, Density-Based Cluster &** Hierarchical Cluster) we have found incorrect instances as follows:

Simple k-Means Cluster = 2790 or 56.962%

Density-Based Cluster = 2916 or 59.5345%

Hierarchical Cluster = 200 or 4.0833%

Therefore, hierarchical cluster accuracy is good for the white wine dataset. Because this algorithm shows only 200 incorrected cluster instances whereas, density-based cluster is the worst accuracy for the dataset, so it shows 2916 incorrectly cluster instances.

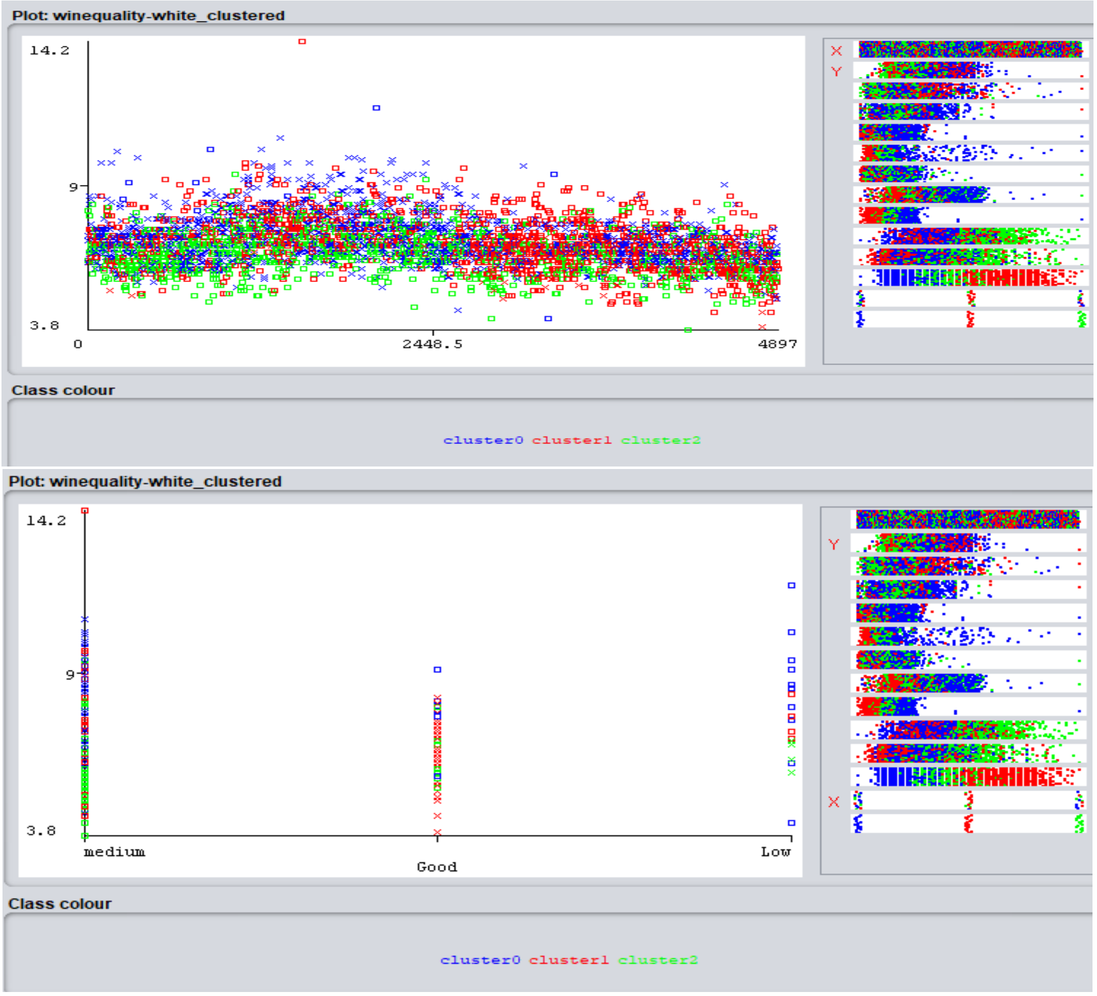

Plot: winequality-white_clustered
14.2
9
3.8
0
2448.5
4897
X
Y
Class colour
cluster0 cluster1 cluster2
Plot: winequality-white_clustered
14.2
9
3.8
medium
Good
Low
Y
X
Class colour
cluster0 cluster1 cluster2

Figure 4: Distribution of Simple k-Means Cluster

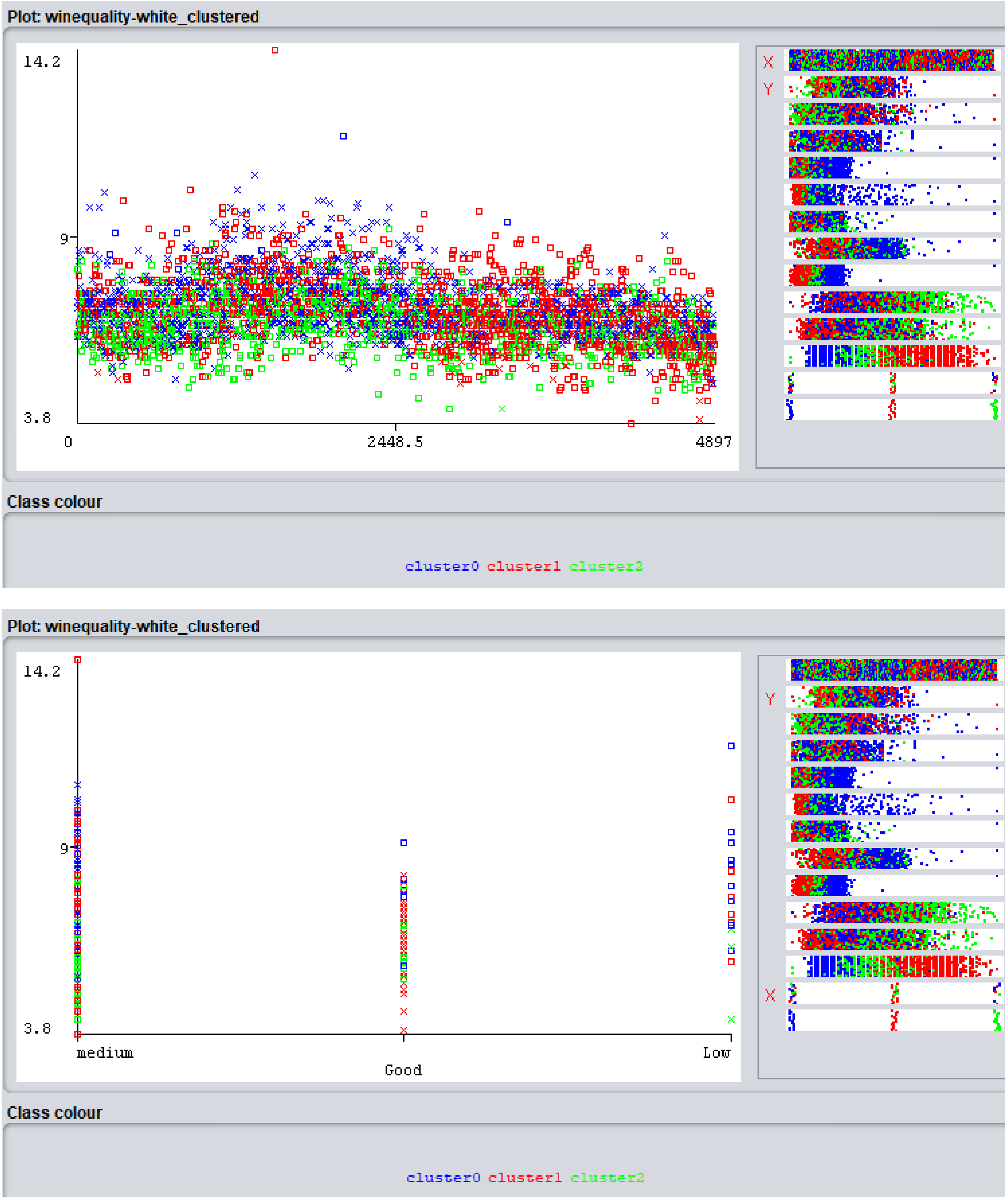
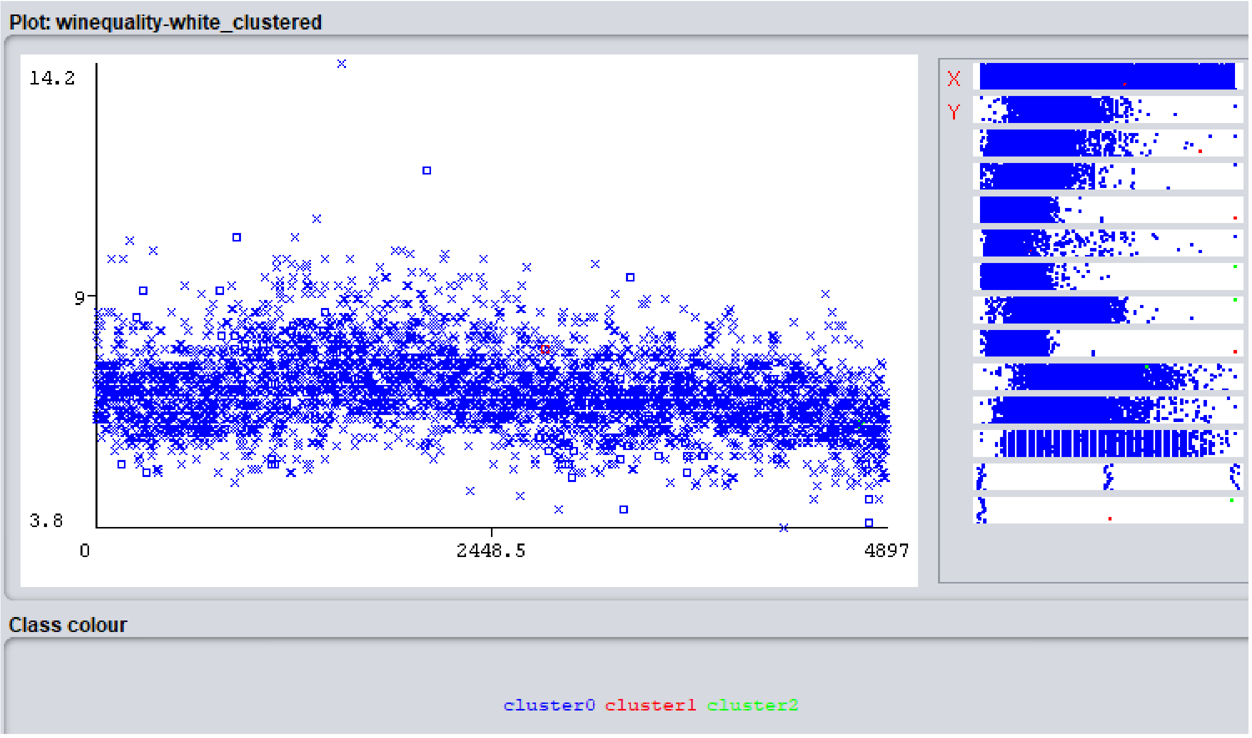

Plot: winequality-white_clustered
14.2
9
3.8
0
2448.5
4897
X
Y
Class colour
cluster0 cluster1 cluster2
Plot: winequality-white_clustered
14.2
9
3.8
medium
Good
Low
Y
X
Class colour
cluster0 cluster1 cluster2

Figure 5: Distribution of Density-Based Cluster



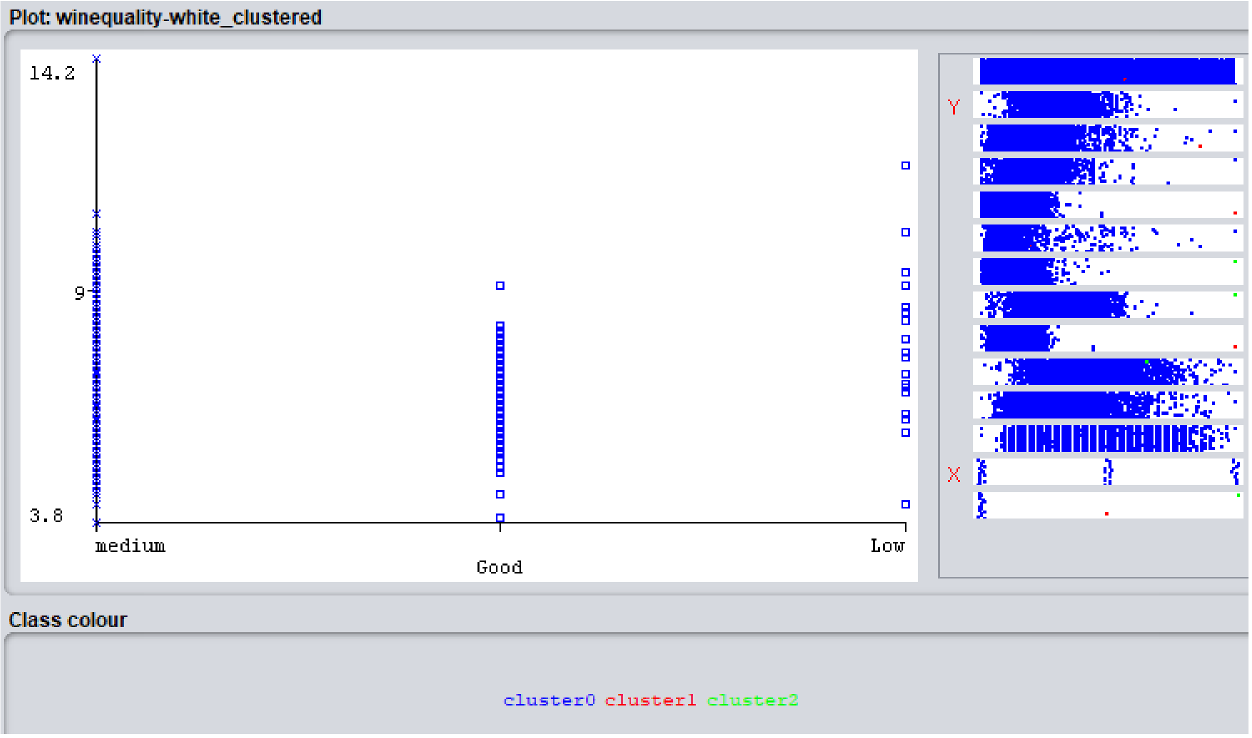


Figure 6: Distribution of Hierarchical Cluster