# HEART DATA ANALYSIS – REPORT

## BY NATHAN CARNEY

### Introduction

The purpose of this report is to carry out exploratory analysis on the Statlog Heart Data Set. This dataset consists of observations of 10 physiological characteristics in 270 patients. The aim of the analysis is to attempt to effectively classify patients into subgroups depending on their characteristics and to develop models to predict the risk of developing heart disease in patients.

Both supervised and unsupervised methods will be used to conduct this analysis.

Unsupervised (or classification) methods are those which have "no output variables to predict. The objective of this class of data science techniques, is to find patterns in data based on the relationship between data points themselves" [1] (Kotu & Deshpande, 2019). In the case of this report, these methods will find patterns based on patients' similar characteristics.

The unsupervised methods chosen were: Principal Components Analysis, Hierarchical Clustering and K-Means Clustering.

Unsupervised methods in data analysis have an output variable which they must predict. In this case, unsupervised methods will be used to predict if any given patient will develop heart disease.

K-Nearest-Neighbours, Linear Discriminant Analysis and Quadratic Discriminant Analysis were the supervised methods chosen for this analysis.

### Summary of Data

The data set consists of 10 variables and 270 observations. Below is a table outlining the first 5 observations (patients) in the raw (unchanged) data set, with their corresponding physiological characteristics:

| Age | Sex | RestBloodPressure | SerumCholestoral | FastingBloodSugar | MaxHeartRate | ExerciseInduced | Slope | MajorVessels | Class |
|---|---|---|---|---|---|---|---|---|---|
| 70 | 2 | 130 | 322 | 1 | 109 | 1 | 2 | 4 | 2 |
| 67 | 1 | 115 | 564 | 1 | 160 | 1 | 2 | 1 | 1 |
| 57 | 2 | 124 | 261 | 1 | 141 | 1 | 1 | 1 | 2 |
| 64 | 2 | 128 | 263 | 1 | 105 | 2 | 2 | 2 | 1 |
| 74 | 1 | 120 | 269 | 1 | 121 | 2 | 1 | 2 | 1 |

The dataset consists of continuous and binary variables. The 'Class' variable signifies the absence (1) or presence (2) of heart disease. For each of the other binary variables, It is not known which value corresponds with which state of the variable (for example, it is not known if '1' or '2 means 'Male' or 'Female' under the 'Sex' column).

The 'Fasting Blood Sugar' variable indicates if a patient's blood sugar was above or below 120 mg/dl, 'Exercise Induced' records if a patient experience Exercise Induced Angina after extended periods of physical activity. The

[1] https://www.sciencedirect.com/topics/computer-science/unsupervised-data#:~:text=In%20unsupervised%20data%20science%2C%20there,relationship%20between%20data%20points%20themselves.&text=Classification%20and%20regression%20techniques%20predict%20a%20target%20variable%20based%20on%20input%20variables.

'Slope' variable is a measure of how quickly a person's heart rate increases or decreases when undergoing exercise or stopping exercise, respectively. 'Major Vessels' is a count of the major vessels present in a patient.

Below is a table displaying the mean and standard deviation of each continuous variable:

| | Age | RestBloodPressure | SerumCholestoral | MaxHeartRate | Slope | MajorVessels |
|---|---|---|---|---|---|---|
| Mean | 54.433333 | 131.34444 | 249.65926 | 149.67778 | 1.5851852 | 1.6703704 |
| Standard Deviation | 9.109067 | 17.86161 | 51.68624 | 23.16572 | 0.6143898 | 0.9438964 |

## Adjustments to Data

### Use of Binary Variables

A major recurring issue encountered throughout the analysis conducted in this report was the presence of binary variables. At every stage, for virtually every method, the inclusion of binary variables seemed to give minimal marginal information about each patient with regard to the results of PCA, HC and K-Means. Also, as stated above, it was not known which implication corresponded with each value for the binary columns (Sex, Fasting Blood Sugar and Exercise Induced). Hence, even if these were included in analysis, their values would be relatively meaningless when trying to extrapolate information about certain groups of patients. One method in which binary variables were included was Linear and Quadratic Discrminant Analysis, as these models are designed to be able to use variables of all data types, without needing to know the meaning behind each variable's value. It is also important to note that the 'Class' variable was not included in any unsupervised methods, for the reason that it is a dependent variable and should only be used for supervised methods.

### Outliers

Outliers were also removed from the full data set. This was done to improve clustering accuracy, as outliers may have resulted, in the case of unsupervised methods, in classifications which were not explanatory of the data being examined. The outliers were identified by calculating the mean value for each continuous variable (for instance Age mean was 54.33, as displayed above) and its standard deviation. If any observation was more than three standard deviations away from the mean, it was removed from the dataset. This process resulted in 7 observations being taken out of the data set of 270 observations. This was a relatively low figure, hence I felt confident that it would not result in analysis being less accurate or mischaracterized.

## Unsupervised Methods

### Principal Components Analysis

Principal Components Analysis is a method of dimension reduction. In short, this means that the number of variables (dimensions) involved in a data set are distilled down into the data set's most explanatory (important) components. In the case of the heart data, I immediately thought it necessary to try and reduce the number of dimensions to the data, as there were 6 variables and each had 263 observations to examine.

The process of Principal Components involves constructing a covariance or correlation matrix which outputs the covariance or correlation of each variable compared to the other for every column. Eigenvalue decomposition is then performed on these matrices and linear combinations (principal components) are produced which explain a certain aspect of the data. These components are uncorrelated (independent). By analysing the coefficients (loadings) of these linear combinations, prominent groups within a data set can be identified.

### Standardization

There were several assumptions and subsequent adjustments made before performing PCA on the Heart Data Set, Firstly, variables were standardized. This was done due to the existence of different units of measurement present in various data columns (for instance Maximum Heart Rate and Cholestoral). To standardize, the 'scale'

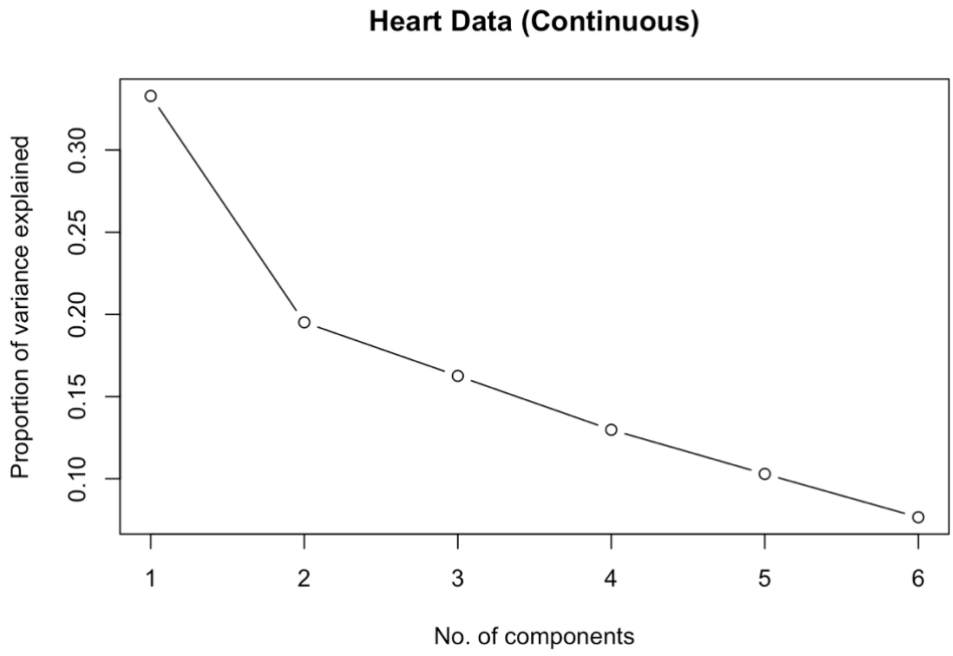
function in R was used which divides each value in a column by its standard deviation which results in every observation's variance being equal to one.

## Results

The following results were observed after conducting PCA Analysis on the continuous data set:

Principal Components for Standardised Continuous Heart Data

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| Age | -0.5458329 | 0.1529892 | -0.1325309 | 0.2673160 | 0.4618729 | -0.6134426 |
| RestBloodPressure | -0.3033283 | 0.4587896 | 0.5826836 | 0.4665142 | -0.1695354 | 0.3340749 |
| SerumCholestoral | -0.2170500 | 0.6462700 | -0.0602605 | -0.7258028 | -0.0692967 | -0.0011299 |
| MaxHeartRate | 0.5106749 | 0.3640163 | 0.0308513 | 0.2115961 | -0.4378089 | -0.6077022 |
| Slope | -0.3560534 | -0.4638160 | 0.5105694 | -0.3084993 | -0.4173671 | -0.3578440 |
| MajorVessels | -0.4187897 | -0.0284298 | -0.6145332 | 0.2101718 | -0.6222999 | 0.1213531 |




Below is a summary of these results:

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6
## Standard deviation     1.413 1.0822 0.9876 0.8824 0.7859 0.67781
## Proportion of Variance 0.333 0.1952 0.1626 0.1298 0.1029 0.07657
## Cumulative Proportion  0.333 0.5282 0.6907 0.8205 0.9234 1.00000
```

Several methods exist for choosing how many Principal Components to use. For my analysis, I decided to do this by visual means. There is a convention whereby the number of PC's to include is done by identifying a 'kink' (or 'elbow') in the curve of the proportion of variance explained by each component against the number of components. In this case, the 'kink' clearly appears after PC2. However, as is evident in the summary above, the first 2 Principal Components only explain roughly 50% of the data. I found it appropriate to include 3 Principal

Components, as these explain just under 70% of the data's variance. When examining components, I chose to consider a 'significant' loading as one with a value of 0.4 or higher.

### Analysis of Principal Components

PC1 – This PC seems to comprise of subjects with relatively high maximum heart rate and a young age, as illustrated by the magnitude of the loadings and their signs for Age (-0.54) and Max Heart Rate (.51). These loadings also point to an inverse relationship between the variables. This intuitively makes sense, as in general when people are younger their heart is capable of reaching higher bpm. These subjects also have fewer major vessels appearing in the fluoroscopy.

PC2 – The group encompassed by this PC's variance are those with high Blood Pressure (.46) and high Cholestoral (0.65), alongside a low value for Slope (-.46). This PC accounts for roughly 20% of the overall variance in the data.

PC3 – This group accounts for 16% of the total variance. Itis made up of patients with a high value for Slope and Blood Pressure, and a low value for Major Vessels.

### Binary Variables

Binary variables were not included in PCA as their loadings (PCA values) did not provide any useful information in analysing the data set. This issue was first mentioned in the 'Use of Binary Variables' section above, where it was outlined that the meaning of each value is not known. In addition, the inclusion of binary variables did not change the loadings of each PC by very much.

### Hierarchical Clustering

### Outline

Hierarchical Clustering (HC) is another unsupervised classification method to identify groups or 'clusters' within data which bring together variables which are similar to each other. The algorithm used is agglomerative, whereby each variable is set as its own cluster initially, and subsequently they are grouped together according to their similarity. The goal of clustering is to have distinct groups in which group members are similar, but each group is different.

Two processes are involved in the development of a HC model:

1. Dissimilarity Measure
2. Linkage Method

Some general explanation of these processes should be given before analyzing results.

A dissimilarity measure is a method of discerning the (dis)similarity of data points compared to one another. Many approaches are available for this, yet the choice of distance measure tends to be relatively unimportant for clustering as most will provide similar results. The output of this measure is entered into a matrix.

Linkage methods are responsible for grouping together observations which have similar characteristics, outlined by the chosen dissimilarity measure. The choice of linkage method heavily influences the result of clustering, as essentially you are deciding how similar observations must be to be grouped together in a given cluster.

### Assumptions

I decided that the clustering would be conducted using the adjusted data set with outliers removed, to ensure accurate clusters, and the continuous variables remained standardized to account for different units of measurement.

From my research of various scholarly papers and from previous work with clustering, it was clear that Euclidean Distance was the most popular measure for continuous data. Euclidean distance takes the square root of the squared difference of two variables, providing a positive numeric value indicating their similarity. This method was chosen for all of the analysis conducted on continuous variables.

For my analysis, I chose to use two different linkage methods to be used with the Euclidean dissimilarity matrix; Complete Linkage and Ward Linkage.

## Binary Variables

The inclusion of binary variables presented the most problems for HC compared with any other method in this report.

Initially, it was attempted to use binary variable whose values were adjusted from 2 and 1, to 1 and 0 respectively (R requires binary values to be presented this way to perform dissimilarity measures).

The dendograms produced by Jaccard and R's Binary dissimilarity measure with Ward and Complete Linkage did not prove useful in analysis. To By joining the two data types together, the clustering would give a broader overview of the groups of patients that may be present within the Heart data set. However, while this seemed an optimal way to approach the problem, I saw it as too problematic. In order to 'join' the two data types together and the perform clustering, an appropriate weight had to assigned to the columns (variables) in both the continuous and binary data sets. Many different methods have been proposed in the past, notably by Huang (1998)[2] and Oppong (2018)[3], by whom various methods of handling mixed data in clustering are discussed. Both argue that a weight should be applied to the data according to the importance of each variable (or column), however, the importance of each variable in this case is not known. s

With this ambiguity surrounding the importance of each variable and the general recommendation by bodies such as IBM [4] not to use binary data in Hierarchical clustering, I decided not to include binary data for clustering. It is important to note that this will cause the results of clustering to explain less of the data. However, I felt it necessary in order to avoid inaccurate and arbitrary results.

I also found that Ward Linkage could not be used with binary variables, as the Ward method computes geometric centroids for the building of clusters, which requires Euclidean distance to be used on the dataset. Complete linkage with this combine data set did not yield significant results.

## Results

The dendograms for Complete and Ward Linkages are displayed below. (Ward on the right, Complete on the left). The dissimilarity between clusters is reflected in the height of the lines joining them. To determine the number of clusters from each dendogram, a 'cut-off line' was calculated to intersect the dendogram a certain number of times, with each intersection representing one cluster present. The cut-off line was calculated by adding the mean height of the dendogram and three times the standard deviation of this height. I immediately observed that the Ward method displays a group of 6 clusters which exhibit the ideal characteristics of a HC result. Clusters are very dissimilar, and cluster members are quite similar. Complete Linkage results in a 7-cluster solution with clusters being more similar to each other compared to Ward. It should be noted that by design the Ward linkage method produces dendograms which seem optimal by visualization. However, due to the fact that my later K-means clustering solution matched the number of clusters for this linkage method, I deemed it correct.

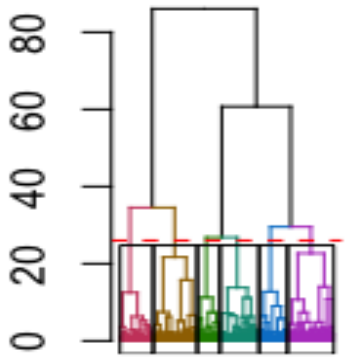
The Rand Index (a measurement of agreement between two linkage methods) of the Ward and Complete Linkages is shown below:

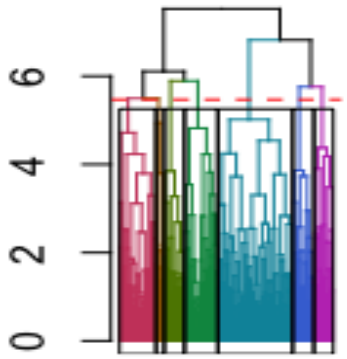
---

[2] (Huang, 1998)
[3] (Oppong, 2018)
[4] https://www.ibm.com/support/pages/clustering-binary-data-should-be-avoided

```
       RI
0.7830668
```

This was interpreted as a positive result, however when the Adjusted Rand Index was computed (the Rand Index with a correction for chance), the result was a value of 0.28. This would be classed as a poor recovery. Hence, the clustering solution should not be taken as completely accurate.



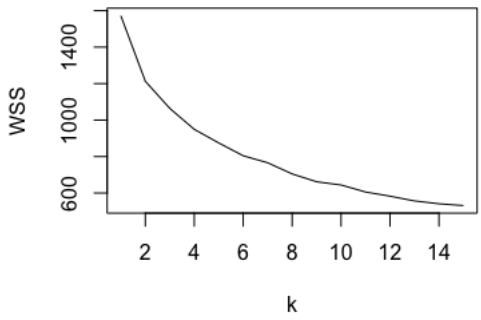




## K-Means Clustering

Unlike HC, which is an agglomerative clustering method, K-Means Clustering is divisive. This is a method which "begins with all patterns in a single cluster and performs splitting until a stopping criterion is met".[5] This method divides a data set into k groups, each with its own 'cluster mean'. The observations are classified into each group by their proximity to each central mean.

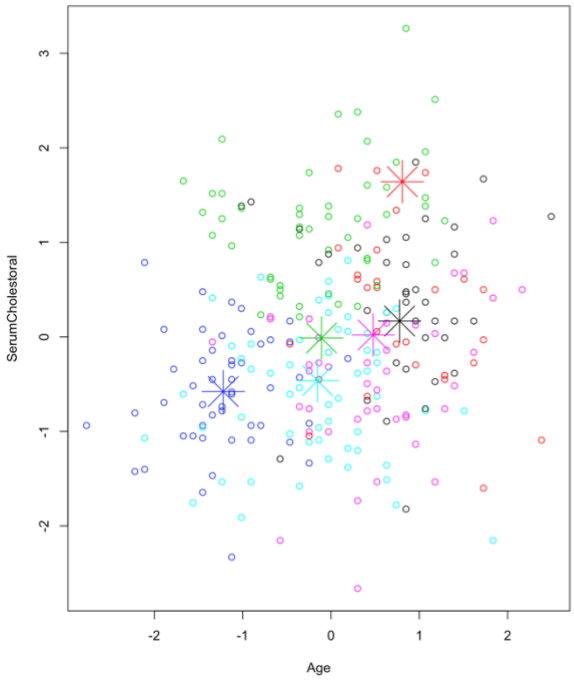
K-means clustering uses a Euclidean distance method to calculate a point's proximity to its group mean. Hence, this analysis was not attempted with the inclusion of binary variables as their distances would be insignificant.

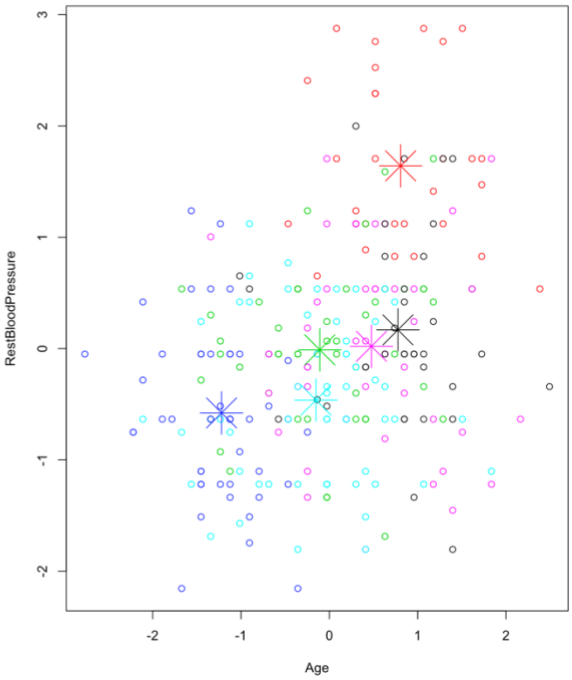
First, I ran the k-means algorithm using 15 values for k (1 to 15). Similar to PCA, the number of groups to be used can be identified visually using a plot of k against its Within Sum Squared error. In the plot below, I identified a 'kink' or elbow at value 6. Hence, I chose 6 for my number for k.




[5] https://www.datacamp.com/community/tutorials/hierarchical-clustering-R

Below is an illustration of the various clusters present in Age against Cholestoral and Blood Pressure. There seems to be a group comprised of young people whose age was low and blood pressure was low, along with a group of young people with low cholestoral in the adjacent plot.



## Supervised Methods

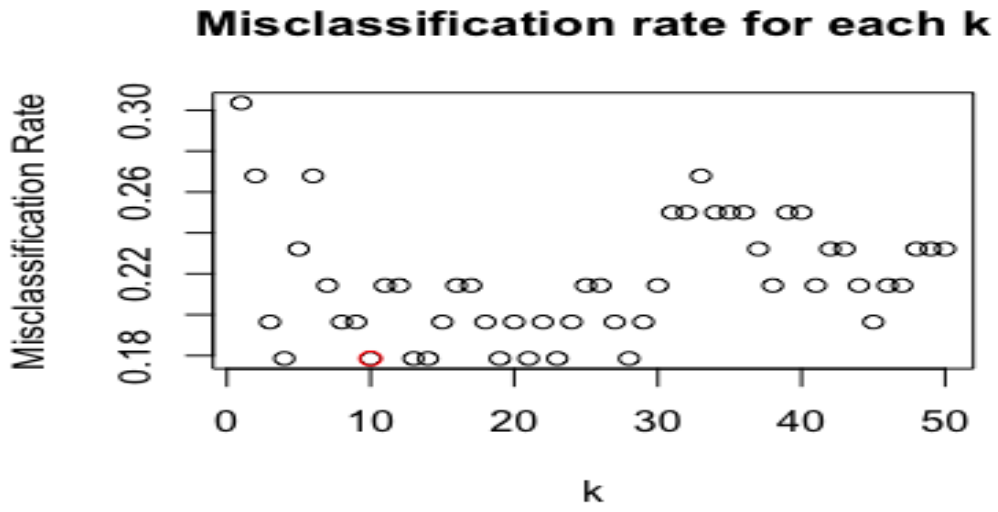
I will now introduce the 'Class' variable to analysis. Using supervised methods, patients can be classified according to whether or not they have heart disease, given their characteristics.

### K-Nearest-Neighbours (KNN)

KNN is an algorithm which classifies observations based on their 'likeness' or distance to their most similar counterparts. Through grouping together in this manner, the outcome (Class) can be estimated for each patient.

To begin, I chose to use a training/test/validation split of 50/25/25. After examining the Heart dataset, I discovered that 115 had heart disease and 148 didn't. It is worth noting that this was after outliers were removed. To split the data appropriately, I had to use an even number of observations from each classification (heart disease or no heart disease). Hence, I removed the last 33 observations from those without heart disease. After splitting the data, I was left with 112, 56 and 56 variables in the training, test and validation sets respectively.

Below can be seen the Misclassification rate of each choice of k neighbours. A rule of thumb exists where the optimum number of neighbours is usually the square root of the number of observations in the training set. Hence, I chose 10 for k. This point in the plot is circled in red. Its misclassification rate in the validation data set was 23%, as shown below.




```
[1] 0.2321429
```
(Missclassification rate of k)

### Linear Discriminant Analysis (LDA)

LDA is a method which uses Bayesian probability and covariance matrices to create linear combinations which can predict an object's classification based on their characteristics. LDA assumes a multivariate normal distribution for datasets and combines these to form models for prediction of classes. For this analysis, all variables were used, as LDA does not depend on distance measures.

Similar to KNN, a training and test data set were compiled to train the algorithm and test its efficacy. A validation set was not used as I opted for cross validation as a validation method. I chose an 80/20 split in the data, and once an even number of observations from each class was included, there were 184 observations in the training set and 46 observations in the test set.

After training the algorithm, I ran it on the test data set. Its misclassification rate is given below:

```
[1] 0.3181818
```
LDA Misclassification on test data

I took this as a positive result, as heart disease was predicted correctly 69% of the time.

### Quadratic Discriminant Analysis (QDA)

I also conducted QDA which, unlike LDA assumes that each class has its own covariance matrix. After conducting analysis using the same training and test sets above, the below misclassification rate was found:

```
[1] 0.7272727
```
QDA Misclassification on test data

This was significantly worse than LDA, and hence I chose LDA as the correct model.

### Conclusion

My analysis provided no clear at-risk groups in the data. Each unsupervised classification method showed no definite separation of the Heart Data Set into those who were more or less at risk of developing heart disease.

Supervised methods provided a more comprehensive result. LDA seemed to be able to predict those who would develop heart disease accurately.

## Bibliography

[1] https://www.sciencedirect.com/topics/computer-science/unsupervised-data#:~:text=In%20unsupervised%20data%20science%2C%20there,relationship%20between%20data%20points%20themselves.&text=Classification%20and%20regression%20techniques%20predict%20a%20target%20variable%20based%20on%20input%20variables.

[2] (Huang, 1998) 'Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values' Data Mining and Knowledge Discovery 2 1998

[3] (Oppong, 2018) 'Clustering Mixed Data: An Extension of the Gower Coefficient with Weighted L2 Distance'

Augustine Oppong, 2018

[4] https://www.ibm.com/support/pages/clustering-binary-data-should-be-avoided

[5]https://www.datacamp.com/community/tutorials/hierarchical-clustering-R