Data Exploration and Clustering

Module Code: **CS3DS19**
Assignment Report Title: **Data Exploration and Clustering**
Student Number: **28010336**
Date Completed: **23/03/2022**
Actual hours spent for the assignment: **17 hours**
Assignment evaluation (3 key points):
1) I have learnt how to apply a clustering algorithm to a dataset in KNIME.
2) I found the implementation aspect of the assignment quite straight forward.
3) The brief was complicated to understand what was required in the assignment.

Introduction

The aim of the assignment was to perform a clustering analysis, on the provided multidimensional Wine dataset, twice – once without normalisation and once with normalisation. The report contains the documentation of the process that was implemented in KNIME. Scatter Plot graphs produced were then compared and discussed, determining if normalisation affects the results.

Clustering Algorithm

K-Means clustering algorithm was used for plot2 in Task 1 and Task 2. It is an unsupervised learning algorithm which divides the dataset into K clusters. In this case, the dataset has been divided into three clusters. The dataset was shuffled, and three data points were randomly selected as centroids. After the centroids were initialised, data points were allocated to the closest centroid, using Euclidean distance. Next, new centroids were generated, in the centre of the clusters, replacing the original ones. Then, the data points were, again, allocated to the closest centroid. The algorithm iteratively repeats the steps until centroids do not change and that the data points remain in the same cluster.

Task 1: Clustering Without Normalisation

In this task, the clustering algorithm was performed without normalisation. Figure 11 displays the workflow.

1.1: plot1

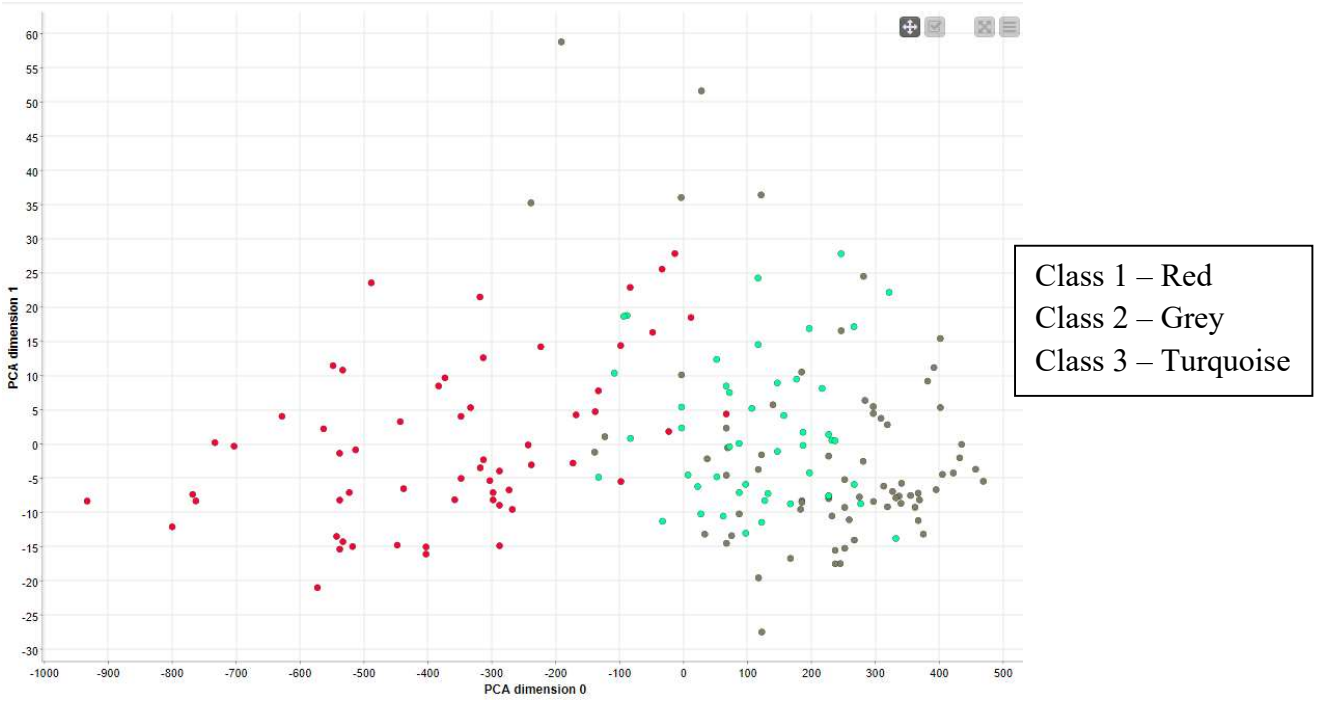


Figure 1: plot1 Graph

Workflow and Node Configurations

The provided dataset was imported, using the CSV Reader node. Principal Component Analysis (PCA) was then applied to the Wine dataset, to produce two dimensions. This eradicates the possibility of the Curse of Dimensionality issue occurring. Next, three colours were applied to the data points, depending on the class. Finally, the data points in the dimension columns were used to generate a Scatter Plot graph, as shown above in Figure 1: plot1 Graph.

1.2: plot2

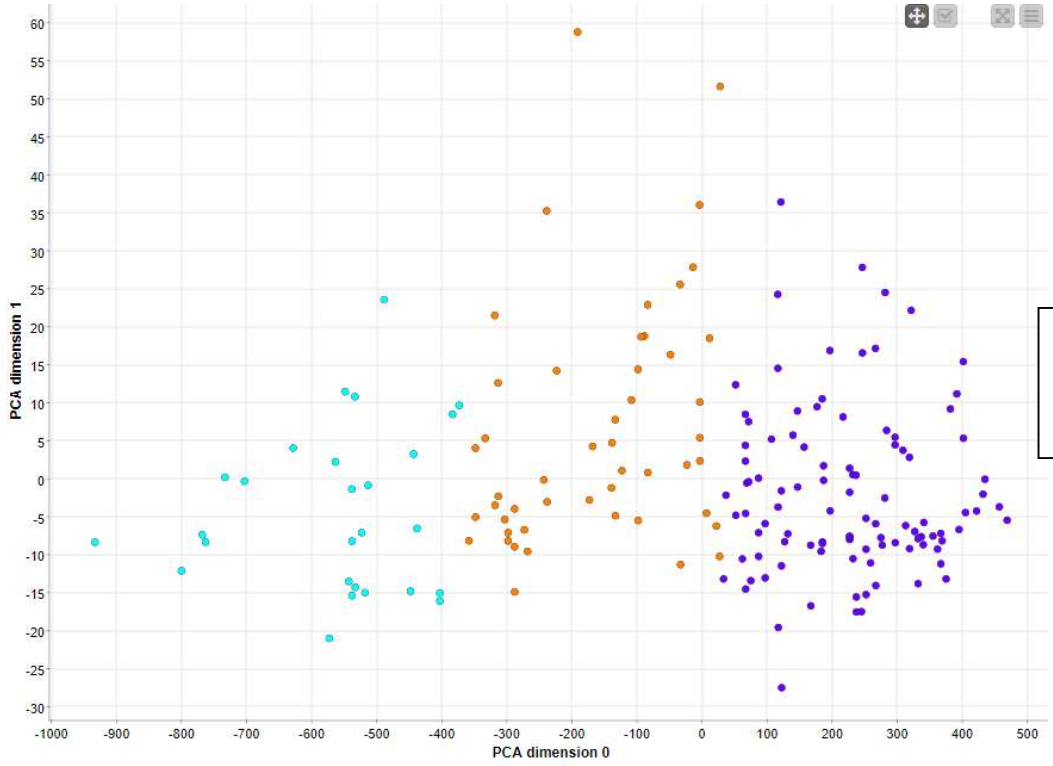


Figure 2: plot2 Graph

Workflow and Node Configurations

K-Means clustering analysis was applied to the dataset, assigning the rows to a cluster label. Colours were assigned to the dataset, based on the cluster labels, before a Scatter Plot graph is produced.

1.3: 2D Plots

In plot1, the data points across all of the classes are spread out and sometimes overlaps data points from other classes. Whereas in plot2, the clusters are less spread out. This indicates the centroids are the closest data point in the cluster.

In plot1, Class 1 data (red) they are more spread out on the left and in plot2, Cluster 1 data (orange) are less spread out overall but are located in the middle.

In plot1, Class 2 data (grey), they are on the right, with a few points in the top middle. In plot2, Cluster 2 data (aqua), they are located on the left.

In plot1, Class 3 data (turquoise), the data points are also on the right but more towards the middle whereas in plot2, Cluster 3 data (purple), the data points are on the right, less spread out compared to other clusters.

1.4: plot3

The graphs are in the appendix.

Workflow and Node Configurations

The data values are split into three, using Row Filter nodes, based on their cluster label. The Colour Manager node then applies colours to the values, based on the class. Finally, the Scatter Plot graphs are created.

Observations

In plot3a, there is a variety of coloured data points, the majority of them being red. This means the first cluster has data points from each of the classes, but the majority of them belonging to Class 1.
All of the data points in plot3b are coloured red, indicating cluster two only contains values from Class 1. The third cluster mainly contains data points from Class 2 and Class 3.
However, as shown in plot3c, only one Red point is seen in this cluster. The rest are coloured grey and turquoise.

Verification

class \ Clus...	1	2	3
1	31	27	1
2	7	0	64
3	11	0	37

Figure 3: Confusion Matrix. Rows IDs Represent Class Numbers. Columns Headers Represent Cluster Numbers.

The figure above show a confusion matrix that verifies the distribution of class labels in each cluster.

1.5: Cluster Validity Measures

Description

Entropy, Purity, and Confusion Matrix are used to evaluate the performance of the K-Means analysis. Entropy is an external evaluation measure which determines if the cluster labels match with the class labels, producing a value. The closer the number is to 0, the better. Purity is also an external evaluation metric that divides the total instances in each cluster by the maximum number of data points per class. This produces a value, representing the purity of the class in the cluster. The closer the value is to 1, the better. The Confusion Matrix produces statistics, such as accuracy, precision, etc, which are helpful in evaluating the performance of K-Means analysis. Accuracy will be the main metric that is used.

Workflow and Node Configurations

The Class and Cluster columns are used as inputs to the Entropy Scorer, producing Entropy and Purity values. To produce the Confusion Matrix statistics, the Cluster attribute is converted to an integer and then to a string. The Class and Cluster attributes are also used as inputs for Scorer and Scorer (JavaScript) nodes.

1.6: Task 1 Results

Results

Clustering statistics

Data Statistics

Statistics	Value
Number of clusters found:	3
Number of objects in clusters:	178
Number of reference clusters:	3
Total number of patterns:	178

Data Statistics

Score	Value
Entropy:	0.942
Quality:	0.4057

Row ID	Size	Entropy	Normali...	Quality
cluster_1	27	0	0	?
cluster_2	102	1.018	0.642	?
cluster_0	49	1.303	0.822	?
Overall	178	0.942	0.594	0.406

Figure 4: Entropy & Purity Values

Scorer View

Confusion Matrix

Rows Number : 178	1 (Predicted)	2 (Predicted)	3 (Predicted)	
1 (Actual)	31	27	1	52.54%
2 (Actual)	7	0	64	0.00%
3 (Actual)	11	0	37	77.08%
	63.27%	0.00%	36.27%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
1	31	18	101	28	52.54%	63.27%	52.54%	84.87%	57.41%
2	0	27	80	71	0.00%	0.00%	0.00%	74.77%	0.00%
3	37	65	65	11	77.08%	36.27%	77.08%	50.00%	49.33%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
38.20%	61.80%	0.109	68	110

Figure 5: Confusion Matrix Statistics

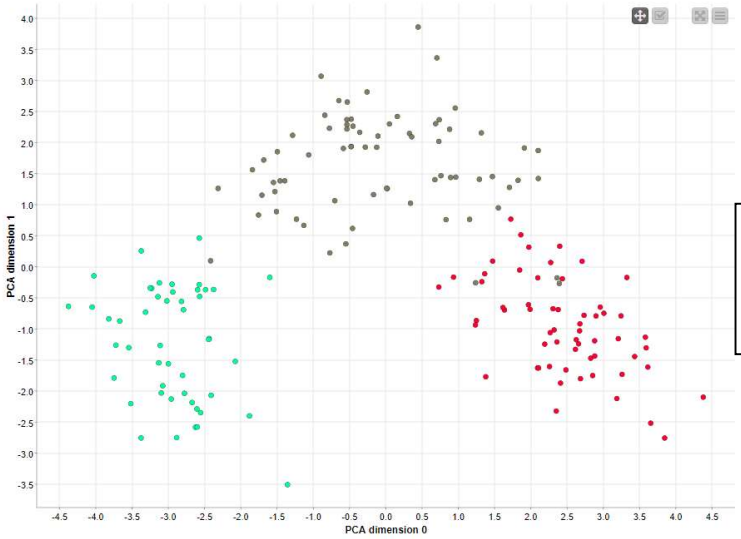
Discussion

The Entropy value is very close to 1 and the Purity value is close to 0 which indicates the K-Means Clustering Analysis performance, for non-normalised data points, is quite inadequate. The statistics, such as accuracy, also suggests the data points were not clustered properly.

Task 2: Clustering With Normalisation

In this task, the clustering algorithm was performed with normalisation. Figure 15 displays the workflow.

2.1: plot1



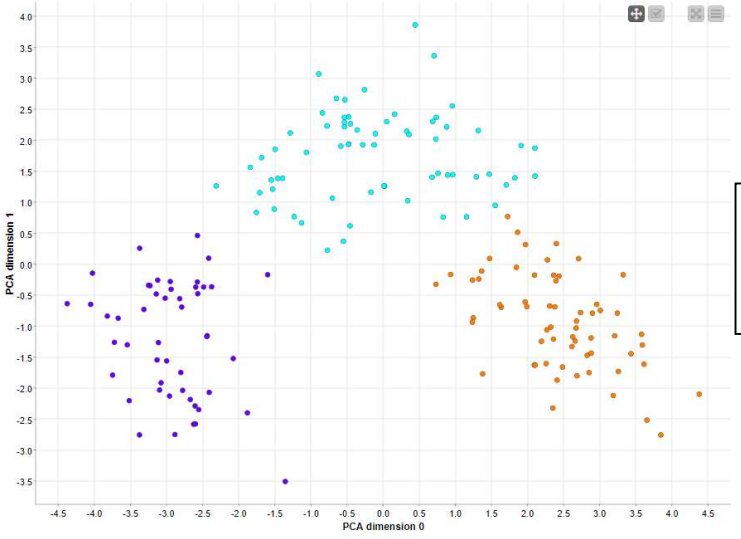
Class 1 – Red
Class 2 – Grey
Class 3 – Turquoise

Figure 6: plot1 Graph

Workflow and Node Configurations

The workflow for Task 2 is almost identical to the workflow of Task 1, but with the exception of the Normaliser node. Specifically, Z-Score normalisation was applied to all of the values, where the mean becomes 0 and the standard deviation becomes 1. The Class column was the only attribute that wasn't normalised. Otherwise, it would greatly affect the Scatter Plot graphs, producing inadequate results. The colours were then assigned to the datapoints, based on the class labels, and the plot1 graph was produced.

2.2: plot2



Cluster 1 – Orange
Cluster 2 – Aqua
Cluster 3 – Purple

Figure 7: plot2 Graph

Workflow and Node Configurations

Clustering Analysis was applied to the normalised values, that belong to the PCA dimension attributes. Colours were allocated to the data points, depending on the cluster labels. Then, the plot2 graph was created.

2.3: 2D Plots

There is not much difference to the shape of the clusters in both plot graphs as the locations of the centroids for all three classes stay the same. Three Class 2 (grey) data points in plot1 became a part of Cluster 1 (orange) in plot2. A Class 2 (grey) data point in plot1 became a part of Cluster 3 (purple) in plot2.

2.4: plot3

The graphs are in the appendix.

Workflow and Node Configurations

The steps that were taken to produce the plot3 graphs in Task 1 was repeated in Task 2. However, this time, the process was applied to normalised values.

Observations

Figure 16 shows that all the data points in plot3a are red apart from 3 grey ones. In plot3b all data points are grey. All data points are green apart from 1 grey data point in plot3c.

Verification

class \ Clus...	1	2	3
1	59	0	0
2	3	67	1
3	0	0	48

Figure 8: Confusion Matrix. Rows IDs Represent Class Numbers. Columns Headers Represent Cluster Numbers.

The figure above show a confusion matrix that verifies the distribution of class labels in each cluster.

2.5: Cluster Validity Measures

Description & Workflow and Node Configurations

Entropy, Purity, and Confusion Matrix are also used in this task. The description of these measures is in section 1.5. The Workflow and Node Configurations are also described in section 1.5.

2.6: Task 2 Results

Results

Clustering statistics

Data Statistics

Statistics	Value
Number of clusters found:	3
Number of objects in clusters:	178
Number of reference clusters:	3
Total number of patterns:	178

Data Statistics

Score	Value
Entropy:	0.1369
Quality:	0.9136

Row ID	I Size	D Entropy	D Normali...	D Quality
cluster_1	67	0	0	?
cluster_2	49	0.144	0.091	?
cluster_0	62	0.28	0.176	?
Overall	178	0.137	0.086	0.914

Figure 9: Entropy & Purity Values

Scorer View

Confusion Matrix

Rows Number : 178	1 (Predicted)	2 (Predicted)	3 (Predicted)	
1 (Actual)	59	0	0	100.00%
2 (Actual)	3	67	1	94.37%
3 (Actual)	0	0	48	100.00%
	95.16%	100.00%	97.96%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
1	59	3	116	0	100.00%	95.16%	100.00%	97.48%	97.52%
2	67	0	107	4	94.37%	100.00%	94.37%	100.00%	97.10%
3	48	1	129	0	100.00%	97.96%	100.00%	99.23%	98.97%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
97.75%	2.25%	0.966	174	4

Figure 10: Confusion Matrix Statistics

Discussion

The Entropy value is very close to 0 and the Purity value is very close to 1 which indicates the K-Means Clustering Analysis performance, for normalised data points, is quite satisfactory. The statistics, such as accuracy, also suggests the data points were clustered properly.

Results and Discussion

This section compares the results from Task 1 and Task 2.

Comparison of Results

The normalised data points in plot1 seemed to have more of a cluster appearance as opposed to the non-normalised data points in plot1 – See Figures 1 and 6. As shown in the plot2 graphs, the locations of the centroids in Task 2 did not change at all as opposed to the locations of the centroids in Task 1. The majority of the data points in all of the plot3 graphs in Task 2 are of the same colour, as compared to the plot3 graphs in Task 1.

Cluster Validity Measures	Task 1 – Without Normalisation	Task 2 – With Normalisation
Entropy	0.942	0.137
Purity	0.406	0.914
Accuracy	38.20%	97.75%

Discussion

In Task 2, the performance of the clustering analysis was much better than in Task 1. As shown in the table above, the accuracy without normalisation was approximately 38% compared to 98% with normalisation. This is because the values are in the same range and hence more accurate results were produced.

Conclusion

To conclude, normalisation does affect the results produced from clustering analysis in a positive way.

References

Dabbura I (2018, September 17). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Pulkit S (2019, August 19). *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need*. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Mohanty A (2018, December 11). *Curse Of Dimensionality and PCA*. https://adityaroc.medium.com/curse-of-dimensionality-and-pca-f90f1258a7f2

Zach (2021, August 12). *Z-Score Normalization: Definition & Examples*. https://www.statology.org/z-score-normalization/#:~:text=Z%2Dscore%20normalization%20refers%20to,the%20standard%20deviation%20is%201.

Appendix

Task 1

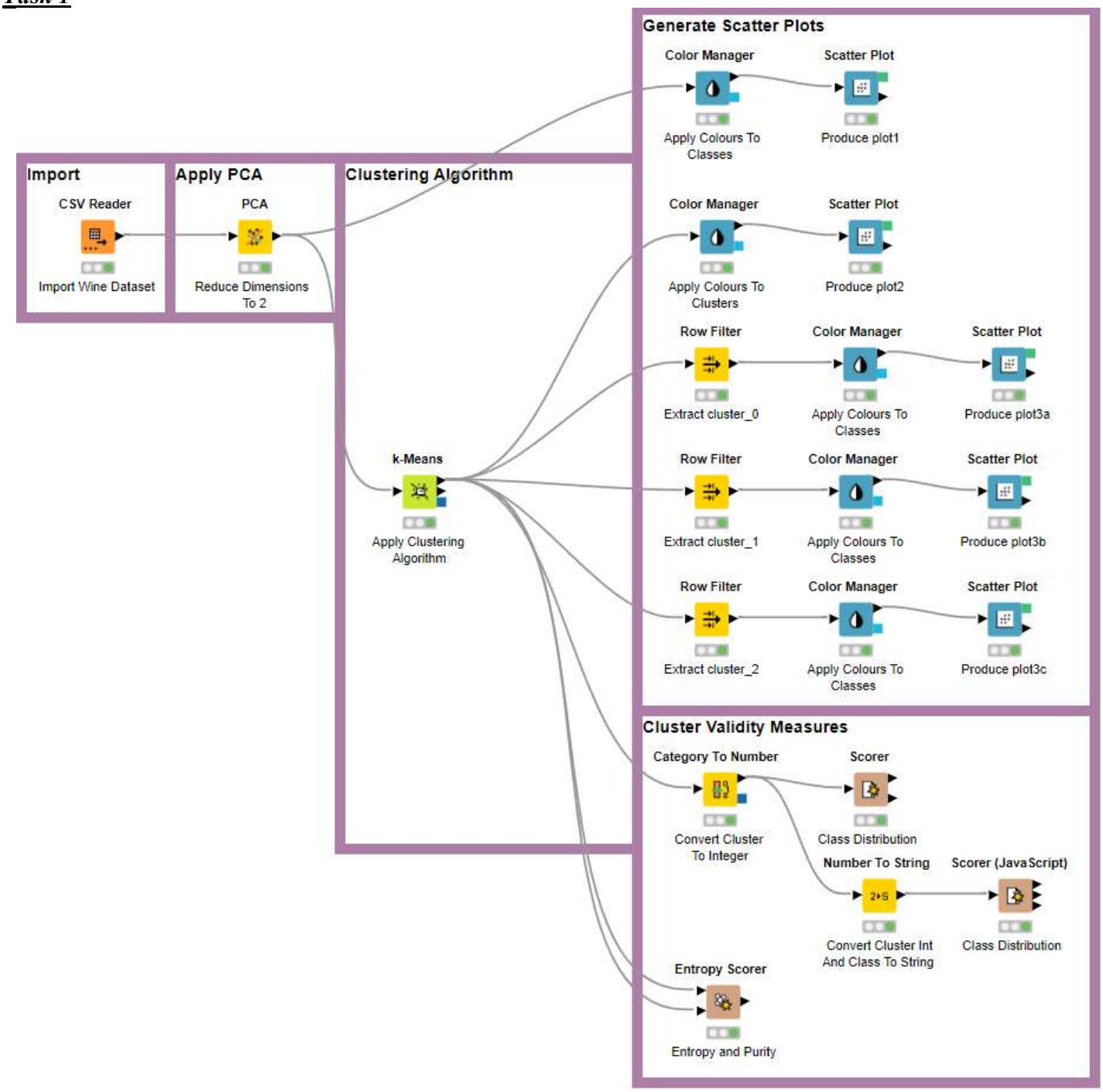
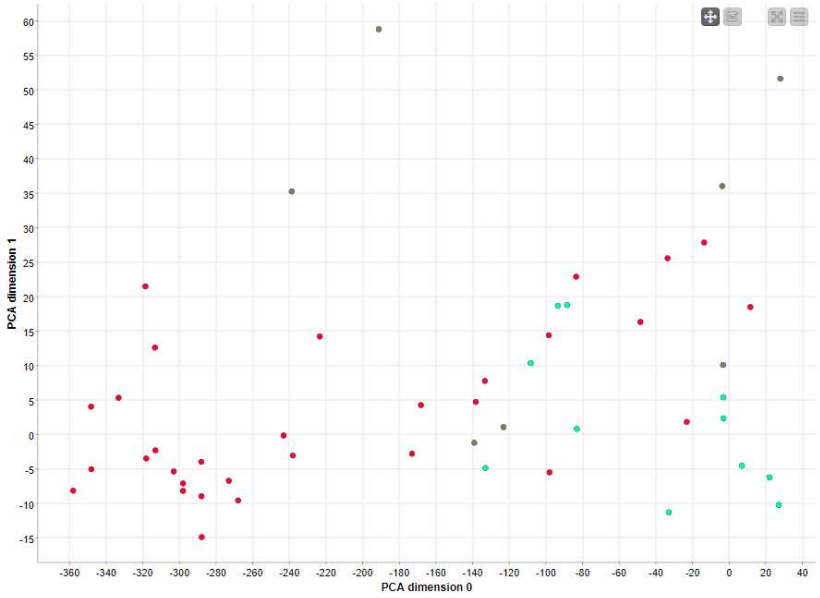
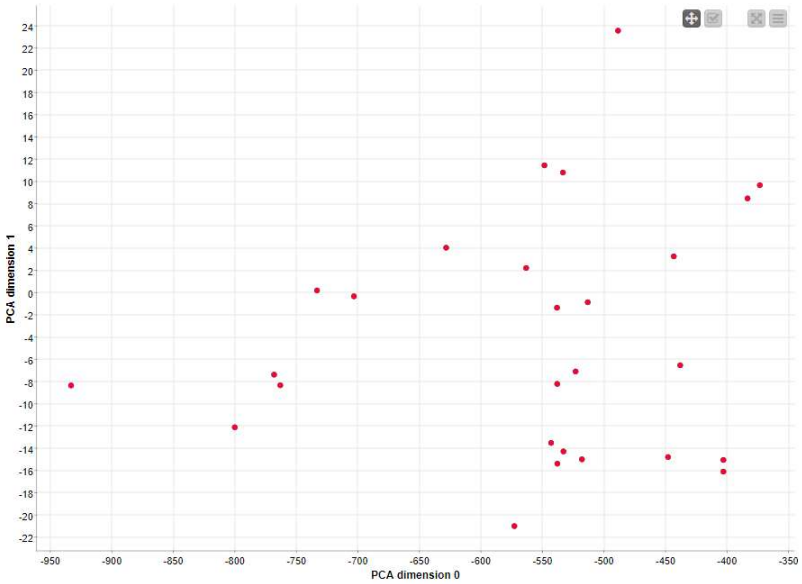


Figure 11: KNIME Workflow



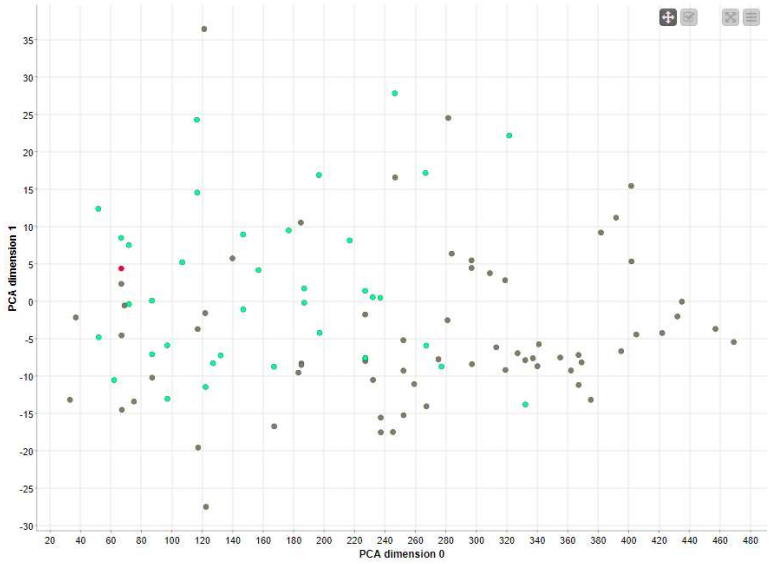
Class 1 – Red
Class 2 – Grey
Class 3 – Turquoise

Figure 12: plot3a Graph



Class 1 – Red
Class 2 – Grey
Class 3 – Turquoise

Figure 13: plot3b Graph



Class 1 – Red
Class 2 – Grey
Class 3 – Turquoise

Figure 14: plot3c Graph

Task 2

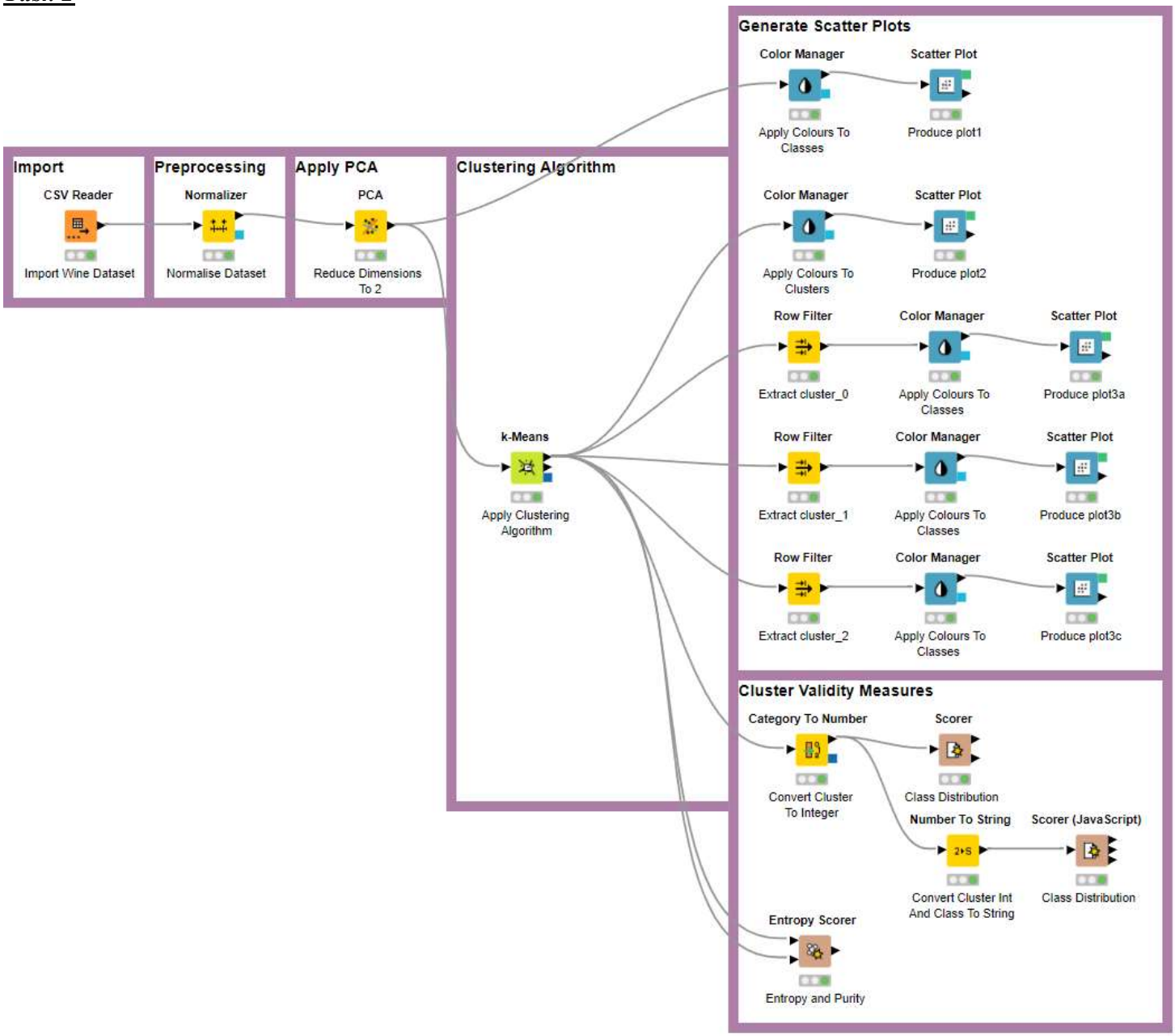


Figure 15: KNIME Workflow

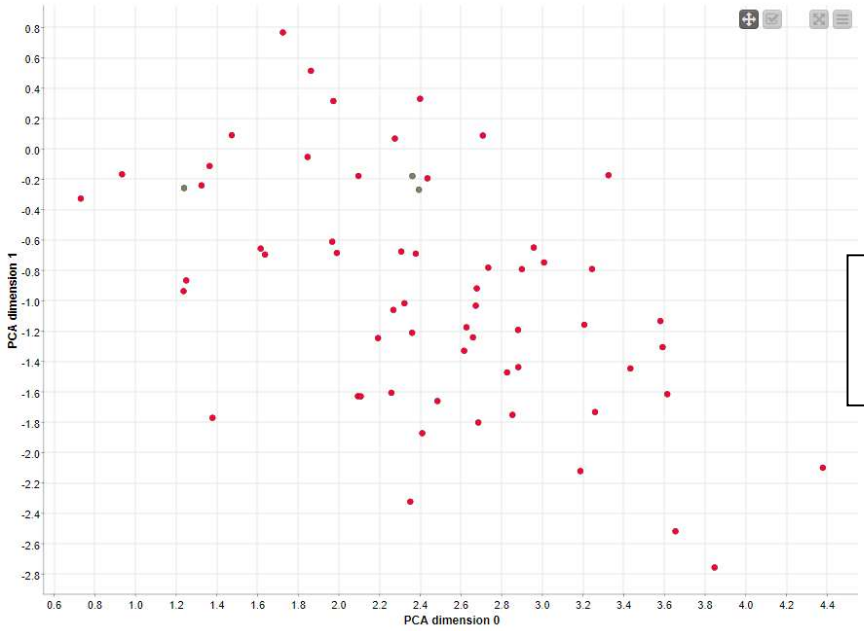


Figure 16: plot3a Graph

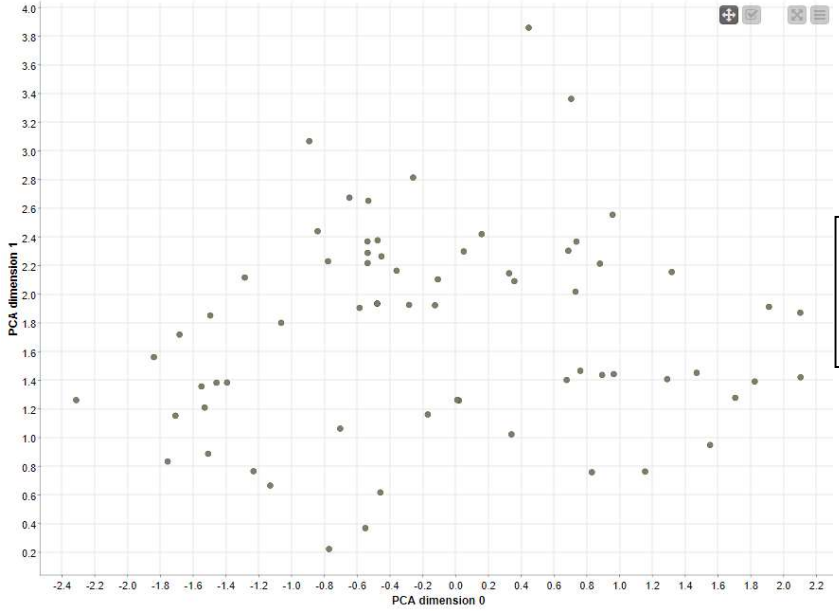


Figure 17: plot3b Graph

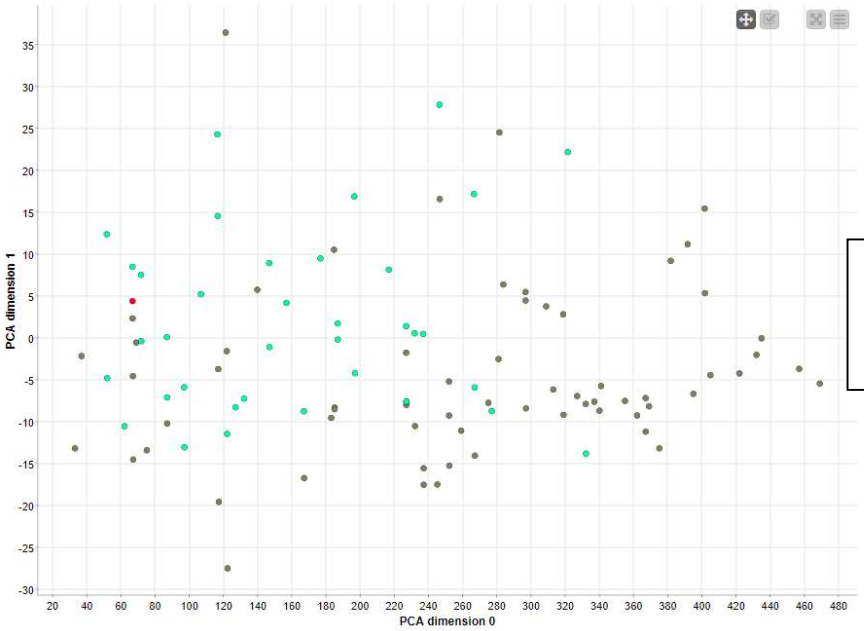


Figure 18: plot3c Graph