

**RIGA TECHNICAL UNIVERSITY**  
**FACULTY OF COMPUTER SCIENCE AND INFORMATION**  
**TECHNOLOGY**  
**INSTITUTE OF APPLIED COMPUTER SYSTEMS**

**Fundamentals of Artificial Intelligence**

**Practical Assignment #2**

**Applying methods of machine learning**

**GitHub: <https://github.com/MustafaKemalV/Practical-2-AI.git>**

**Author: Mustafa Kemal VURAL**

**Student ID: 201ADB076**

## **CONTENT:**

<b>Heart Failure Prediction Dataset</b>	<b>3</b>
<b>Description of the dataset</b>	<b>3</b>
<b>Description of the content of the dataset</b>	<b>5</b>
<b>The Classes in the Dataset</b>	<b>6</b>
<b>1. Data Exploration</b>	<b>9</b>
<b>1.1 Scatter Plot</b>	<b>11</b>
<b>1.2 Histogram</b>	<b>13</b>
<b>1.3 Distributions</b>	<b>15</b>
<b>2. Unsupervised Learning</b>	<b>18</b>
<b>2.1 Hierarchical clustering</b>	<b>18</b>
<b>2.3 K-Means clustering</b>	<b>21</b>
<b>3. Supervised Learning</b>	<b>28</b>
<b>3.1 Neural Network</b>	<b>28</b>
<b>3.2 K-Nearest Neighbor</b>	<b>31</b>
<b>Workflow</b>	<b>33</b>
<b>References</b>	<b>35</b>

## **Heart Failure Prediction Dataset**

### **Description of the dataset**

Last year (2022), 67.1 million people died worldwide, according to data from Our World In Data. Every year, the number of deaths increases in direct proportion to population growth. There are thousands of different causes of death. The most common cause of death is cardiovascular disease (CVD). Cardiovascular disease (CVD) is the No. 1 cause of death, claiming an estimated 17.9 million lives each year and accounting for 31% of all deaths worldwide. There are 5 most common Cardiovascular diseases. In short, we can call them the 5 CVDs. These are Heart Attack, Stroke, Heart Failure, Arrhythmia and Heart Valve Complications. Heart attack and stroke represent four out of every five CVD deaths and a third of these deaths occur before the age of 70. Heart failure is a common complication of CVDs and this dataset contains 11 factors that can be used to predict the development of a potential heart condition.

A machine learning model, based on the context of the dataset Federico Soriano Palacio has provided us with, could be very helpful in the early detection and management of people with cardiovascular disease or at high cardiovascular risk.

We got the dataset we will use in this practical Assignment from the Kaggle website, from the source [1] edited and published by Federico Soriano Palacios. I also present the information "<https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>" that the author shared with us and showed as the sources of this dataset. On the other hand, the information shared with us for license is "<https://opendatacommons.org/licenses/odbl/1-0/>".

As our author explained to us, in this dataset, 5 heart datasets are combined over 11 common features, all of which are independently found datasets. Putting it all together and organizing it makes it the largest heart disease dataset ever available. The five datasets used for curation are, along with their creators and hospital names: [1]

Cleveland: 303 observations

Hungarian: 294 observations

Switzerland: 123 observations

Long Beach VA: 200 observations

Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

Final dataset: 918 observations

The creators and relevant hospital names of these observations are as follows:  
[1]

Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

## Description of the content of the dataset

To briefly describe the content of our dataset, we have 918 data objects. In addition to this, our author's Attribute Information has presented us item by item with their meanings and value ranges information etc. These are the following: [1]

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

## **The Classes in the Dataset**

We have two classes in this dataset. The classes we have represent the possibility of patients having heart failure.

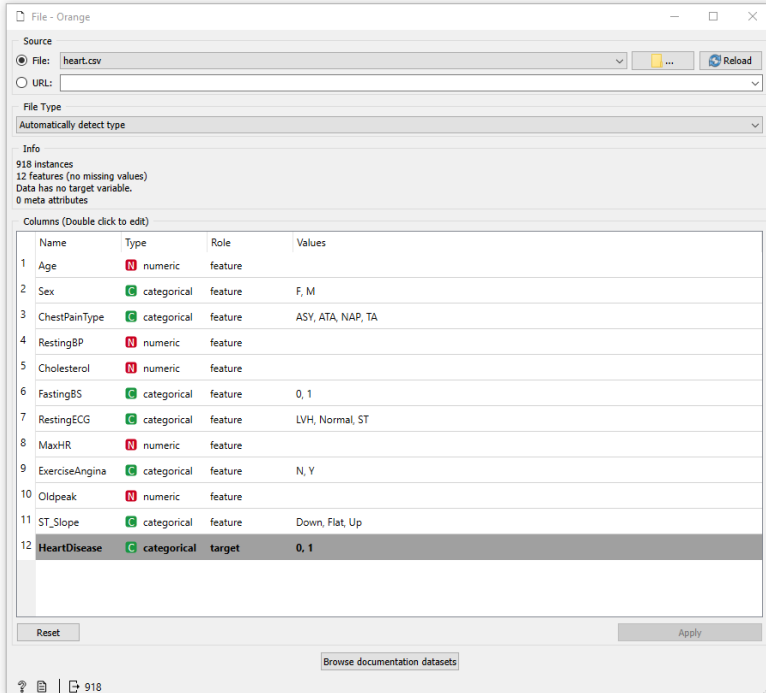
Starting from here, we can define our two classes as Class 0 and Class 1. Class 1 represents heart failure while Class 0 represents absence of heart failure, so no heart failure. Based on this information, we can say that for the label and class relationship, we can say that "0" represents our Class 0 and "1" labels represent our Class 1.

Feature	Meaning	Value Type	Value Range
age	Age of the patient	Numeric	[28, 77]
sex	Gender if the patient	Binary	0**, 1**
ChestPaintType	Determination of the type of pain	Numeric	[1,4]
RestingBP	Blood Pressure	Numeric	[0, 200]
Cholesterol	Blood cholesterol levels	Numeric	[0, 603]
FastingBS	blood sugar after an overnight fast	Binary	0*, 1*
RestingECG	detect abnormalities	Numeric	[1,2]
MaxHR	maximum heart rate	Numeric	[60, 202]
ExerciseAngina	pressure in the chest, jaw or arm.	Binary	0*, 1*
Oldpeak	depression induced by exercise relative to rest	Numeric	[-2,6, 6,2*]
ST_Slope	shift relative to exercise-induced increments in heart rate	Numeric	[1,3]
HeartDisease	Disease	Binary	0*, 1*

- 0\* meaning is "No" and 1\* meaning is "Yes".

- 0\*\* meaning is "Female", 1\*\* meaning is "Male".
- ChestPainType: Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic
- ST\_Slope: → Value1: [Up: upsloping], Value2: [Flat: flat], Value3: [Down: downsloping]
- RestingECG: [Value 1 Normal: Normal, Value 2 ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

Figure 1.1 Here is a screenshot of how our dataset looks in the orange tool. we also determined our features and target



File

File - Orange

Source

File: heart.csv

File Type

Automatically detect type

Info

918 instances  
12 features (no missing values)  
Data has no target variable.  
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	Age	numeric	feature	
2	Sex	categorical	feature	F, M
3	ChestPainType	categorical	feature	ASY, ATA, NAP, TA
4	RestingBP	numeric	feature	
5	Cholesterol	numeric	feature	
6	FastingBS	categorical	feature	0, 1
7	RestingECG	categorical	feature	LVH, Normal, ST
8	MaxHR	numeric	feature	
9	ExerciseAngina	categorical	feature	N, Y
10	Oldpeak	numeric	feature	
11	ST_Slope	categorical	feature	Down, Flat, Up
12	HeartDisease	categorical	target	0, 1

Reset Apply

Browse documentation datasets

918

Figure 1.1



Figure 1.2 An example table of how our dataset looks in excel

The image shows the Orange Data Table widget interface. On the left, a 'File' icon is connected to a 'Data Table' icon. The main window displays a table with 22 rows and 8 columns. The columns are: HeartDisease, Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, and RestingBP. The data is as follows:

	HeartDisease	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingBP
1	0	40	M	ATA	140	289	0	Normal
2	1	49	F	NAP	160	180	0	Normal
3	0	37	M	ATA	130	283	0	ST
4	1	48	F	ASY	138	214	0	Normal
5	0	54	M	NAP	150	195	0	Normal
6	0	39	M	NAP	120	339	0	Normal
7	0	45	F	ATA	130	237	0	Normal
8	0	54	M	ATA	110	208	0	Normal
9	1	37	M	ASY	140	207	0	Normal
10	0	48	F	ATA	120	284	0	Normal
11	0	37	F	NAP	130	211	0	Normal
12	1	58	M	ATA	136	164	0	ST
13	0	39	M	ATA	120	204	0	Normal
14	1	49	M	ASY	140	234	0	Normal
15	0	42	F	NAP	115	211	0	ST
16	0	54	F	ATA	120	273	0	Normal
17	1	38	M	ASY	110	196	0	Normal
18	0	43	F	ATA	120	201	0	Normal
19	1	60	M	ASY	100	248	0	Normal
20	0	36	M	ATA	120	267	0	Normal
21	0	43	F	TA	100	223	0	Normal
22	0	44	M	ATA	120	184	0	Normal

Figure 1.2

## 1. Data Exploration

Figure 2.2 When we uploaded our dataset to the orange tool, we did not have any lost data or missing data, but we still examined this fix part and put the necessary settings.

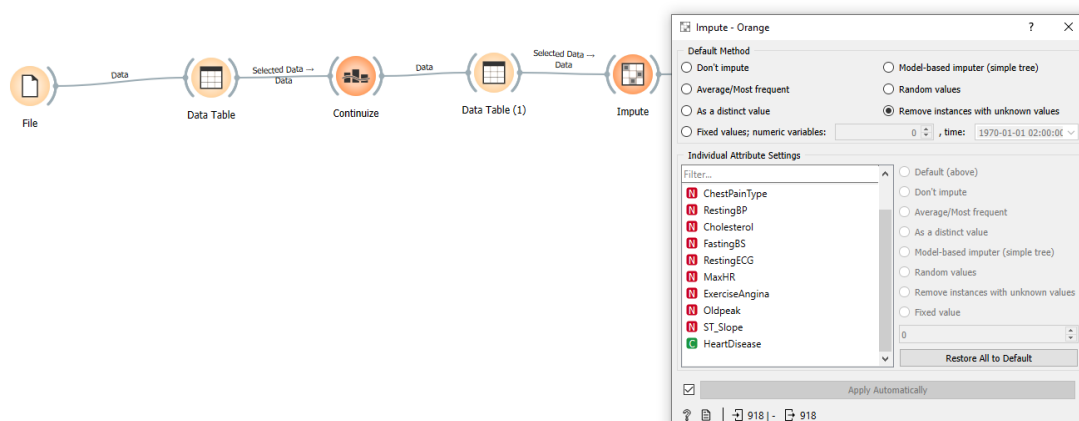


Figure 2.2

Figure 2.3 A scheme showing the final state of our table. We don't have any changes because we didn't have any missing data before.

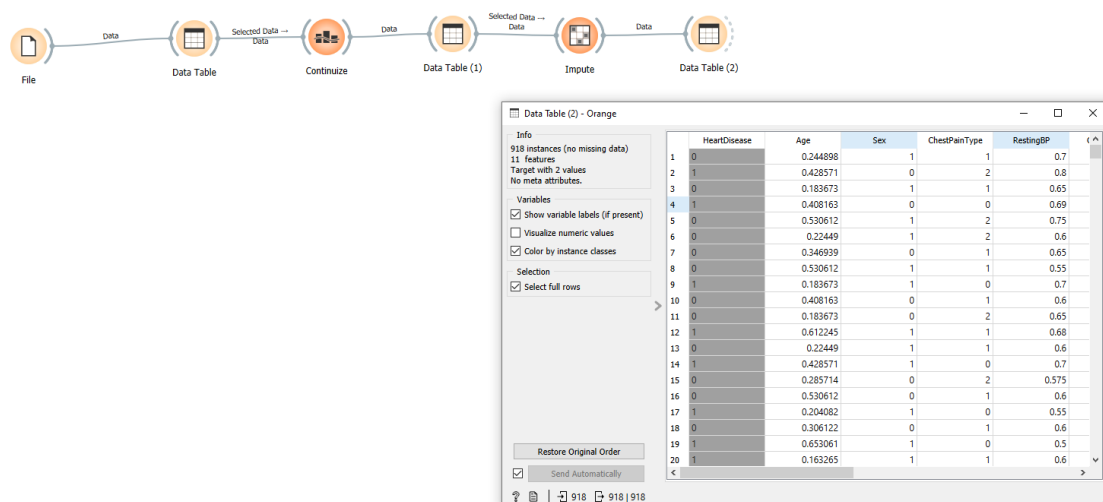


Figure 2.3

## 1.1 Scatter Plot

Scatter plots depict the relationship between two variables within a data set. It represents data points on a Cartesian system or a two-dimensional plane. The independent variable or attribute is depicted along the X-axis, and the dependent variable along the Y-axis. These diagrams are frequently referred to as scatter graphs or scatter diagrams. [2]

Figure 2.4 We created a scatter plot and analyzed age and MaxHR over it. The red dots we see here represent heart failure, while the blue dots represent the absence of heart failure. When we examine the resulting image, the maxhr value decreases as the age increases, therefore, the heart disease data also increases as the age increases.

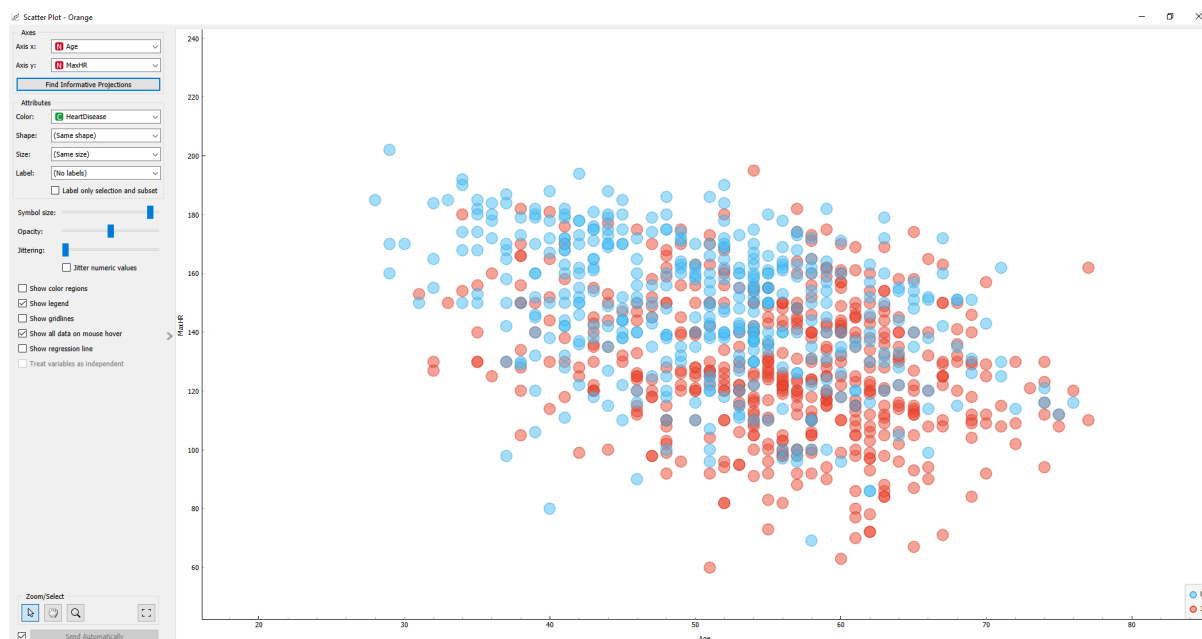


Figure 2.4

Figure 2.5 Here, we examined the MaxHr feature together with blood pressure, namely RestingBP. In order to better understand this, we need to know the RestingBP value range of 120-140, which is considered normal or healthy, and the MaxHr value range of 140 and above. Starting from here, there is a high blue dot at the intersection of these two values. In fact, some blood pressure values are so common in many patients that we can see them as a line from afar.

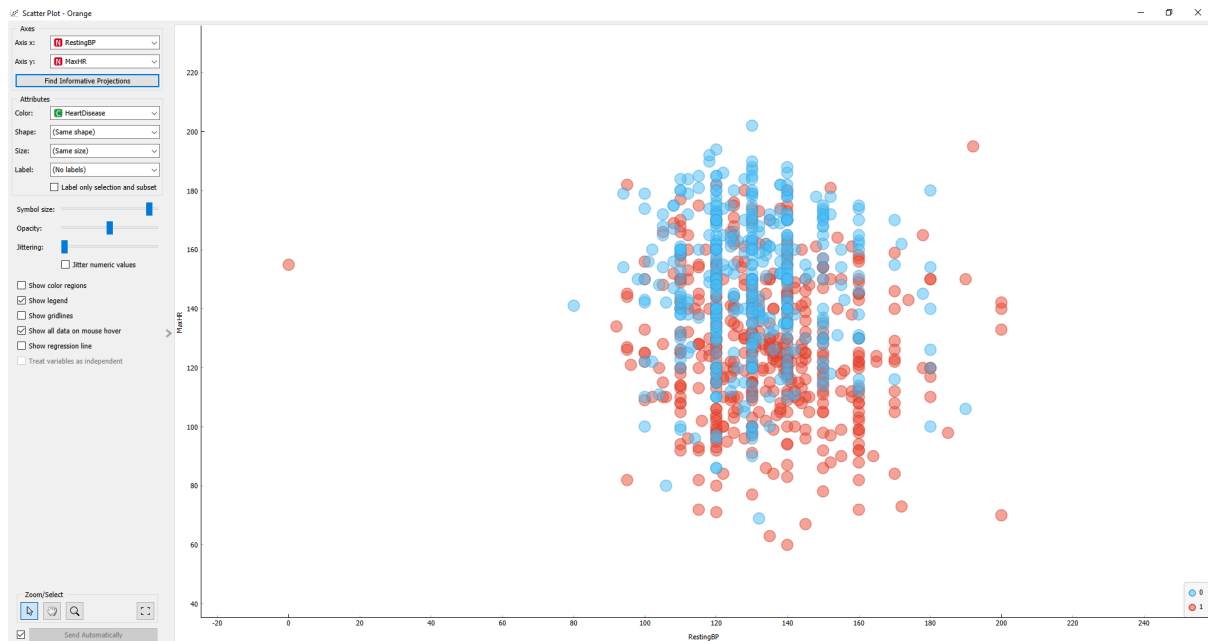


Figure 2.5

## 1.2 Histogram

A histogram is a graph used to represent the frequency distribution of several data elements for one variable. For continuous attributes, the attribute values are also presented in the form of a histogram. Different forms of distribution tables could be derived from the orange program. Although it is recommended to use relative frequency rather than absolute frequency when comparing two data sets' distributions. [4]

Figure 2.6 Here we are actually examining a painting that every person can know. The increase in diseases that can occur in humans as the age increases. Since heart failure is also a basic disease, we see that the data included in this group dominate the table after a certain age. On the other hand, it is seen that

the data without heart failure problems and represented in blue is more in the younger age group.

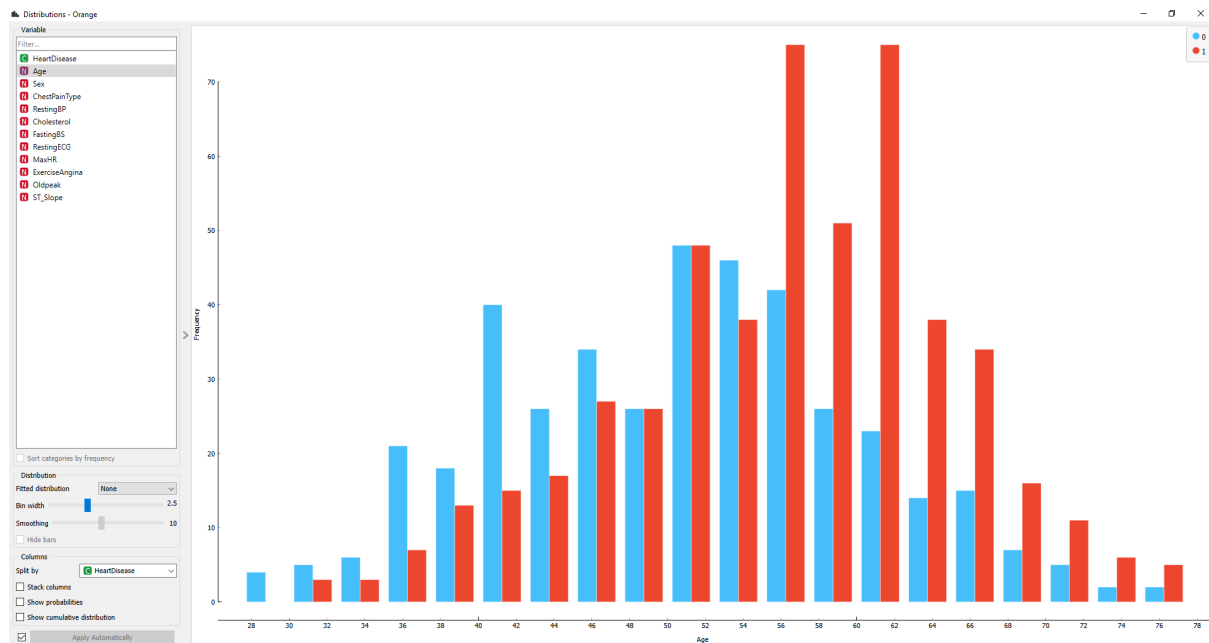


Figure 2.6

Figure 2.7 In fact, the information we gave while examining the scatter plot part is also verified here. The MaxHr value, which is considered healthy, is 140 and above. As we can see in this table, when MaxHr is 140 and above, we can see that the Blue, that is, there is no heart failure problem, increases.

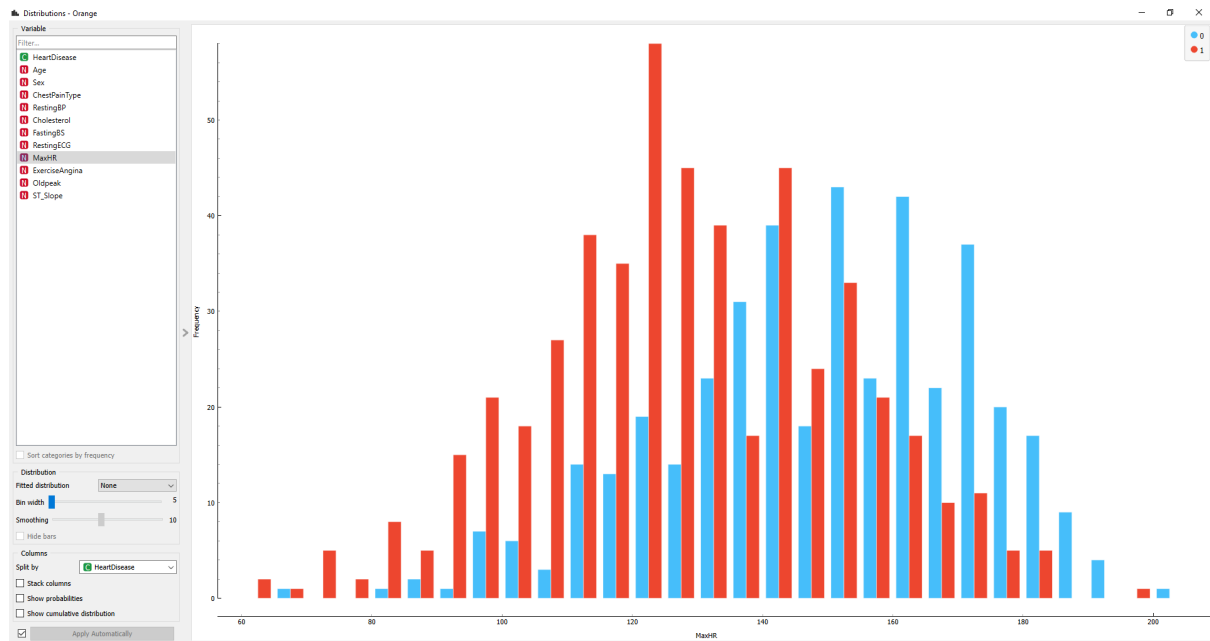


Figure 2.7

Figure 2.8 Here, we have a feature statistics table where we can see the distribution of all features together. By separating our features with two different values with two columns, we can examine whether they appear in large groups.



Figure 2.8

### 1.3 Distributions:

Distributions displays the value distribution of discrete and continuous attributes. Distributions may be conditional on the class if the data contains a class variable. [4]

Figure 2.9 People belonging to the heart failure class mean 175.94 and other classes have a value of 227.12 . They have standard deviations of 126.27 and



74.54 degrees. For a distribution to be perfect, these values must be the same, but we can't see it here.

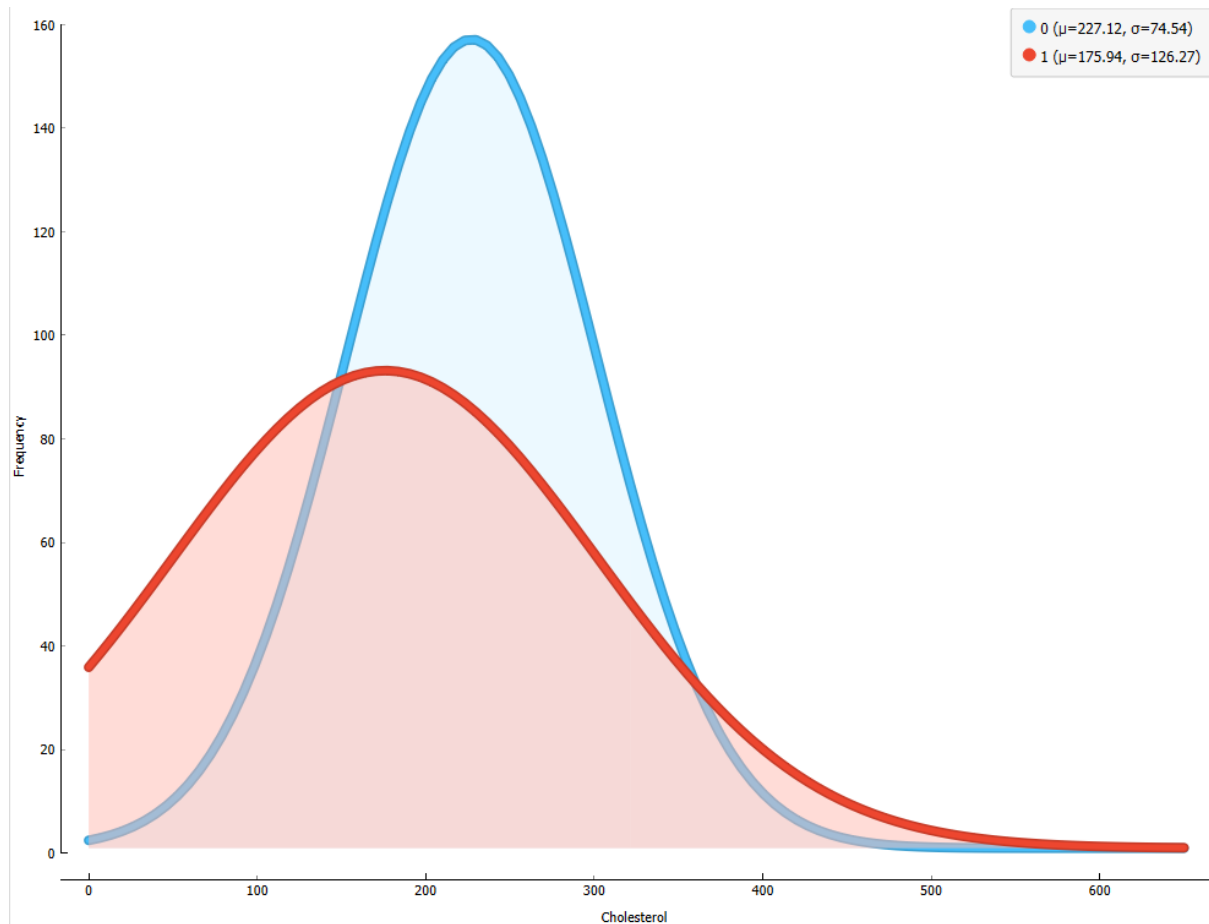


Figure 2.9

Figure 2.10 Here we see that our standard deviation values are almost the same. We also see that where the MaxHr value is 140, we see the cross-section of two classes, the decrease in people with heart failure and the increase in people who are considered healthy because, as we mentioned above, the MaxHr healthy band gap starts from 140.

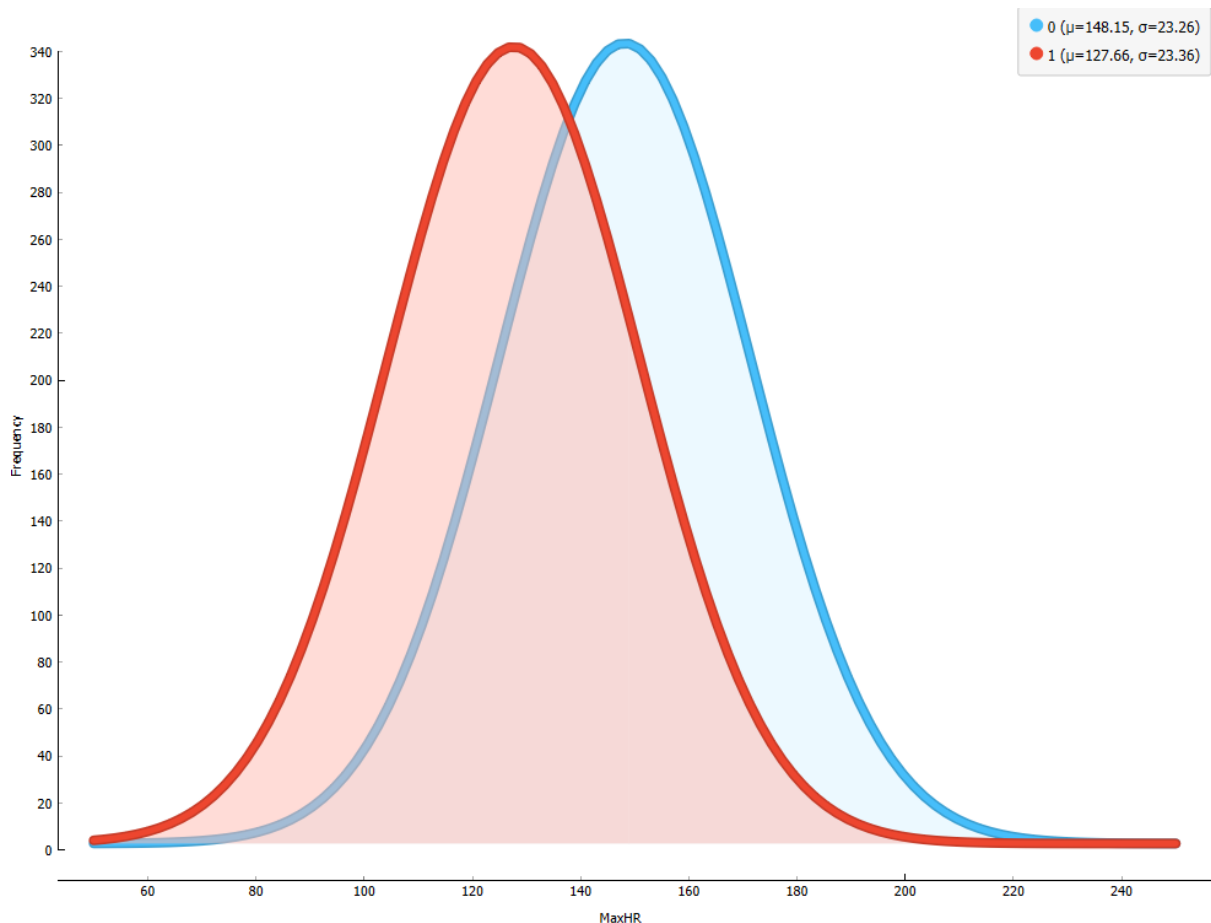


Figure 2.10

## 2. Unsupervised Learning

### 2.1 Hierarchical clustering:

Hierarchical clustering is another unsupervised learning algorithm used to arrange unlabeled data points with similar attributes. We begin by treating each data point as its own cluster. Then, we join clusters together that have the shortest distance between them to create larger clusters. This step is repeated until one large cluster is formed containing all of the data points. [3]

Figure 3.1 Here is a table where we can examine the binary clustering example. Of course we can change this number of clusters by playing the cutting line. Here, with the cutting line reaching its limit, we can see two large clusters. We can also see that the shortest distance between two clusters is used as a connection. Also, the vertical lines in the dendrogram help us define the distance between the clusters.

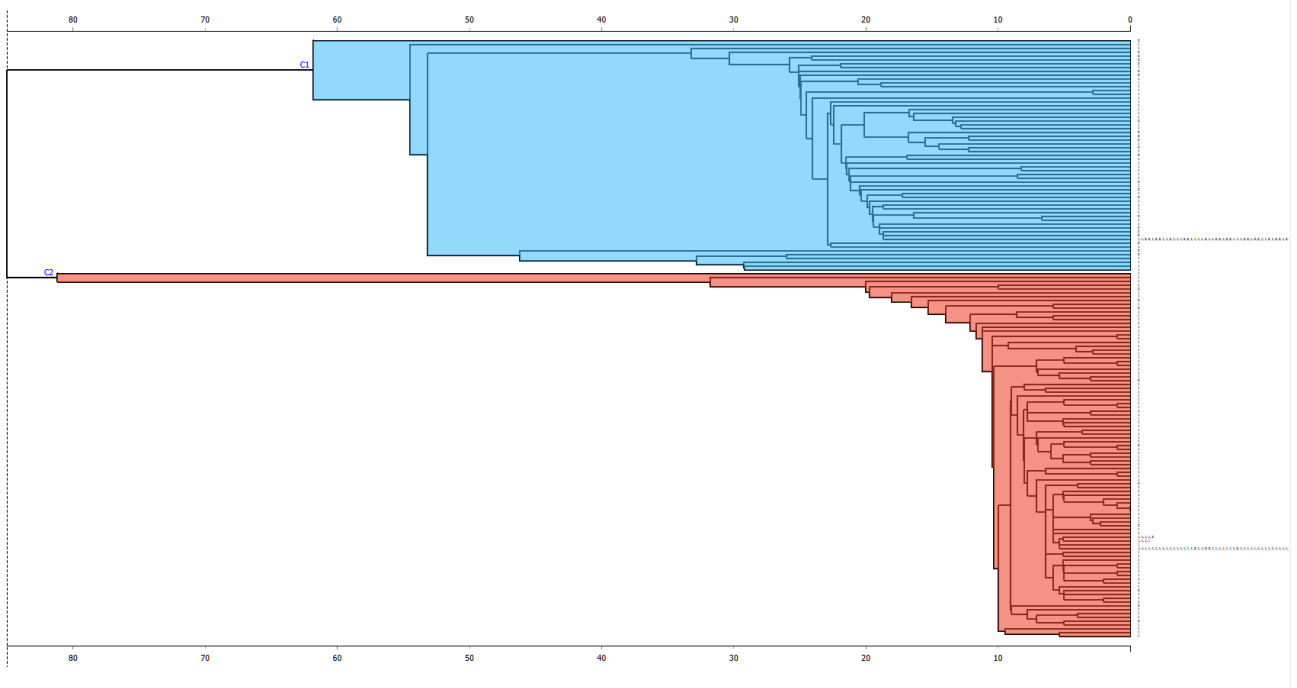


Figure 3.2

Figure 3.3 There is another experiment with linking to the mean where the maximum depth is set to 10 and the distance between two pairs in each cluster is added up and divided by the number of pairs. When the cut is placed at 5, we get 6 clusters.

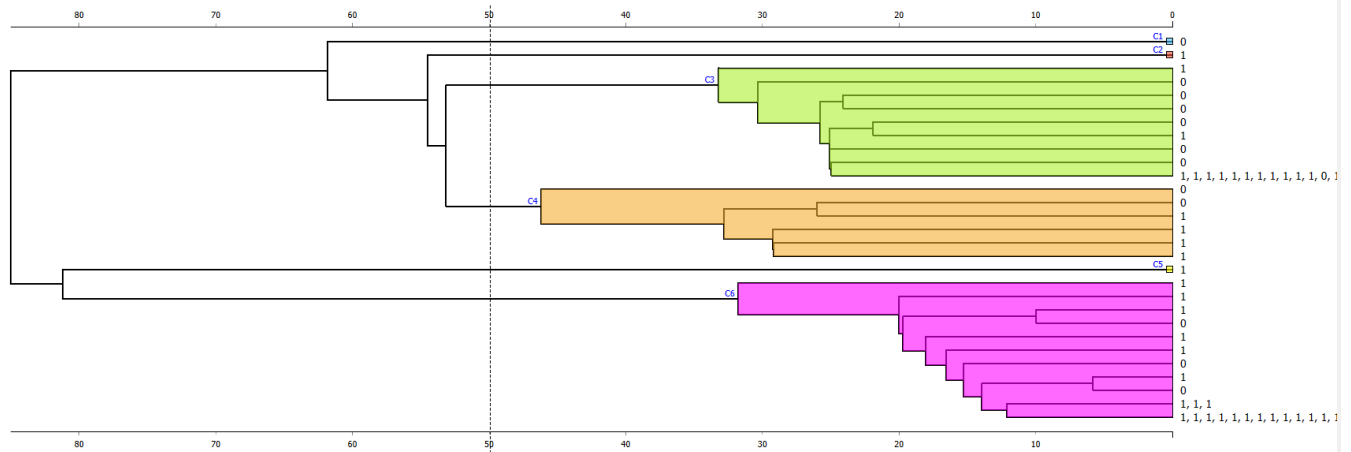


Figure 3.3

Figure 3.4 Demonstrates another dendrogram with different hyperparameter, average linkage method. the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Using the same max depth gives more clusters in this method on the same cut off.

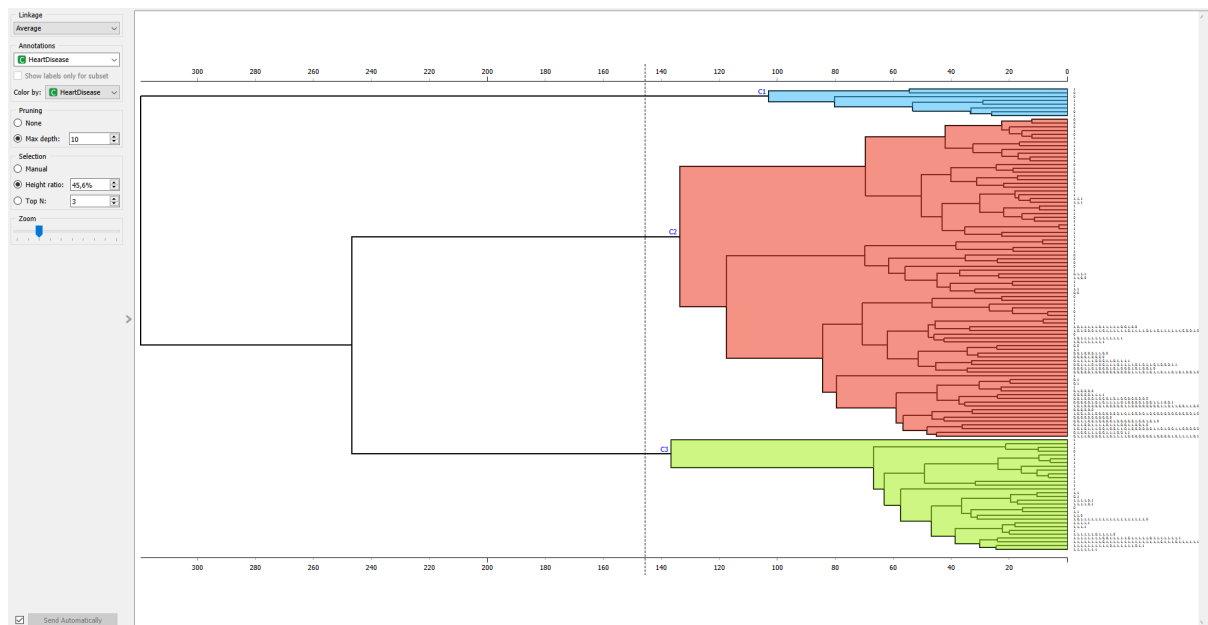


Figure 3.4

## 2.2 K-Means clustering:

k-means is a rapid iterative algorithm that has been implemented in a variety of clustering applications. It is a point-based clustering procedure in which the cluster centers are initially positioned arbitrarily and are subsequently relocated at each stage to reduce clustering error.

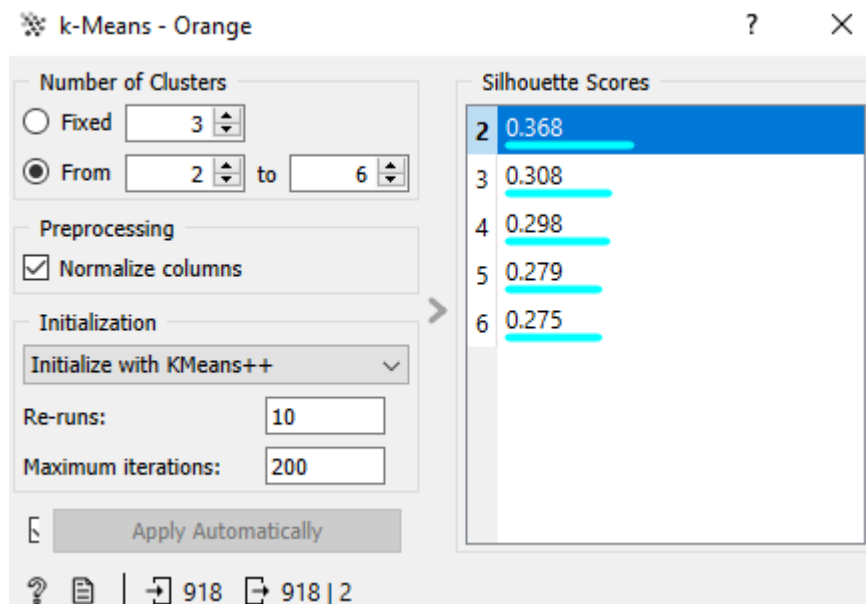


Figure 3.5

We have a dataset with two classes. We will learn the clustering and their values with the k-means algorithm. the results we got are the highest value of 0.368 also we can see in figure 3.5 . Then we will see this value in the graph. figure 3.6 Since the graph is too long, I present the image of the change point. We can see that those who do not have heart health problems score higher.

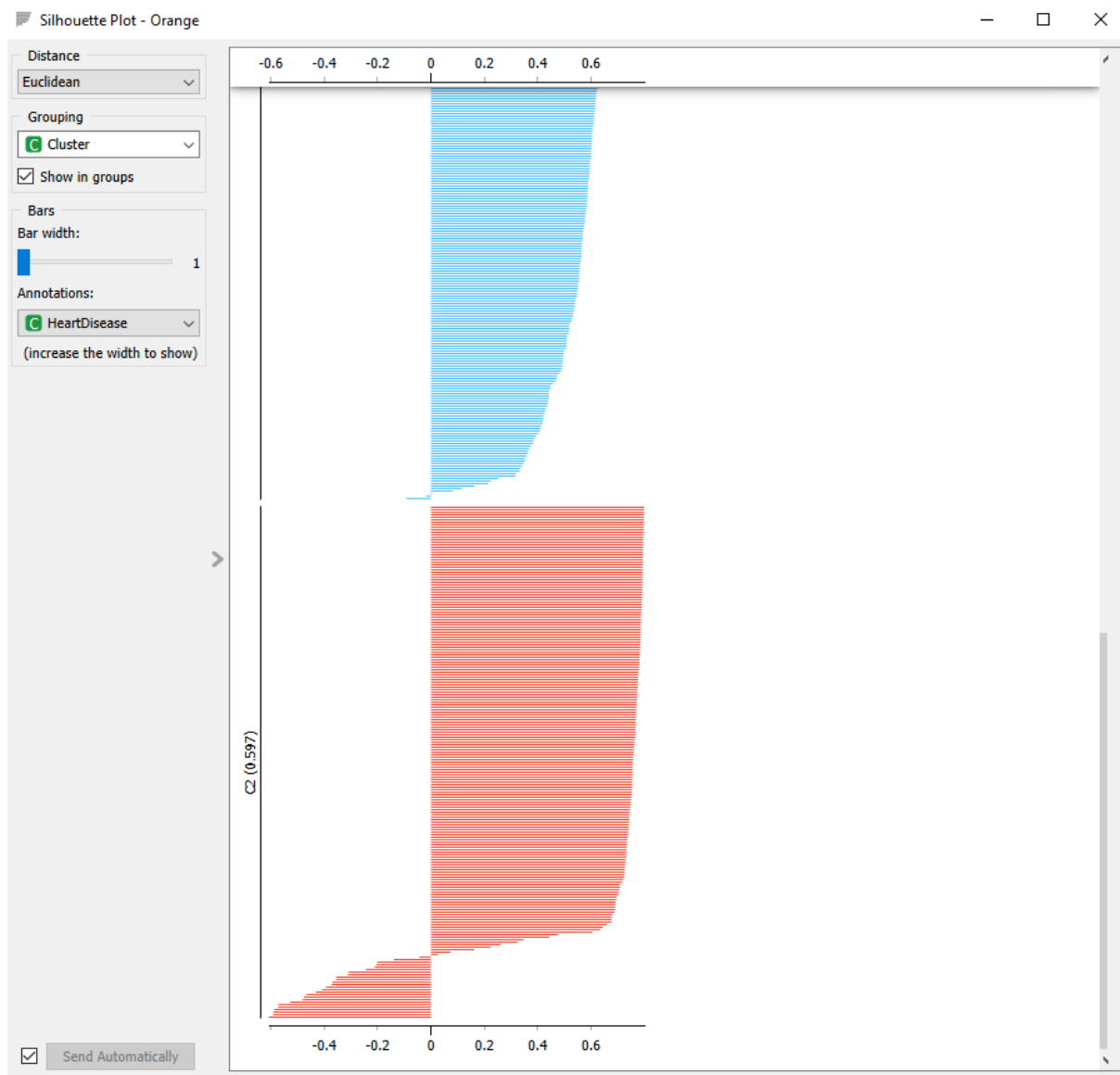


Figure 3.6

Figure 3.7 Here, while leaving the K-means algorithm settings the same, we see the graph that appears when we set the grouping as our target and we can see that all the values are a little closer to the zero point.

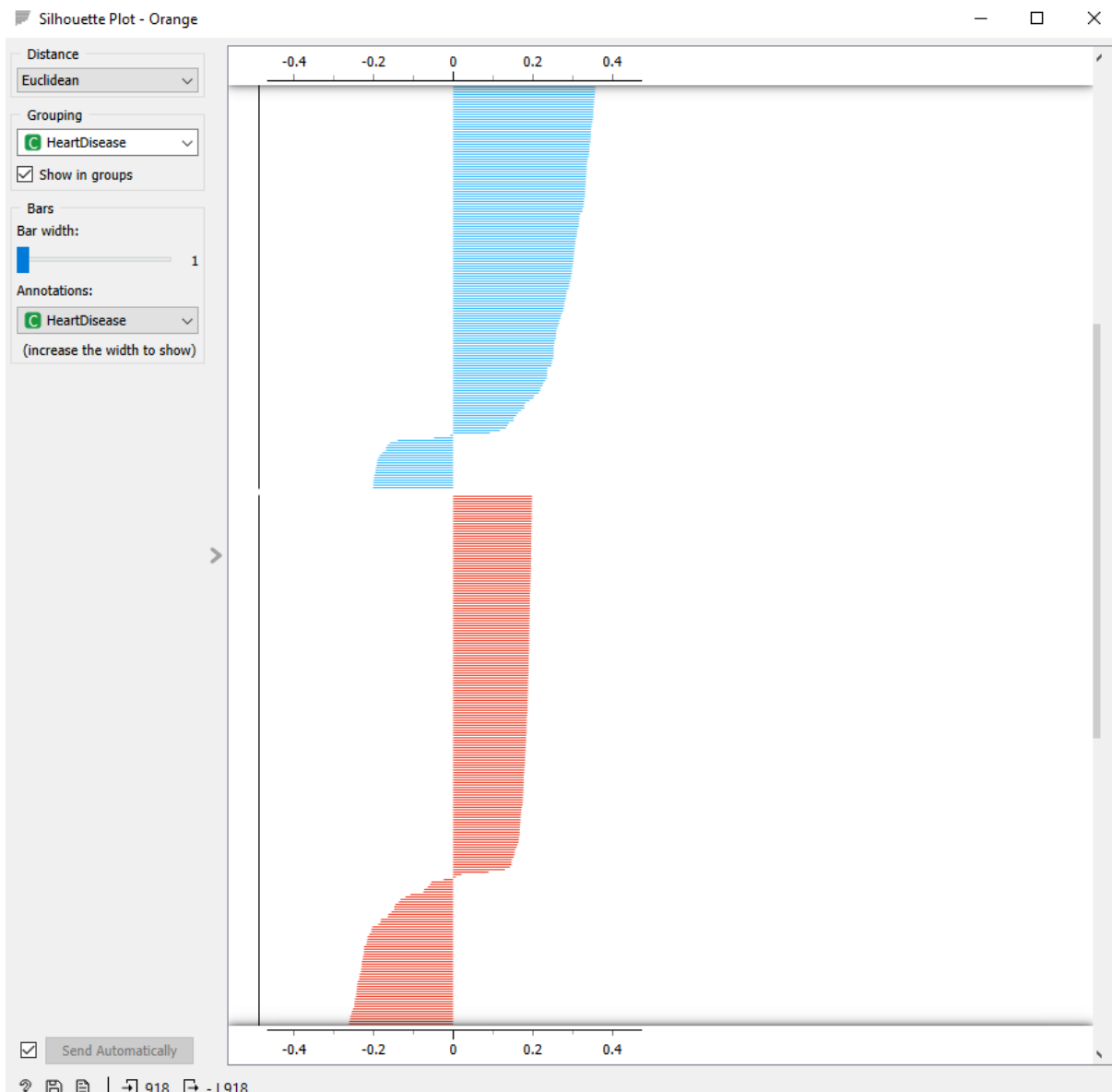
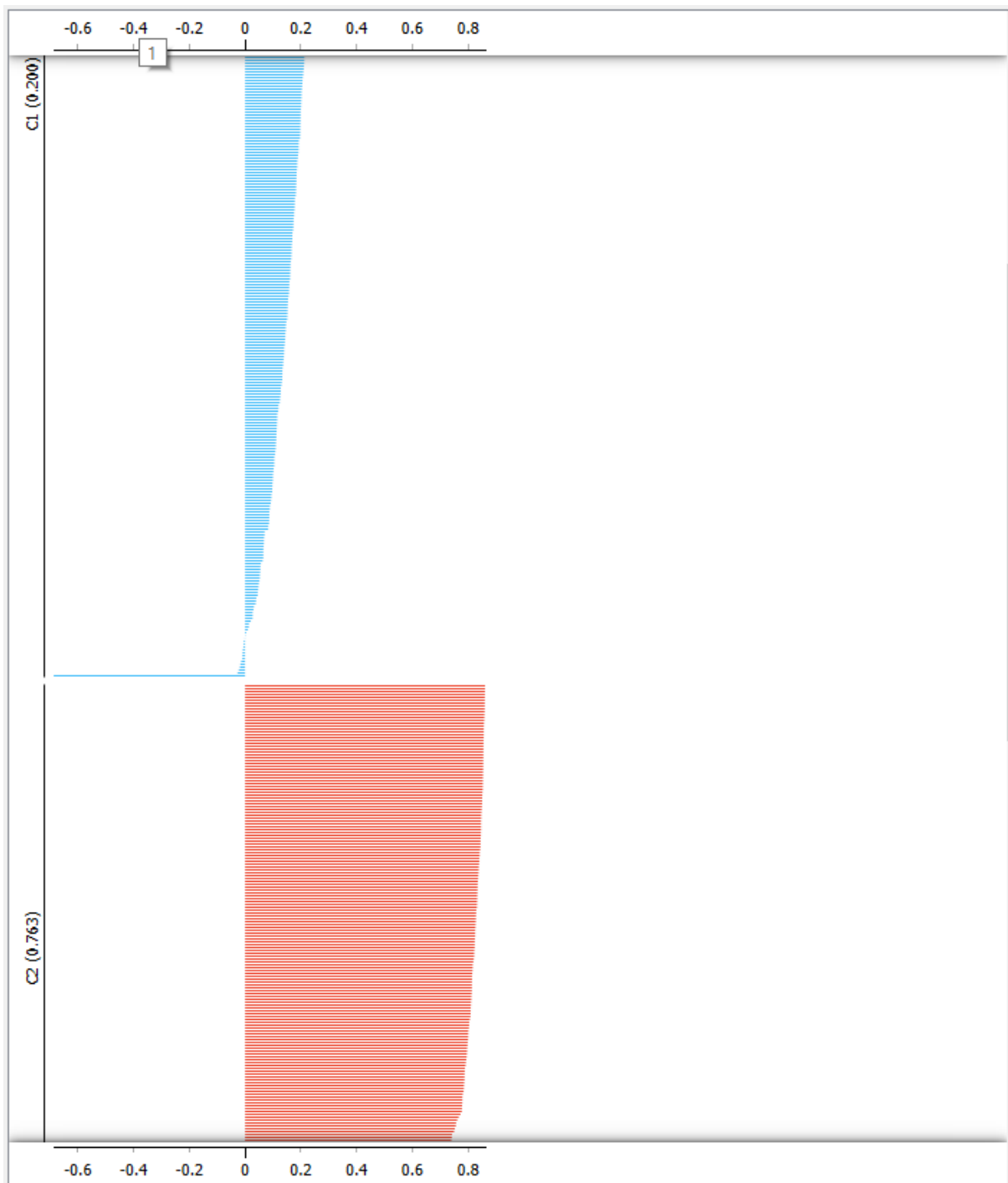


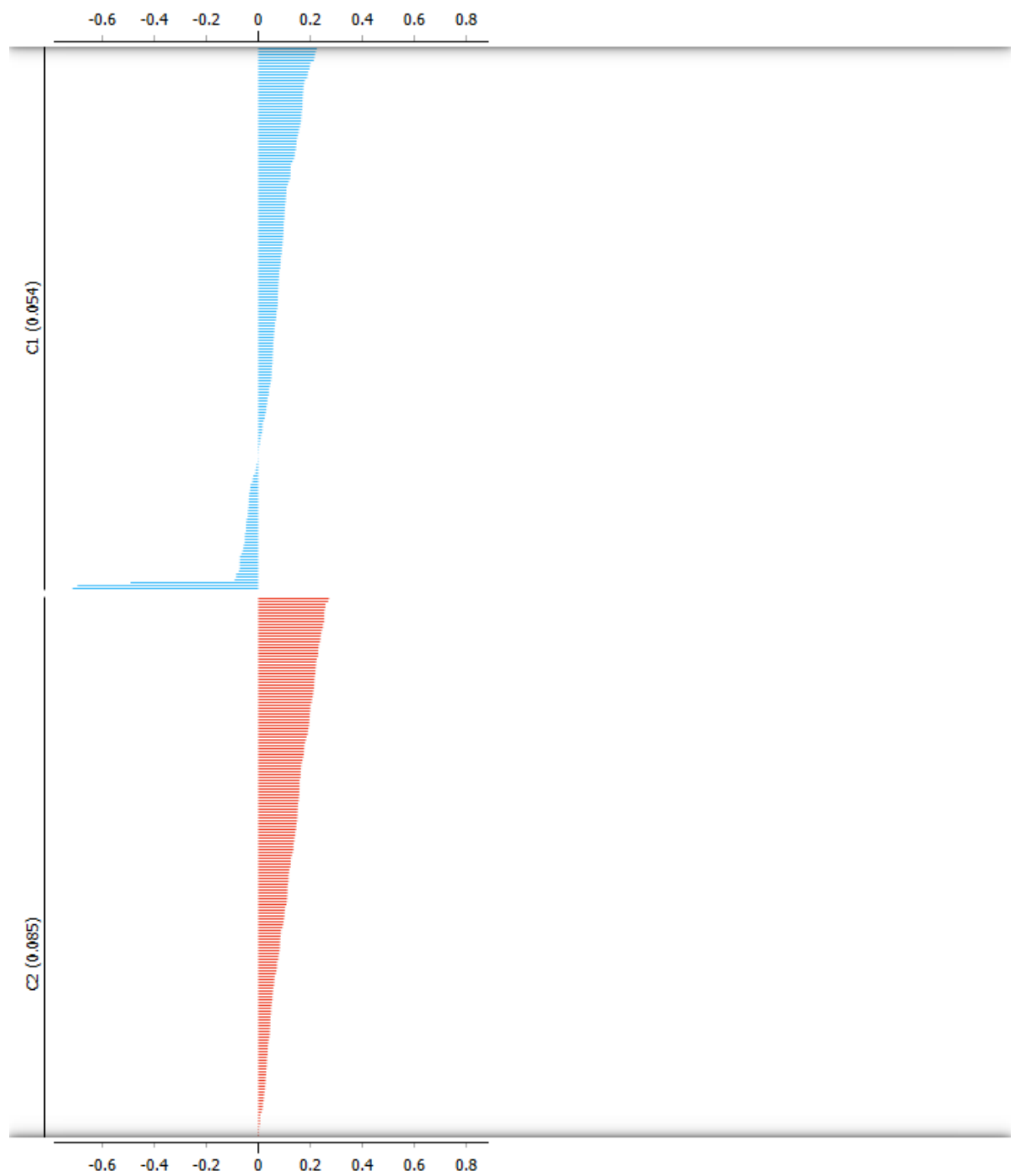
Figure 3.7

We have a value of  $k$  and we examined it. Next we will examine four more different values. We also have the results from two different algorithm features from the same  $k$  value

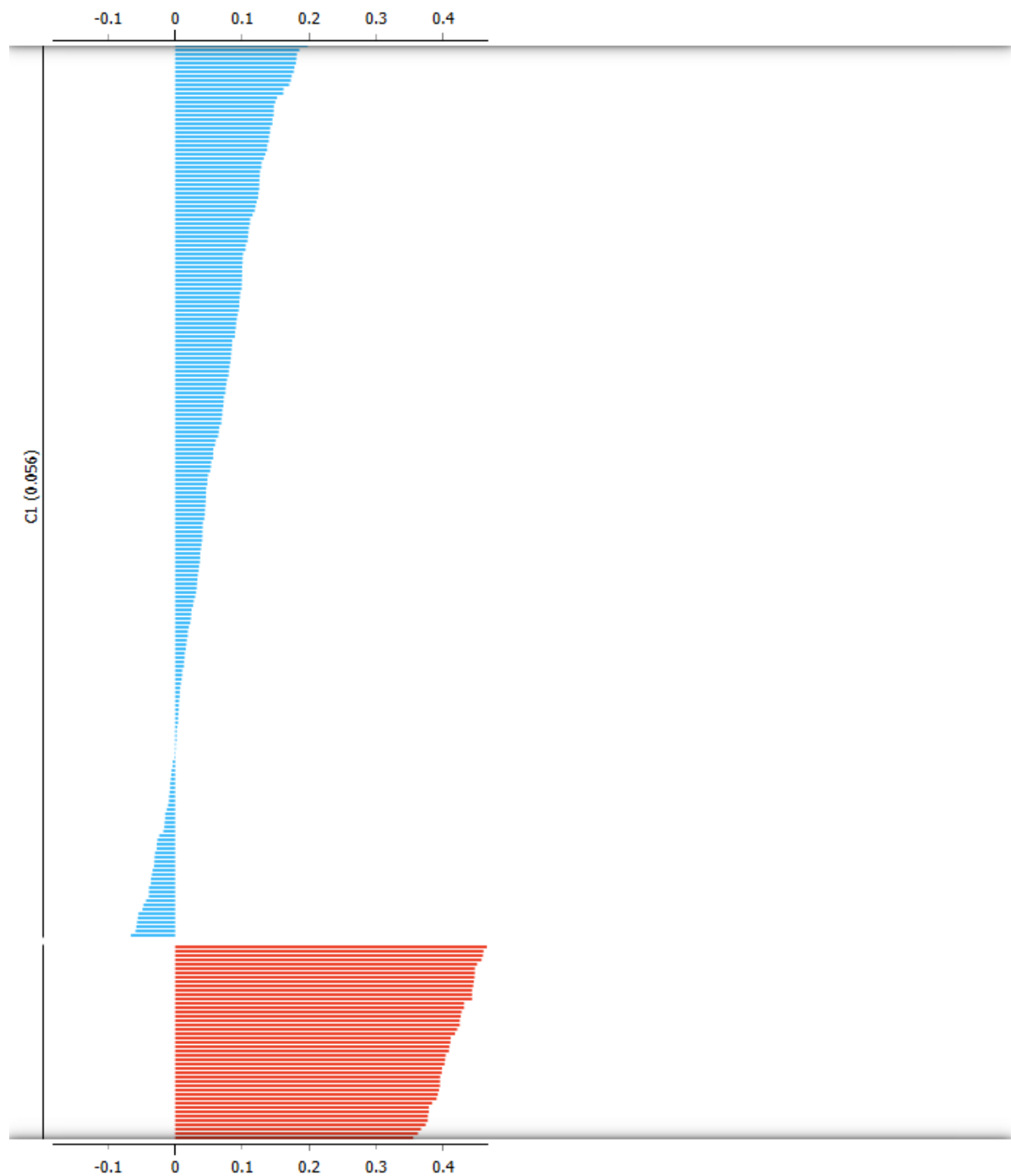




k values is 3



k value is 4



k value is 5

To sum up, the first thing I can say when we examine the data we have is that we did not see a perfect separation. The highest of our values is 0.368. This value is far from one and close to zero, and most of our values are also close to

zero, but since we have predominantly positive values, the assignment process to the clusters is not satisfactory, but it shows that the correct assignments are made at a high rate.

## **3. Supervised Learning**

### **3.1 Neural Network:**

Artificial Neural Networks (ANNs) are distributed, adaptive, generally nonlinear learning machines built from many different processing elements (Processing Elements/PEs). Each PE receives connections from other PEs and/or itself. The interconnectivity defines the topology. The PEs sum all these contributions and produce an output that is a nonlinear (static) function of the sum. The PEs' outputs become either system outputs or are sent to the same or other PEs. [4]

Each algorithm has its own hyperparameters. Our Neural Networks also have hyperparameters such as activation function, learning rate, hidden layers, momentum. By changing these hyperparameters values, we can achieve different results. We will try to get efficient results by changing the first hidden layers and then the activation function hyperparameters.

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.705	0.513	0.348	0.263	0.513

Hidden layer: 5, 140, 65

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.742	0.658	0.649	0.672	0.658

Hidden layer: 100

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.715	0.513	0.348	0.263	0.513

Hidden layer: 80, 20, 70

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.743	0.655	0.637	0.683	0.655

Hidden layer: 9

When I give different values, I realized that when I give only a single number or number to the hidden layer, I get higher efficiency. Apart from that, the ones I tried by giving three different numbers are the efficient results I got by increasing and decreasing the numbers in order.

Now I will do the same experiments by changing the activation function without changing the value of the hidden layer.

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.742	0.658	0.649	0.672	0.658

Activation: Logistic

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.738	0.665	0.660	0.672	0.665

Activation: Identity

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.738	0.673	0.667	0.681	0.673

Activation: tanh

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.754	0.676	0.673	0.680	0.676

Activation: ReLu

ReLu gave the best results when I changed the activation without changing the value I got with 100 in the hidden layer.

### 3.2 K-Nearest Neighbor:

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier that employs proximity to make predictions or classifications regarding the aggregation of a single data point. Although it can be used for both regression and classification problems, it is typically employed as a classification algorithm, based on the premise that similar points are typically found in close proximity. [5]

There are also some hyperparameters in our kNN algorithm. There are number of neighbors and metric, which we will consider. We will try to get the most efficient results by trying both in turn.

Model	AUC	CA	F1	Precision	Recall
kNN	0.746	0.673	0.669	0.678	0.673

K = 33

Model	AUC	CA	F1	Precision	Recall
kNN	0.707	0.647	0.644	0.650	0.647

K = 5

Model	AUC	CA	F1	Precision	Recall
kNN	0.719	0.669	0.669	0.669	0.669

K = 10

Model	AUC	CA	F1	Precision	Recall
kNN	0.740	0.669	0.664	0.675	0.669

K = 15

As our k value increased, I saw that we got more efficient results, and when I searched for the closest result to 1, the most efficient result was 33. I saw that after the value of 10, it took the same height at the value of 33

Now we will examine it by changing the metric value without changing the k value and keeping the k value constant, which gives the highest efficiency.

Model	AUC	CA	F1	Precision	Recall
kNN	0.746	0.673	0.669	0.678	0.673

Metric: Manhattan

Model	AUC	CA	F1	Precision	Recall
kNN	0.734	0.658	0.654	0.662	0.658

Metric: Euclidean

Model	AUC	CA	F1	Precision	Recall
kNN	0.740	0.662	0.657	0.668	0.662

Metric: Chebyshev



Model	AUC	CA	F1	Precision	Recall
kNN	0.742	0.658	0.651	0.668	0.658

Metric: Mahalanobis

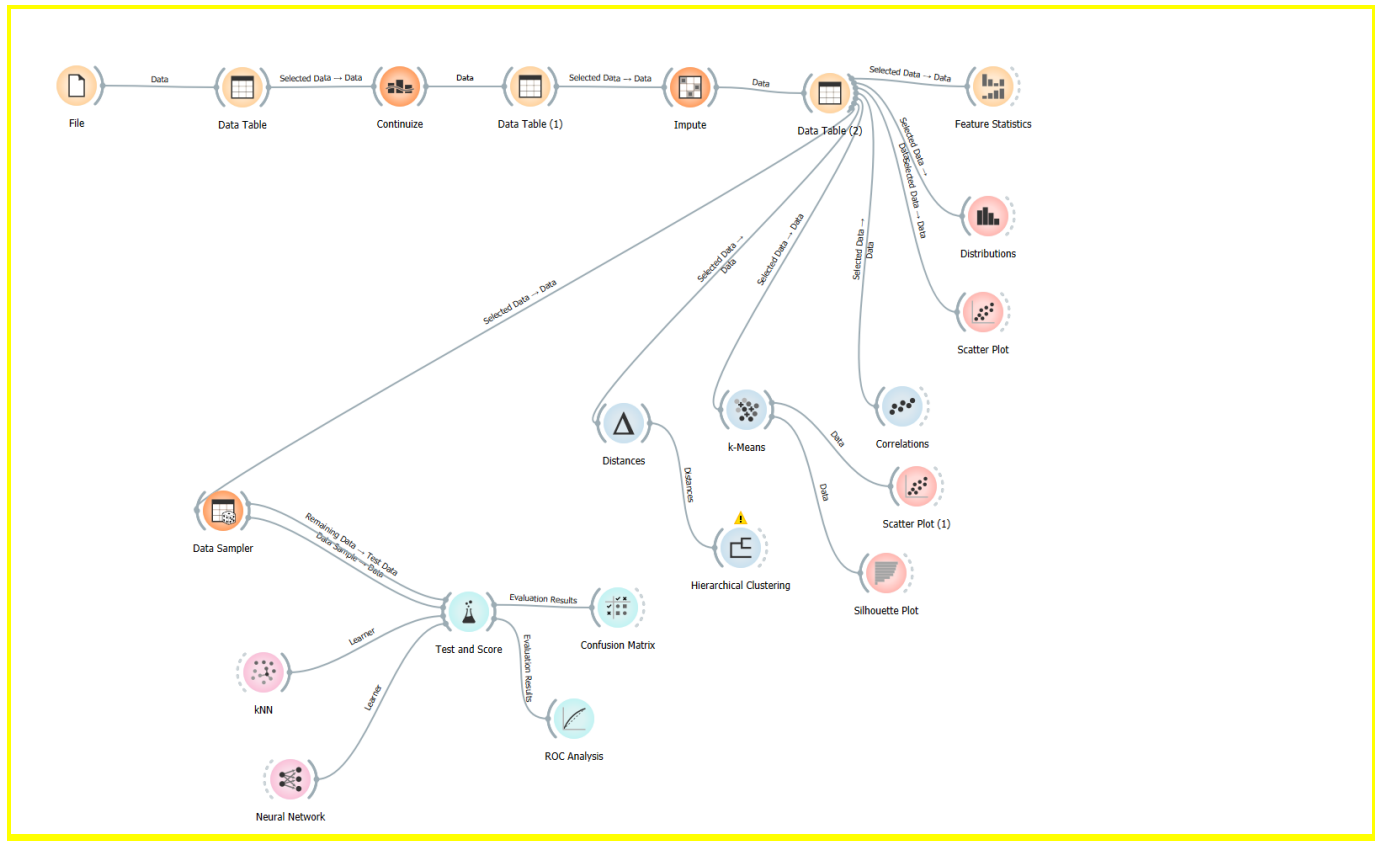
Based on the data we received by giving different metric values, we can say that Manhattan gives the best value.

Model	AUC	CA	F1	Precision	Recall
kNN	0.746	0.673	0.669	0.678	0.673
Neural Network	0.768	0.673	0.670	0.675	0.673

Overall

To make a short summary, it cannot be said that we got very excellent values, but we can see that we did not get bad values. These are the highest values, while our F1 value is above the average with 0.678, our AUC value goes up to 0.768.

# Workflow



## REFERENCE:

1. [Heart Failure Prediction Dataset | Kaggle](#)
2. [\*\*Scatter Plot | Definition, Graph, Uses, Examples and Correlation\*\*](#)
3. [Machine Learning - Hierarchical Clustering](#)
4. [Distributions — Orange Visual Programming 3 documentation](#)
5. [The global k-means clustering algorithm](#)
6. [What is the k-nearest neighbors algorithm? | IBM](#)