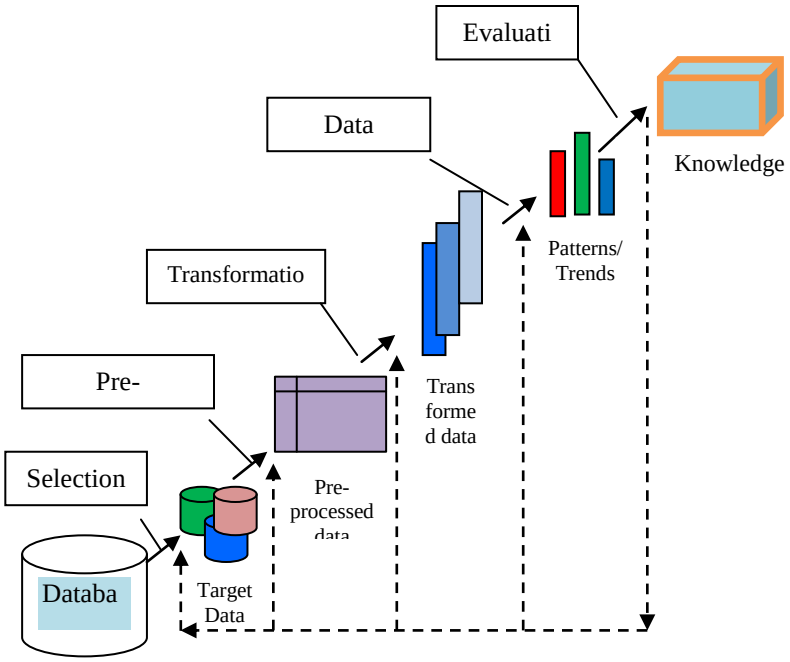
P1. DATA MINING TECHNIQUE FOR RETAIL INDUSTRY

by Sameer

Note: *This is a summary of detailed project. Only the Solution and Results are shown.*

DESCRIPTION OF SOLUTION

This study is performed in data mining tool 'Orange'. It is an open source data analysis and visualization tool, which performs data mining through Python scripting and visual programming. The work is performed in two steps. First is data mining and second is machine learning. In this study data mining techniques such as K-means clustering and classification are used in extracting knowledge from database.



1.1. KDD process for data mining in this work

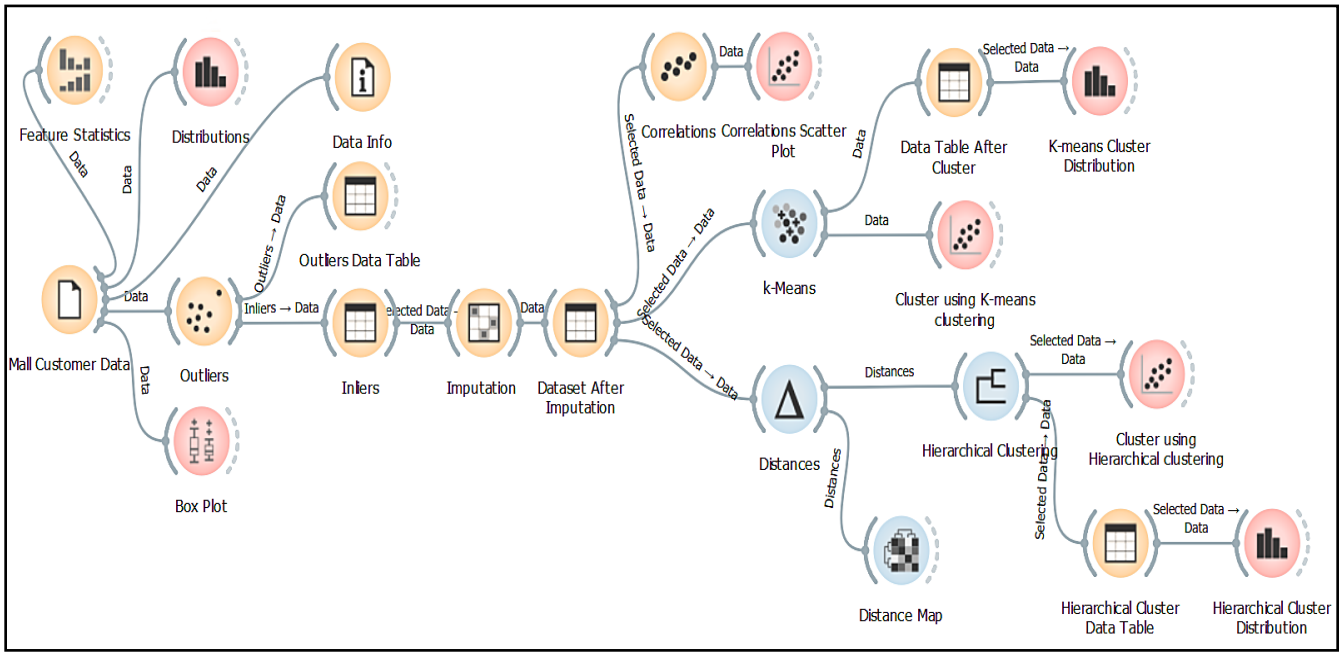
The data mining task consist of the KDD process such as regression, classification, and clustering, etc. In the second step, the same dataset is used to predict the presence of heart disease, using different classification algorithms.

Steps in KDD process:

1. **Data Selection:** It is defined as the step where relevant data is retrieved from database and used to analysis. A target variable must be focus here.

2. **Data Pre-processing:** Data cleaning and preprocessing comprises removing noise or outliers in order to acquire the information needed to model. Also we decide techniques for handling missing data values.
3. **Data Transformation:** In this step data dimension reduction or data transformation takes place. In this step data is transformed into appropriate form required by mining procedure like regression or clustering.
4. **Data Mining:** It is defined as the techniques that are applied to extract hidden patterns and trends potentially useful.
5. **Pattern / Trends Evaluation:** This step is defined as identifying increasing patterns representing knowledge based on given measures. This can be done by finding score of each pattern, visualization, and report summary.
6. **Knowledge Representation:** This is defined as methods or a technique which utilizes visualization tools to represent data mining outcomes.

The proposed methodology of this study comprises of following steps as data collection, data understanding, data pre-processing, feature transformation and data mining.



1.2. Flowchart of the proposed study (Orange Data Mining Tool)

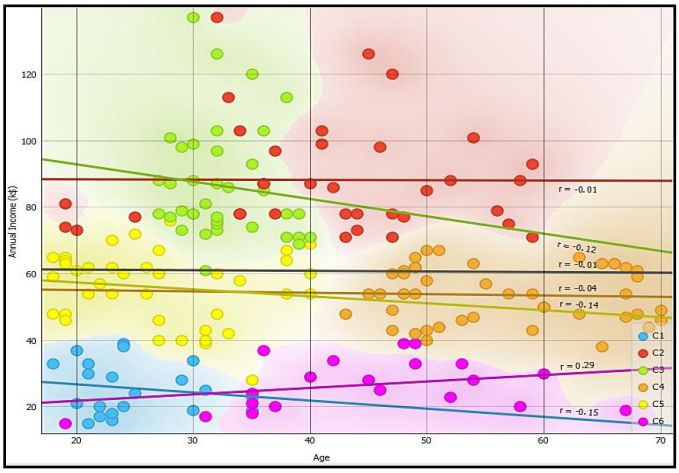
RESULTS

To find the number of clusters I have used silhouette coefficient score. Out of the 8 scores, the 6th number gives the higher silhouette score. So, K= 6 will be the optimal value to find clustering.

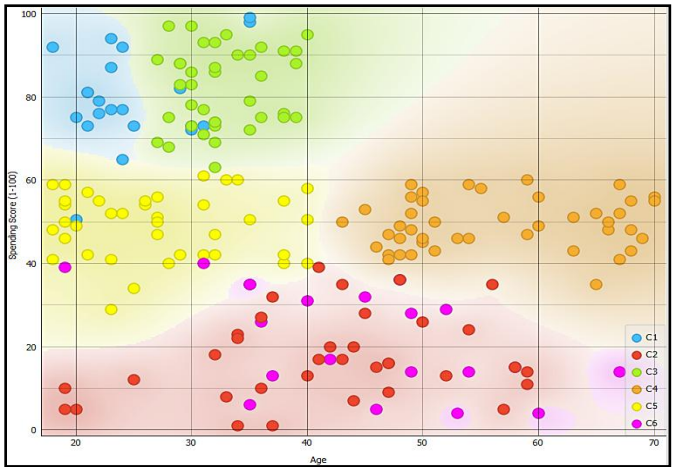
Silhouette Scores	
2	0.298
3	0.389
4	0.406
5	0.430
6	0.448
7	0.434
8	0.420

1.3. Silhouette Score for K number of cluster

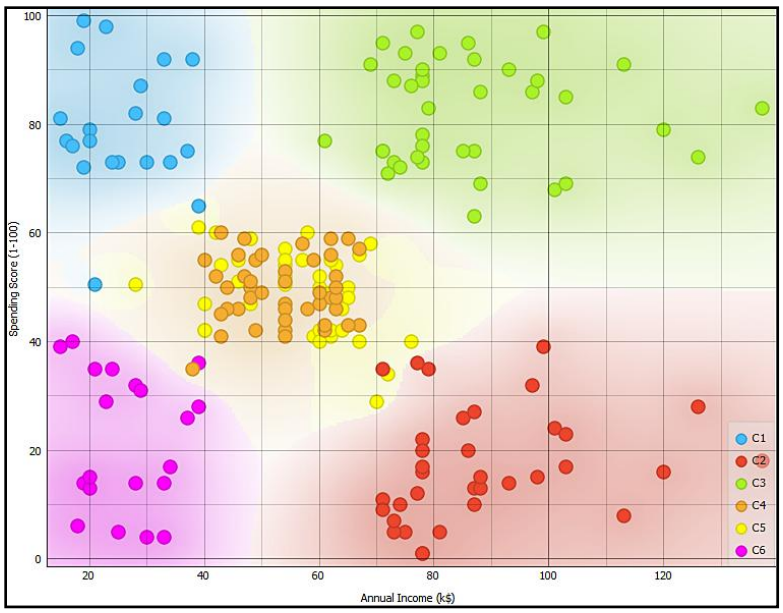
From K-means clustering:



1.4. K-means clustering between age and annual income

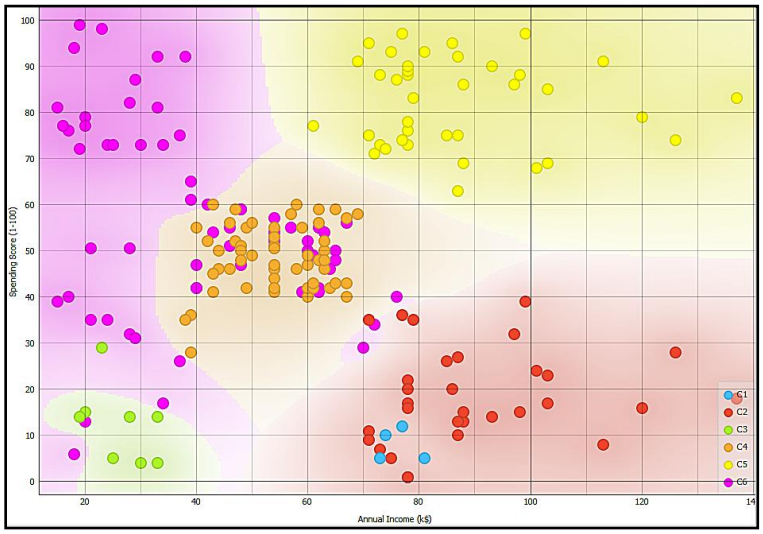


1.5. K-means clustering between age and spending score



1.6. K-means clustering between annual income and spending score

For comparative analysis between two clustering algorithms, I have visualized the hierarchical clustering as shown below.



1.7. Hierarchical clustering between annual income and spending score

DISCUSSION

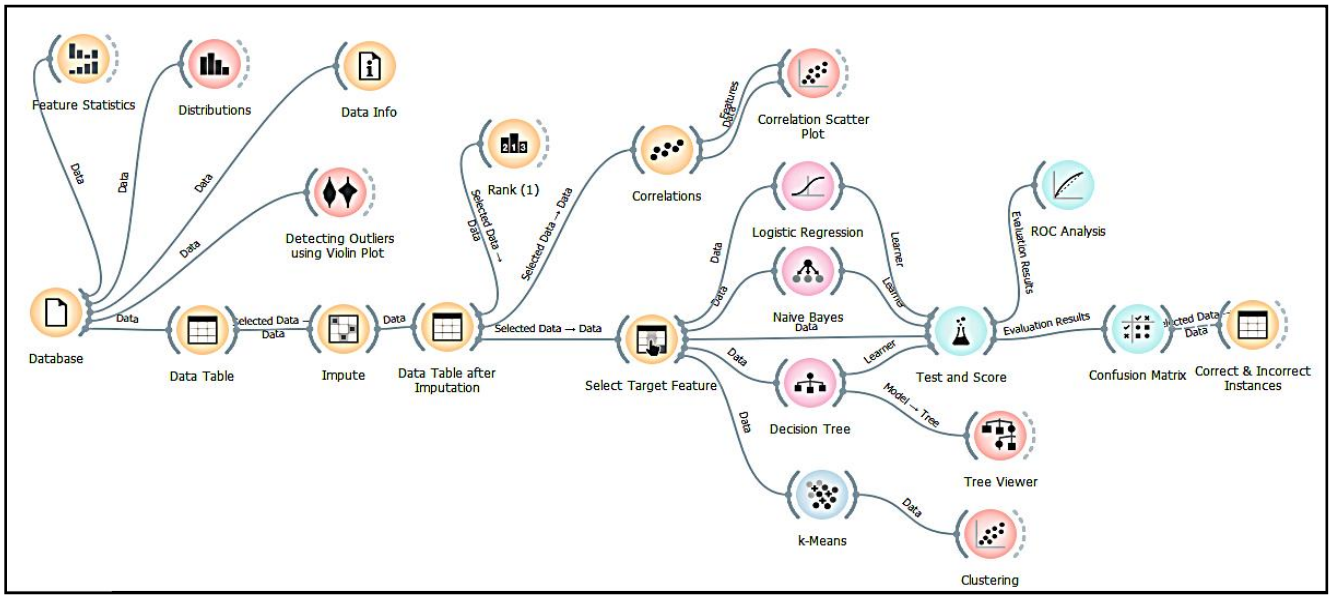
We can see that Fig 1.6 and Fig 1.7 having similar features but the cluster is more accurate and objects are more tightly bound in K-means clustering. The centroid in K-means clustering works well while the Euclidean distance metric is not able to measure the distance between two objects in hierarchical clustering.

P2. DATA MINING APPROACH FOR HEART DISEASE PREDICTION

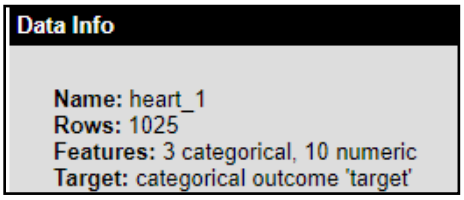
DESCRIPTION OF SOLUTION

Similar as applied in P1.

The proposed methodology of this study comprises of following steps as data collection, data understanding, data pre-processing, feature transformation and data mining.



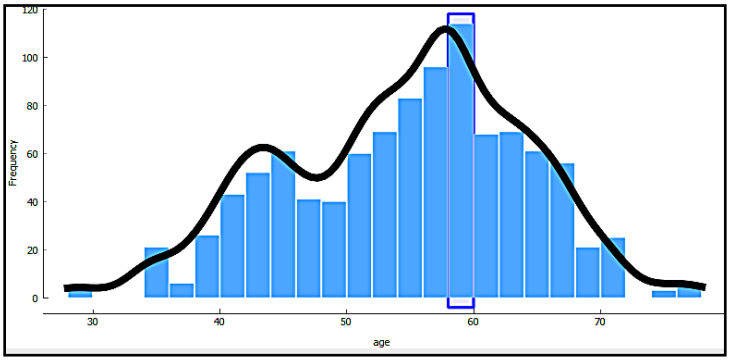
2.1 Flowchart of the proposed study (in Orange tool)



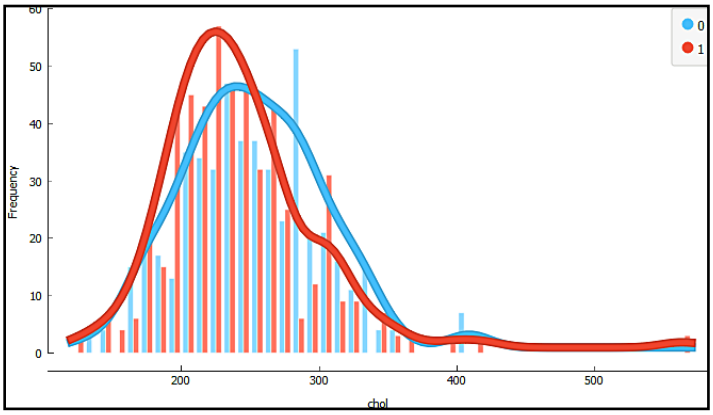
2.2 Data info of Heart Disease dataset

	Name	Type	Role	Values
1	age	numeric	feature	
2	sex	categorical	feature	0, 1
3	cp	numeric	feature	
4	trestbps	numeric	feature	
5	chol	numeric	feature	
6	fbs	categorical	feature	0, 1
7	restecg	numeric	feature	
8	thalach	numeric	feature	
9	exang	categorical	feature	0, 1
10	oldpeak	numeric	feature	
11	slope	numeric	feature	
12	ca	numeric	feature	
13	thal	numeric	feature	
14	**target**	**categorical**	**target**	**0, 1**

2.3 Snapshot of data type and role

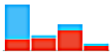
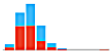


2.4 Distribution of age in dataset



2.5 Distribution of cholesterol in different category of patients

Feature Statistics:

	Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
N	age		54.43	56	0.17	29	77	0 (0%)
N	cp		0.94	1	1.09	0	3	0 (0%)
N	trestbps		131.61	130	0.13	94	200	0 (0%)
N	chol		246	240	0.21	126	564	0 (0%)
N	restecg		0.53	1	1.00	0	2	0 (0%)
N	thalach		149.11	152	0.15	71	202	0 (0%)
N	oldpeak		1.072	0.8	1.096	0.0	6.2	0 (0%)
N	slope		1.39	1	0.45	0	2	0 (0%)
N	ca		0.70	0	1.35	0	3	18 (1%)
N	thal		2.34	2	0.25	1	3	7 (0%)
C	sex			1	0.615			0 (0%)
C	fbs			0	0.421			0 (0%)
C	exang			0	0.639			0 (0%)
C	target			1	0.693			0 (0%)

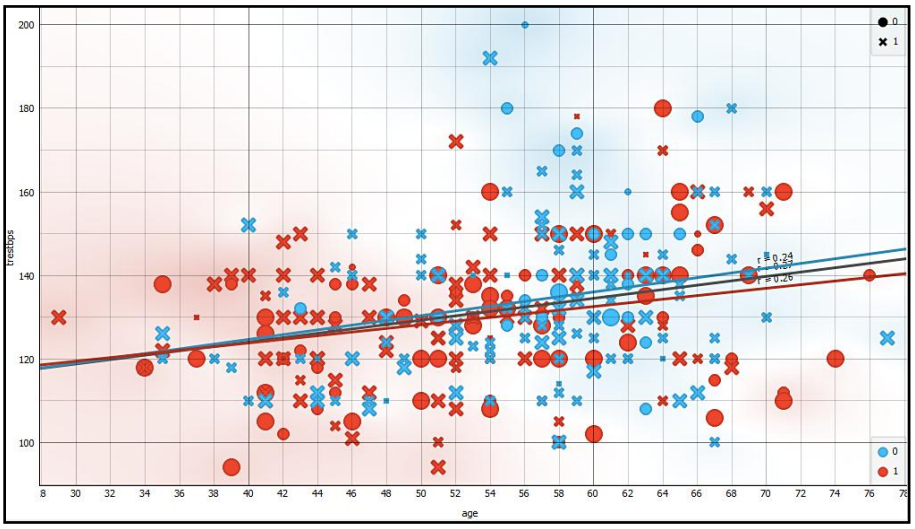
2.6 Feature statistics of the dataset

Correlation between features:

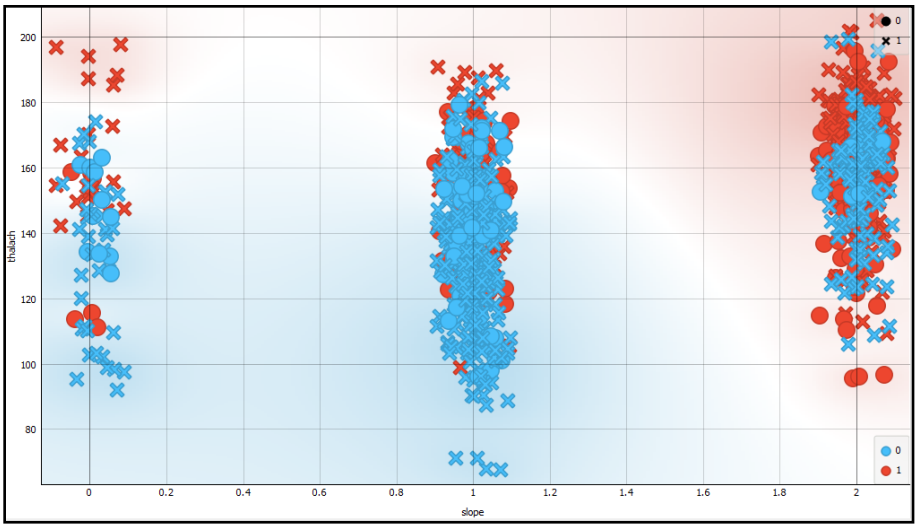
Pearson correlation

1	-0.575	oldpeak	slope	23	-0.129	ca	restecg
2	+0.395	slope	thalach	24	+0.128	chol	trestbps
3	-0.390	age	thalach	25	-0.124	restecg	trestbps
4	+0.371	age	ca	26	-0.120	slope	trestbps
5	-0.350	oldpeak	thalach	27	-0.114	thal	thalach
6	+0.307	cp	thalach	28	-0.108	ca	slope
7	+0.305	ca	oldpeak	29	+0.104	ca	trestbps
8	+0.271	age	trestbps	30	-0.097	slope	thal
9	-0.266	ca	thalach	31	+0.089	chol	thal
10	-0.230	ca	cp	32	+0.086	restecg	slope
11	+0.220	age	chol	33	-0.082	chol	cp
12	+0.208	age	oldpeak	34	-0.072	age	cp
13	+0.203	oldpeak	thal	35	+0.071	age	thal
14	+0.187	oldpeak	trestbps	36	+0.065	chol	oldpeak
15	-0.175	cp	oldpeak	37	+0.058	thal	trestbps
16	-0.174	cp	thal	38	-0.050	oldpeak	restecg
17	-0.169	age	slope	39	+0.048	restecg	thalach
18	-0.147	chol	restecg	40	+0.044	cp	restecg
19	+0.147	ca	thal	41	-0.039	thalach	trestbps
20	+0.141	ca	chol	42	+0.038	cp	trestbps
21	-0.133	age	restecg	43	-0.022	chol	thalach
22	+0.132	cp	slope	44	-0.020	restecg	thal
				45	-0.014	chol	slope

2.7 Correlation between the numerical features

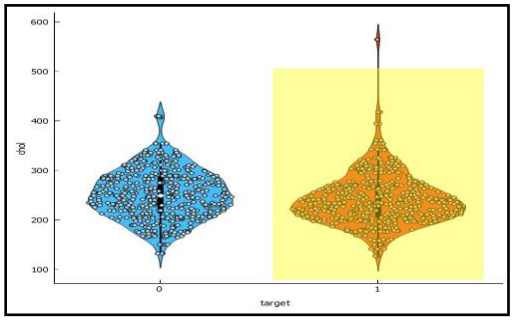


2.8 Correlation between patient age and resting blood pressure



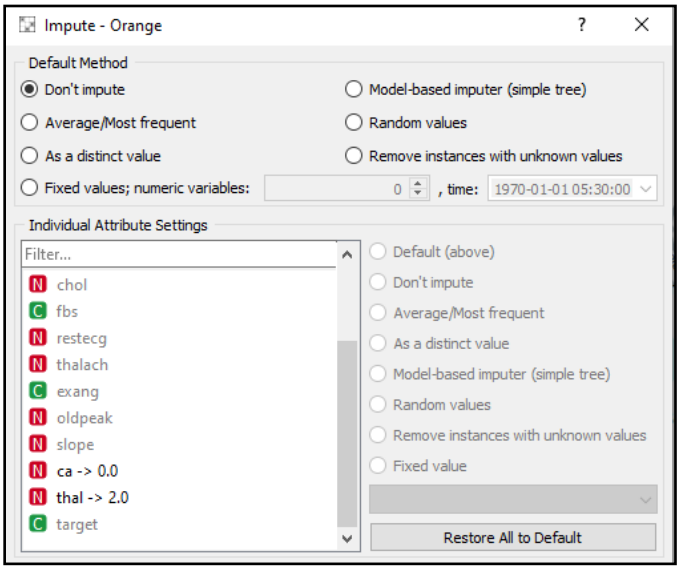
2.9 Correlation between slope of exercise and heart rate

OUTLIER DETECTION



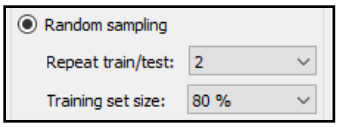
2.10 Detecting the outliers in 'chol' attributes

MISSING VALUES



2.11 Handling missing values

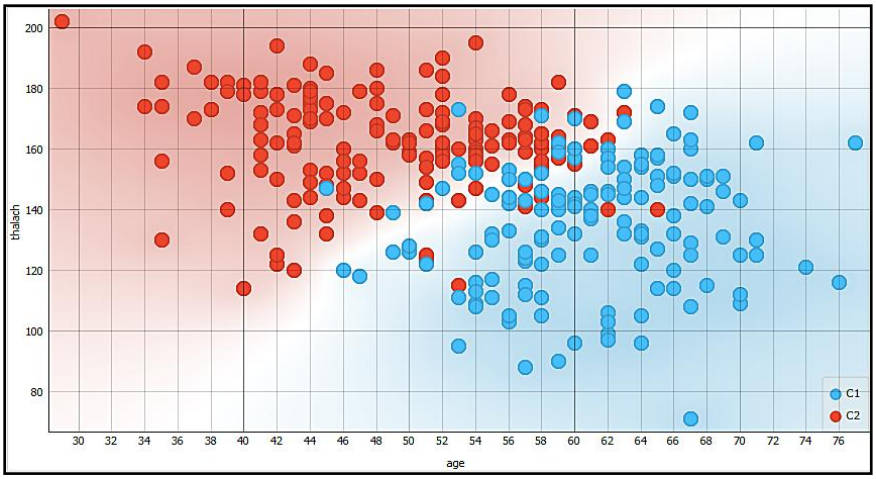
TRAIN/TEST SPLIT VALIDATION



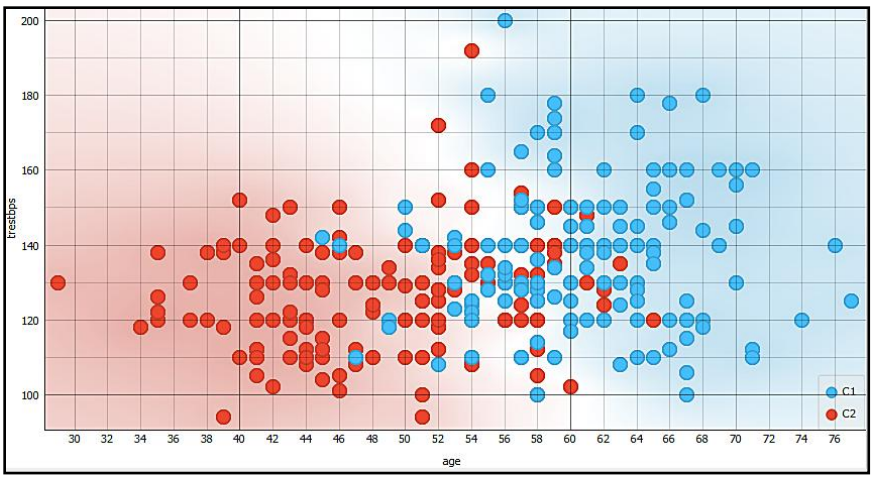
2.12 Train/Test split validation

RESULTS

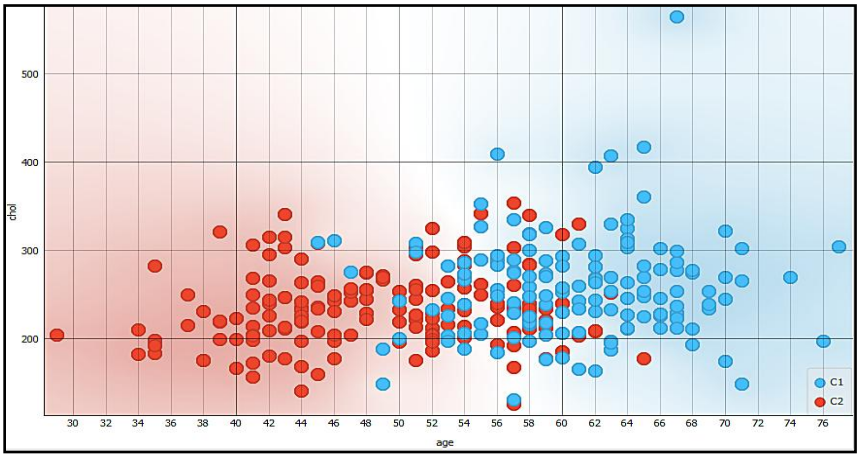
K-means clustering on Heart disease dataset:



2.13 Clustering in age and maximum heart rate (thalach)



2.14 Clustering in age and resting blood pressure (trestbps)



2.15 Clustering in age and cholesterol (chol)

Test and Score

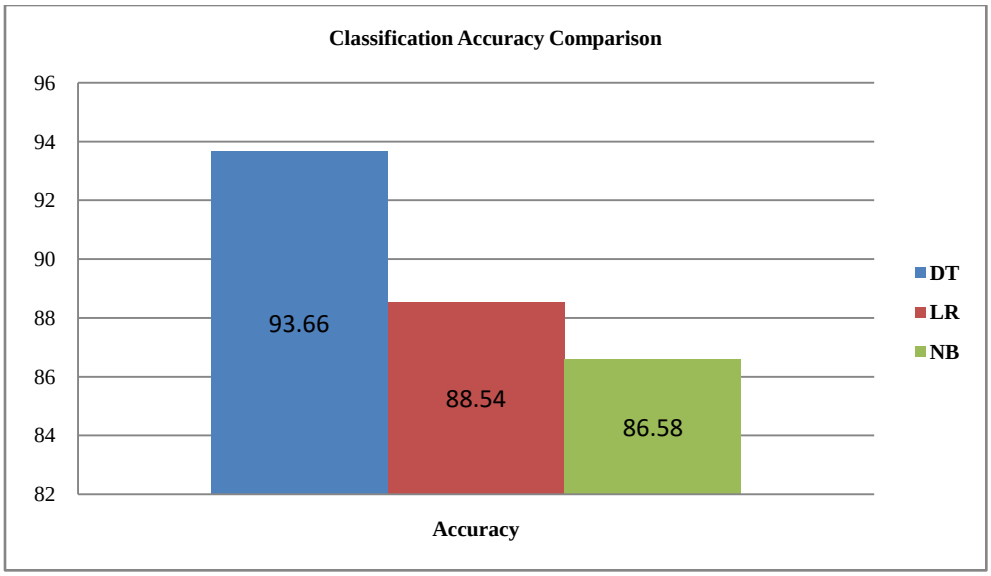
Settings

Sampling type: Stratified Shuffle split, 2 random samples with 80% data
Target class: Average over classes

Scores

Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.9294285714285714	0.8658536585365854	0.8658608441350782	0.8658783338529664	0.8658536585365854
Logistic Regression	0.9480952380952381	0.8853658536585366	0.8849882152974671	0.8884594585529441	0.8853658536585366
Decision Tree	0.955654761904762	0.9365853658536586	0.9365763117939574	0.9381737299362701	0.9365853658536586

2.17. Snapshot of accuracy and performance evaluation measures



2.18 Accuracy of machine learning algorithm

DISCUSSION:

The most suitable model in building a Heart disease predictor is based on the performance of the machine algorithm. The performance summary is given in Fig 2.17.

From Fig 2.17 we can see that DT outperforms all the ML algorithms. DT with train/test ratio of 80:20 has the best predictive accuracy of 93.66%.