Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

A Deductive Learning of Heart Disease Dataset by using K Means Clustering

Arulanantham Zechariah Jebakumar[1*], Dr. R. Ravanan[2]

[1]Lecturer, Prince Sultan Military College of Health Sciences, Dhahran PO Box: 33048, Dammam – 31448, Kingdom of Saudi Arabia.

[2]Joint Director of Collegiate Education, Chennai region, Chennai-15, Tamilnadu, India

Corresponding author: Arulanantham Zechariah Jebakumar

Email: zechariah@psmchs.edu.sa

Abstract

Cardiovascular diseases is one of the most significant causes of mortality in today's world. Cardiovascular diseases are the number one cause of death globally with 17.9 million death cases each year. CVDs are concertedly contributed by hypertension, diabetes, overweight and unhealthy lifestyles. Exploratory Data Analysis is a pre-processing step to understand the data. There are numerous methods and steps in performing EDA, however, most of them are specific, focusing on visualization and distribution. If the number of cluster is 2, this model has 43% & 57% of cluster instances for full training set and 46% & 54% of cluster instances for 66% training set, if the number of cluster is 3, this model has 18% 48% & 34% of cluster instances for full training set and 25%,50% & 25% of cluster instances for 66% training set, if the number of cluster is 4, this model has 21%,40%,10% & 28% of cluster instances for full training set and 24%,13%,26% and 37% of cluster instances for 66% training set, If the number of cluster is 5, this model has 17%,31%,11%,19% & 21% of cluster instances for full training set and 23%,14%,20%,33% &11% of cluster instances for 66% training set, If the number of cluster is 6, this model has 10%,31%,15%,20%,6% &18% of cluster instances for full training set and 16%,18%,15%,22%,13% &15% of cluster instances for 66% training set. In this system proposes the optimal results for build the deductive learning model. Based on the time consumption the system recommends that cluster 2, 3 and 5 have zero second taken the time consumption for build the model in 66% training set. 0.01 seconds for cluster 6 and 0.03 seconds for cluster 4 in 66% training set models. Cluster 5 and 6 have low sum of squared errors for full training and 66% training set comparatively other models.

Keywords: K Means clustering, Centroids, Sum of Squared Errors, Iterations.

Introduction

In this section presents introduction of this research work. 17.9 million people die every year due to heart diseases accounting for 31% of all the deaths in the world. [1]Thus, it is important for early and accurate detection of heart diseases.[2] 4 out of 5 Heart disease patients die due to a heart attack or a stroke, and one-third of the deaths occur prematurely in patients below 70 years of age.[3] Heart disease proves to be the leading cause of death for both men and women.[4] This affects the human life very badly. The diagnosis of heart disease in most cases depends on a complex combination and huge volume of clinical and pathological data. Machine learning has been shown to be effective assisting in making decisions and predictions from the large quantity of data produced by the health care industry.[5] Patients with a high risk of heart diseases demonstrate

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

raised blood pressure, glucose, and lipids along with overweight, and obesity. [6]Lifestyle also plays an important factor in heart diseases along with physiological factors. [7]Tobacco use, unhealthy diet, excessive alcohol intake, and inadequate physical activity are leading reasons for heart diseases.[8]Identifying such people and ensuring they are given appropriate treatment could prevent premature deaths.

In this paper presents section 2 of this paper explains the detail on the related works. In section 3 presents the materials and methods adopted and section 4 presents the details of the experiments and discussions. Finally section 5 concludes the paper by sharing our inferences and future plans.

Related Works

In this section presents focuses the related works of this research work.The accuracy and precision statistics for different algorithms such as Support Vector machines, KNN, Decision Trees, and Neural networks being most popular.[9]TheUCI dataset for comparison of different classifiers such as Multilayer perceptron ,Naive Bayes, KNN etc. and validated that SVM with boosting hyper parameters outperformed others.[10] The machine learningtechniques providing the accuracy of 88.7% in prediction of cardiovascular diseases with a hybrid random forest and linear model.[11]New selection features and methods can be adopted to get broader perception of performance.[12]The traditional machine learning algorithms that aim in improving the accuracy of heart disease prediction. [13]The UK Biobank dataset observed that rather than complex models, information gain was better by consideration of different risk factors.a south African dataset consisting of 462 instances for analyzing algorithms such as Naive Bayes, SVM , and decision trees.[14]Naive bayes obtained good accuracy results however specificity and sensitivity results can be improved with more instances.[15]the accuracy of decision trees in the prediction of heart diseases with the help of a dataset consisting of 573 instances. More number of attributes and hyper parameters can result in better performance classification.[16] Association rules, clustering and other data mining algorithms prove to be useful to mine huge amounts of unstructured data. [17]Various kernel implementations with certain rulebased classifiers.[18]It concludes that the RBF kernel is best for infinite data and Hyper parameter tuning can be added to make the model more effective. [19]

Materials and Methods

In this section presents the materials and methods of this research work. This research work focuses exploratory data analysis and using Weka 3.8.3. The dataset used in this work is UCI Heart Disease dataset. It has 76 features (attributes) from 303 patients. This work uses the dataset consisting of 270 patients with 14 features set.

Table 1: Meta Data Description

S.No	Attribute	Description of the Attribute	Type of the Attribute	Range
1	age	Age in years	Continuous	Minimum = 29 Maximum = 77
2	sex	0=Female ; 1=Male	Binary	Male=183

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

				Female=87
3	chest pain type (cp)	Type of the chest pain	Categorical	Asymtomatic=129 Non Angina=79 Atypical Angina=42 Typical Angina=20
4	resting blood pressure (restbps)	in mm Hg on admission to the hospital	Continuous	Minimum=94 Maximum=200 Mean= 131.34 StdDeviation=17.86
5	serum cholestoral (chol)	serum cholestoral in mg/dl	Continuous	Minimum=126 Maximum=564 Mean= 249.66 StdDeviation=51.69
6	fasting blood sugar (fbs)	0=false; 1=true	Binary	False=230 True=40
7	Resting ECG (restecg)	(fbs>120 mg/dl) 0=Normal ; 1=Having ST-T wave abnormality; 2=Showing probable or define left ventricular hypertrophy	Categorical	Normal=131 ST-T Wave Abnormality=2 Left Ventriclar Hypertrophy=137
8	maximum heart rate achieved (thalach)	maximum heart rate reached	Continuous	Minimum=71 Maximum=202 Mean= 149.68 StdDeviation=23.17
9	exercise induced angina (exang)	0=No; 1=Yes	Binary	No=181 Yes=89
10	oldpeak	ST depreve to restserelatissin induced by exercise relative to rest	Continuous	Minimum=0 Maximum=6.2 Mean= 1.05 StdDeviation=1.145
11	slope	the slope of the peak exercise ST segment 0=upsloping; 1=Flat; 2=Downsloping	Categorical	Flat=122 Upsloping=130 Downsloping=18
12	ca	number of major vessels(0-3) colored by flourosopy 0=Typical Angina 1=Atypical Angina 2=Non Anginal Pain 3=Asymptomatic	Categorical	Typical Angina=160 Atypical Angina=58 Non AnginalPain=33 Asymptomatic=19
13	thal	1=normal; 2=fixed defect; 3=reversible defect	Categorical	Normal=152 Fixed Defect=14 Reversible Defect=104
14	diagnosis of heart disease (Target or Class)	angiographic disease status 0=Disease; 1=No Disease	Binary	Disease=151 No Disease=119

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

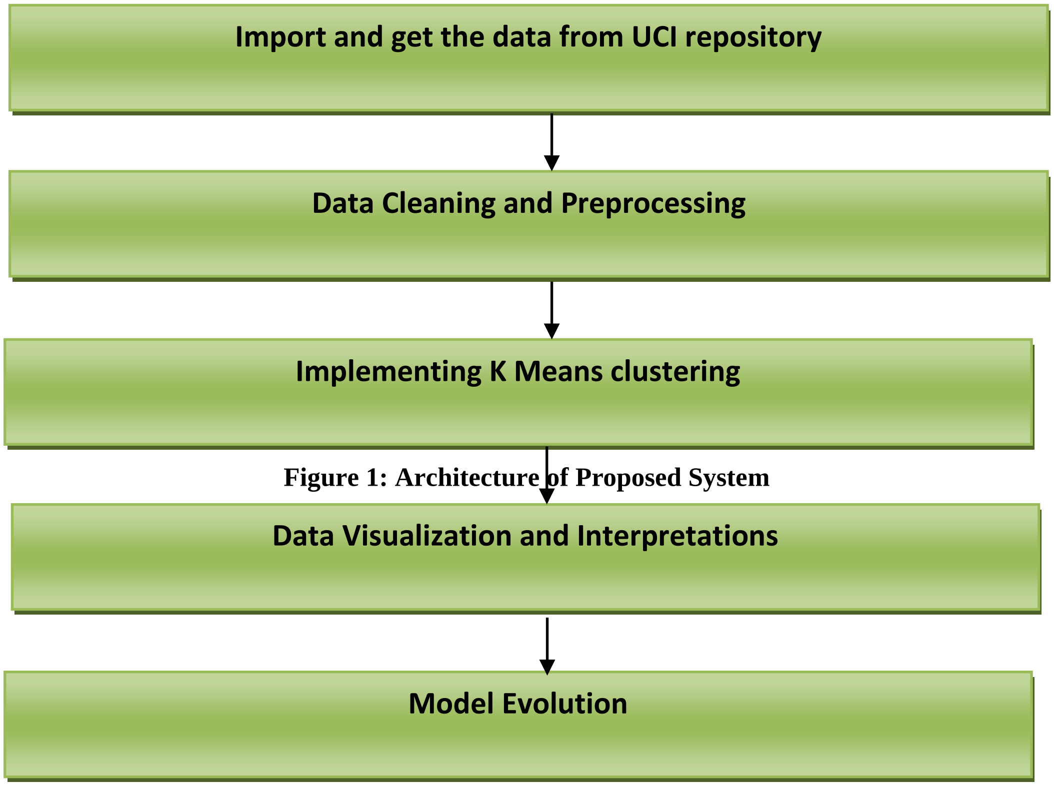

Figure 1: Architecture of Proposed System

Results and Discussion

In this section focuses the results and discussions of this research work. This project covers exploratory data analysis like data visualization and implementing K means clustering approaches by using Weka 3.8.3.

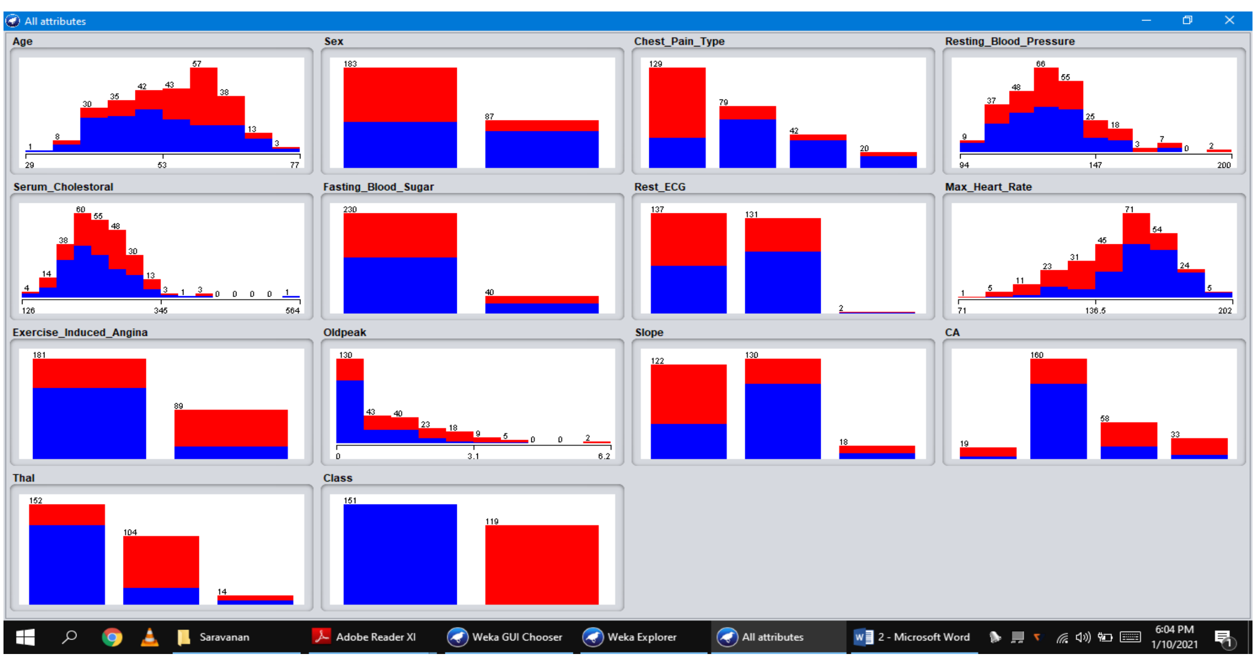

Figure 2: Data Visualization for all Attributes

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

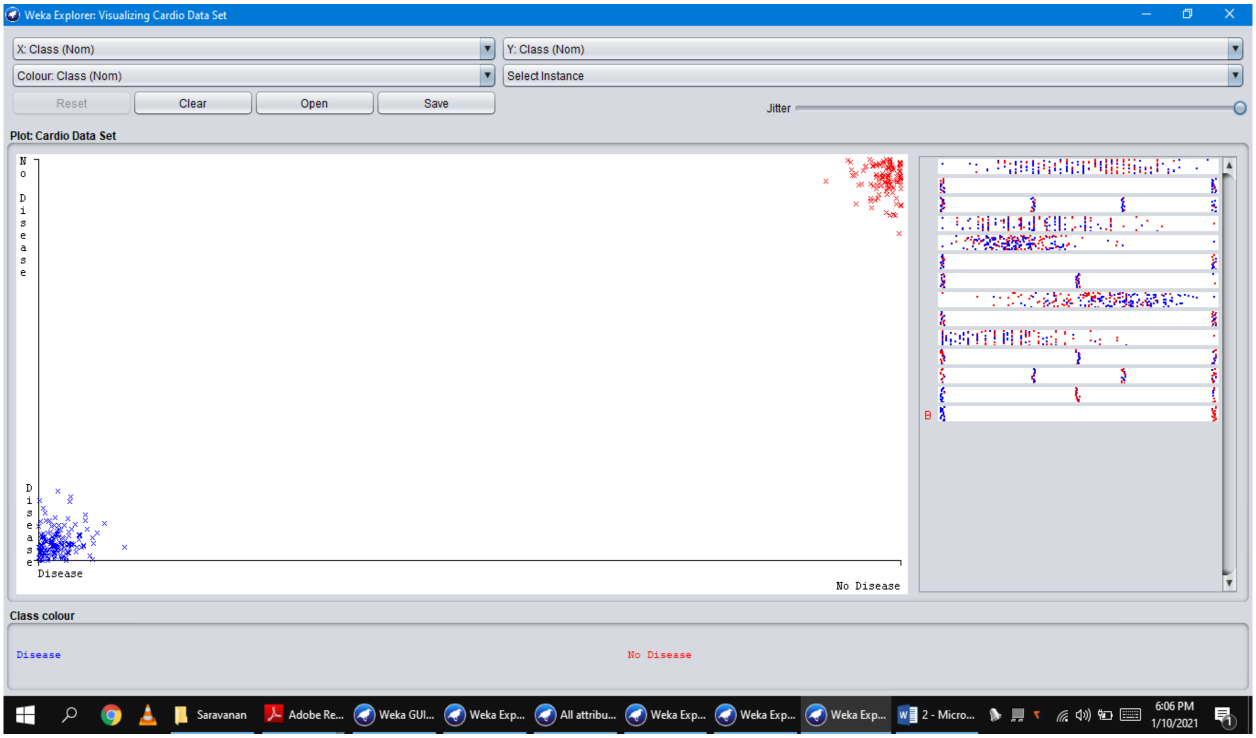

Figure 3: Visualization of Target Attribute

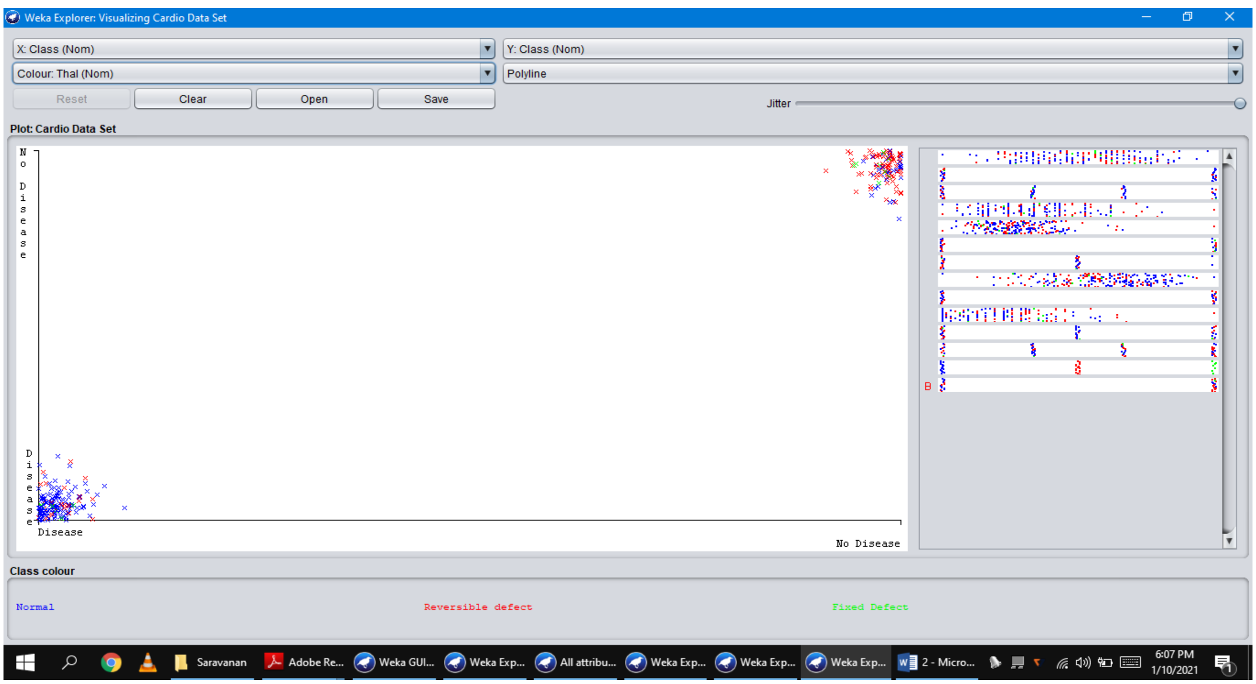

Figure 4: Visualization of ThalAttribute

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

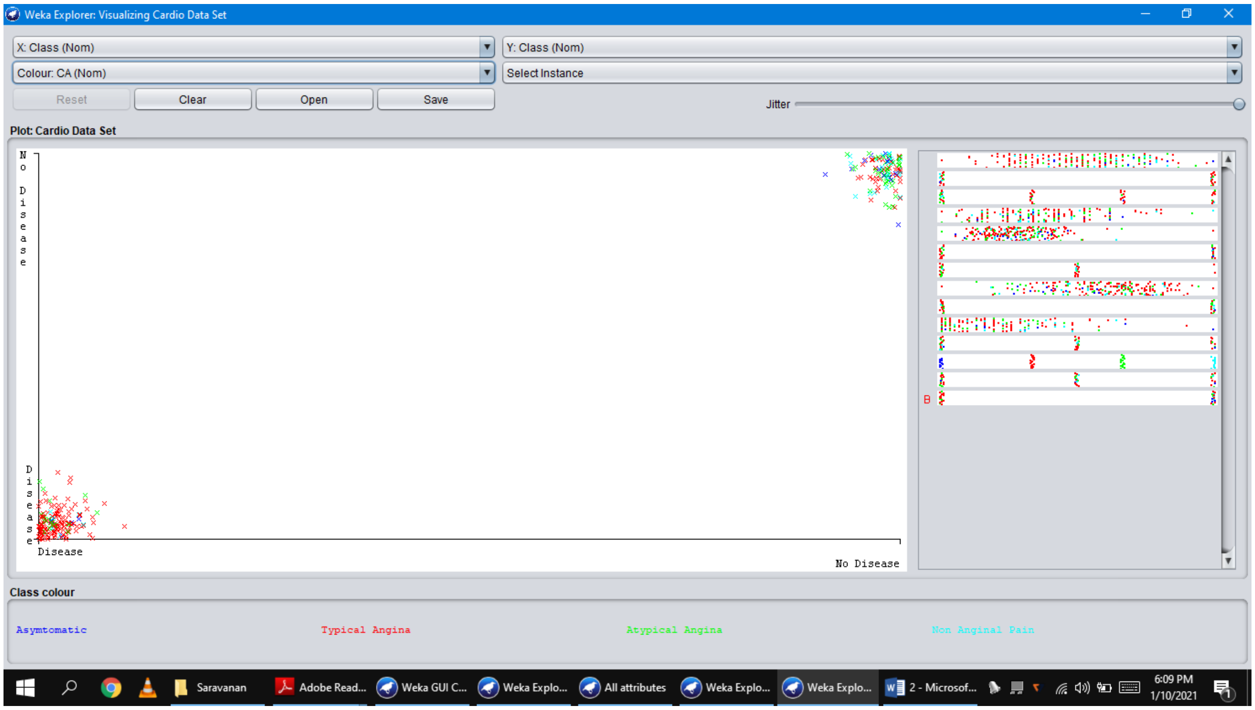

Figure 5: Visualization of CA Attribute

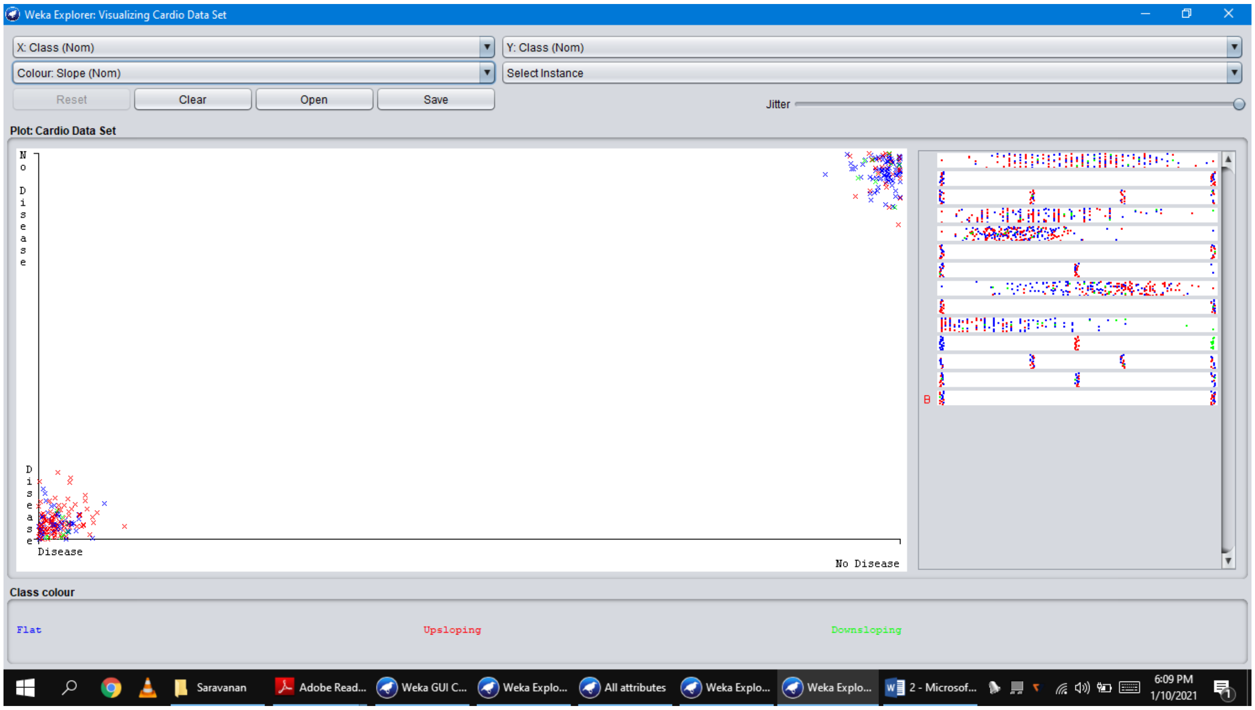

Figure 6: Visualization of Slop Attribute

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

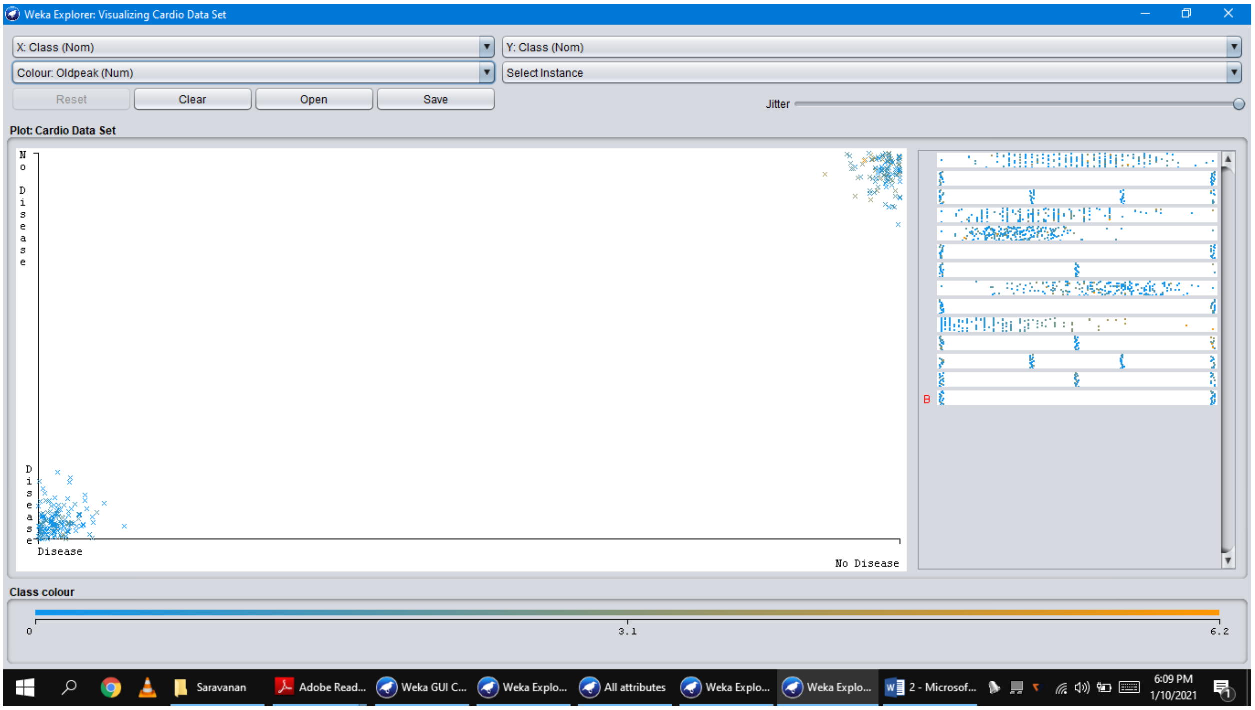

Figure 7: Visualization of Old Peak Attribute

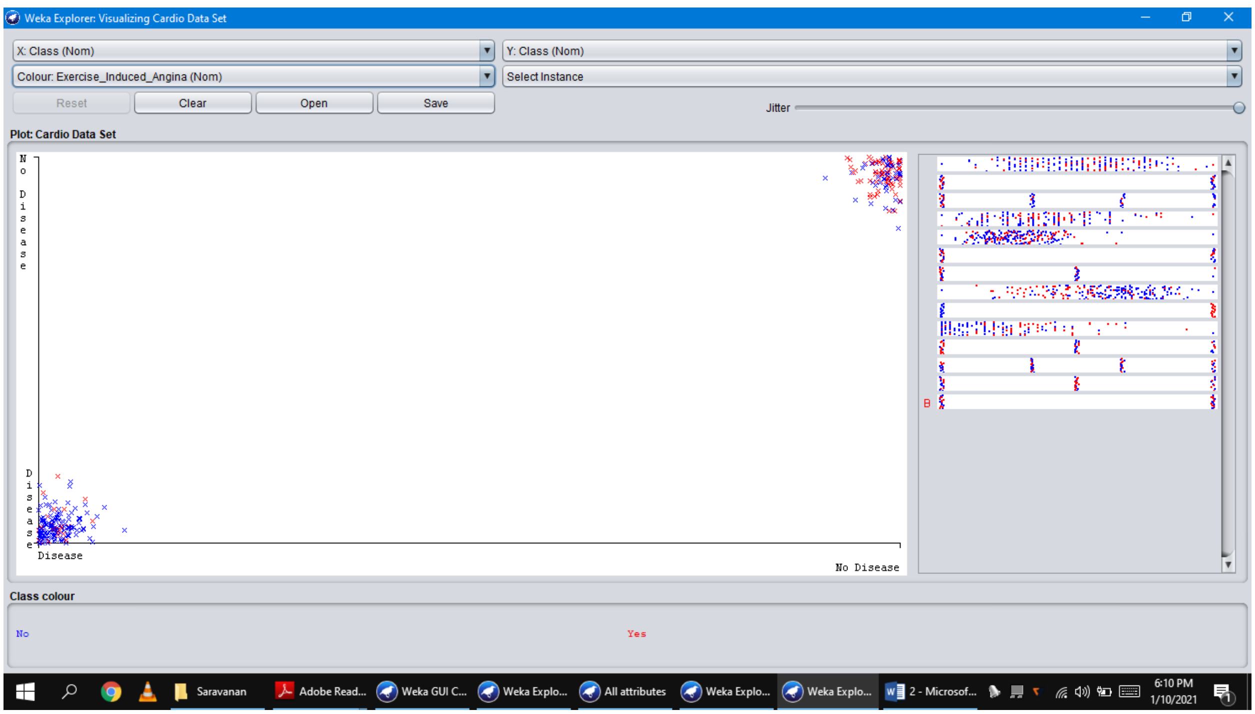

Figure 8: Visualization of Exercise_Induced_Angina Attribute

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

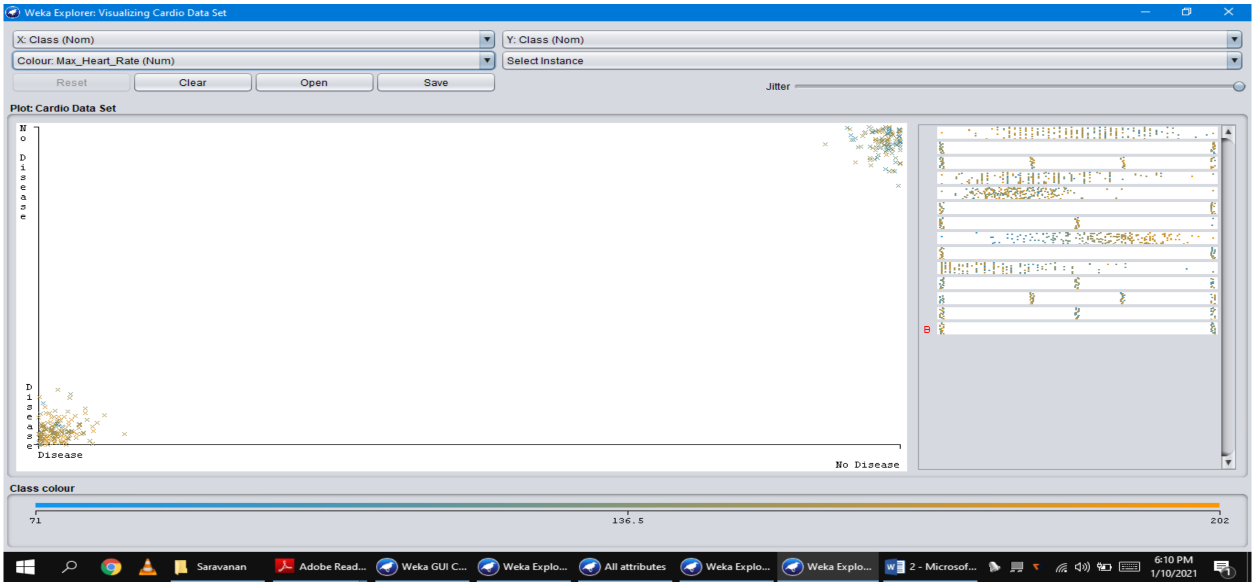

Figure 9: Visualization of Max_Heart Rate Attribute

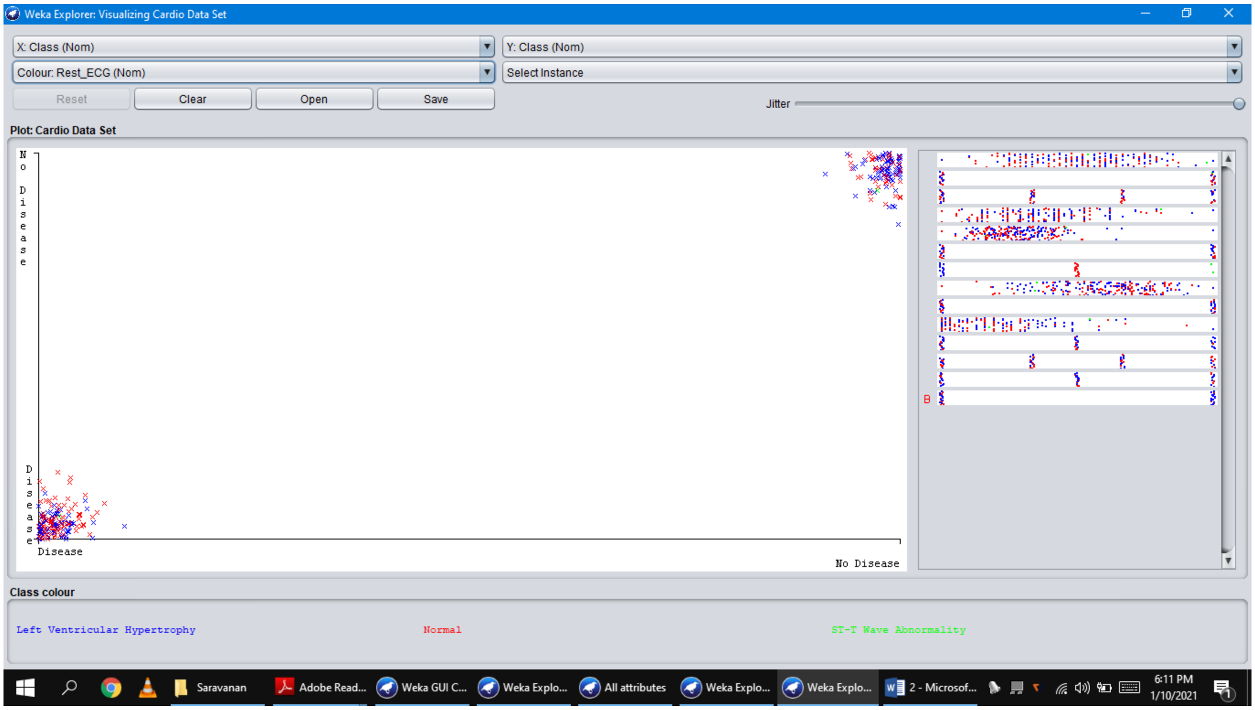

Figure 10: Visualization of Rest_ECG**Attribute**

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

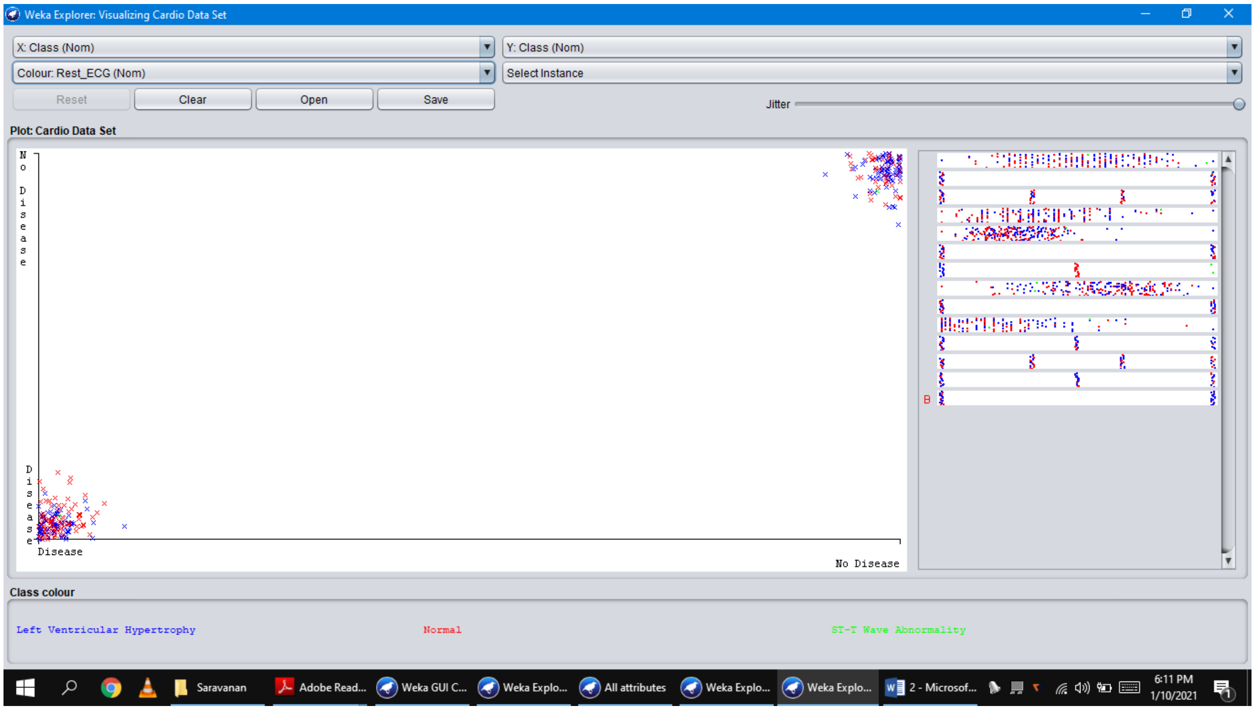

Figure 11: Visualization of Fasting_ECG**Attribute**

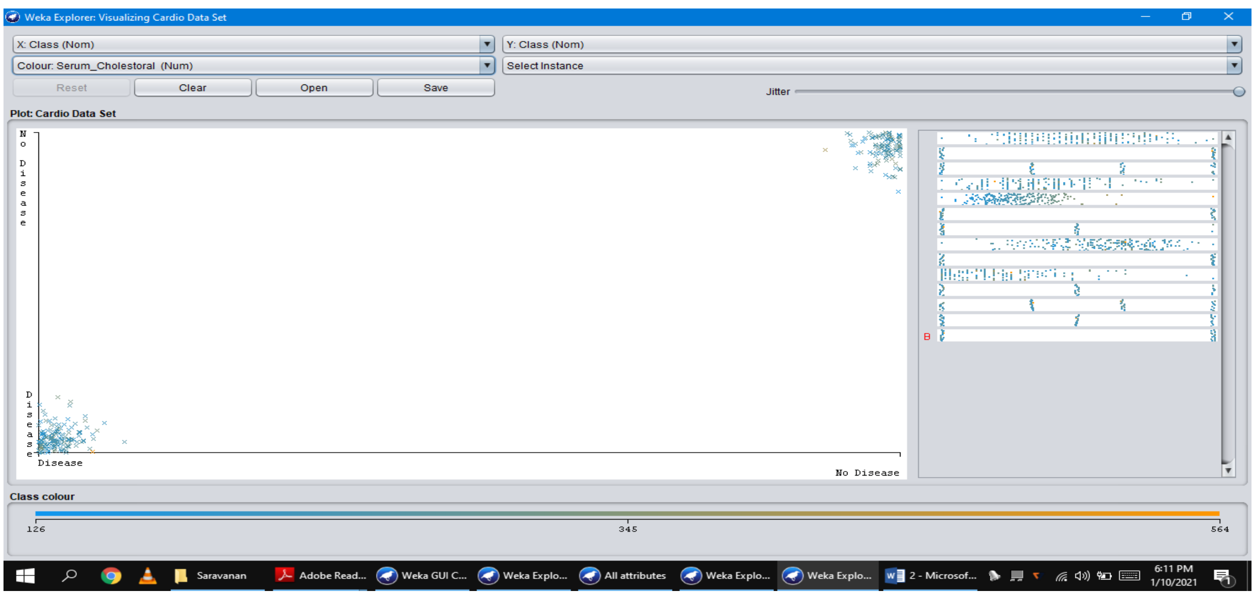

Figure 12: Visualization of Serum_Cholastral**Attribute**

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

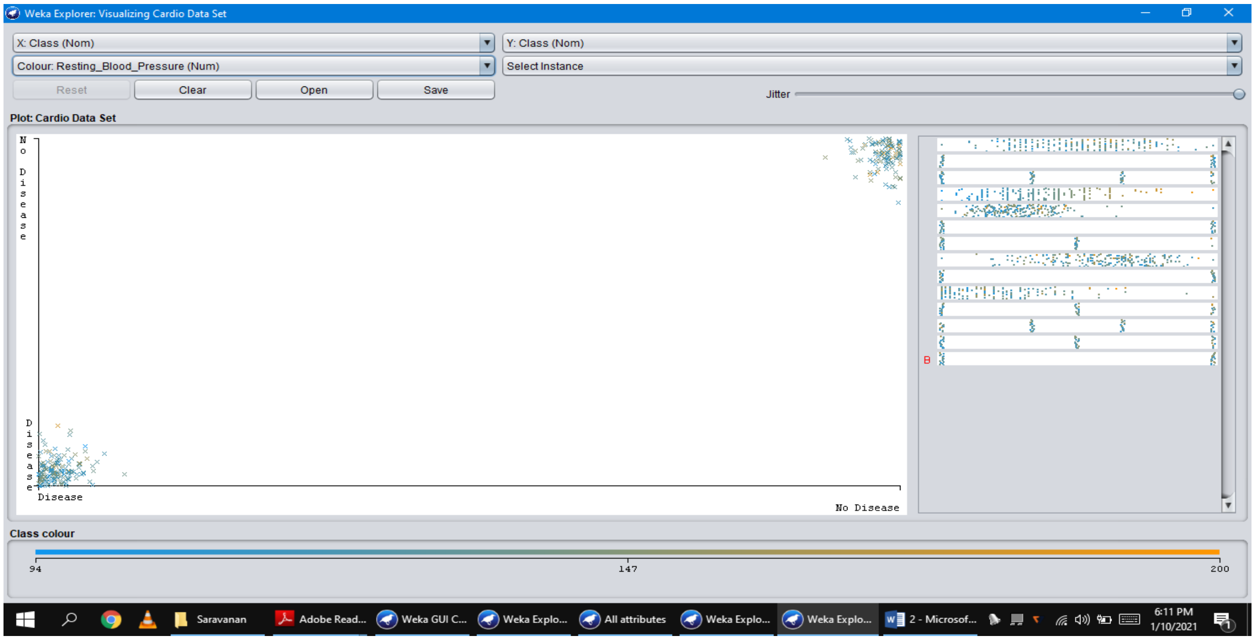

Figure 13: Visualization of Resting_Blood Pressure **Attribute**

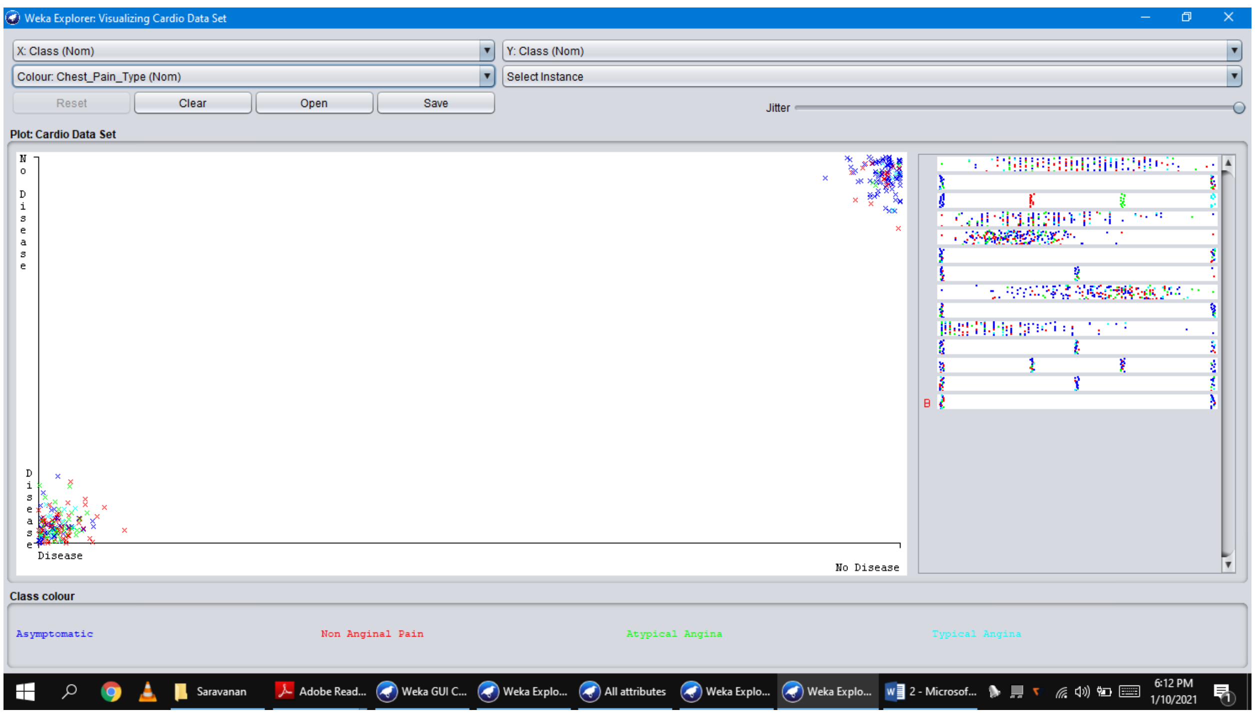

Figure 14: Visualization of Chest PainAttribute

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

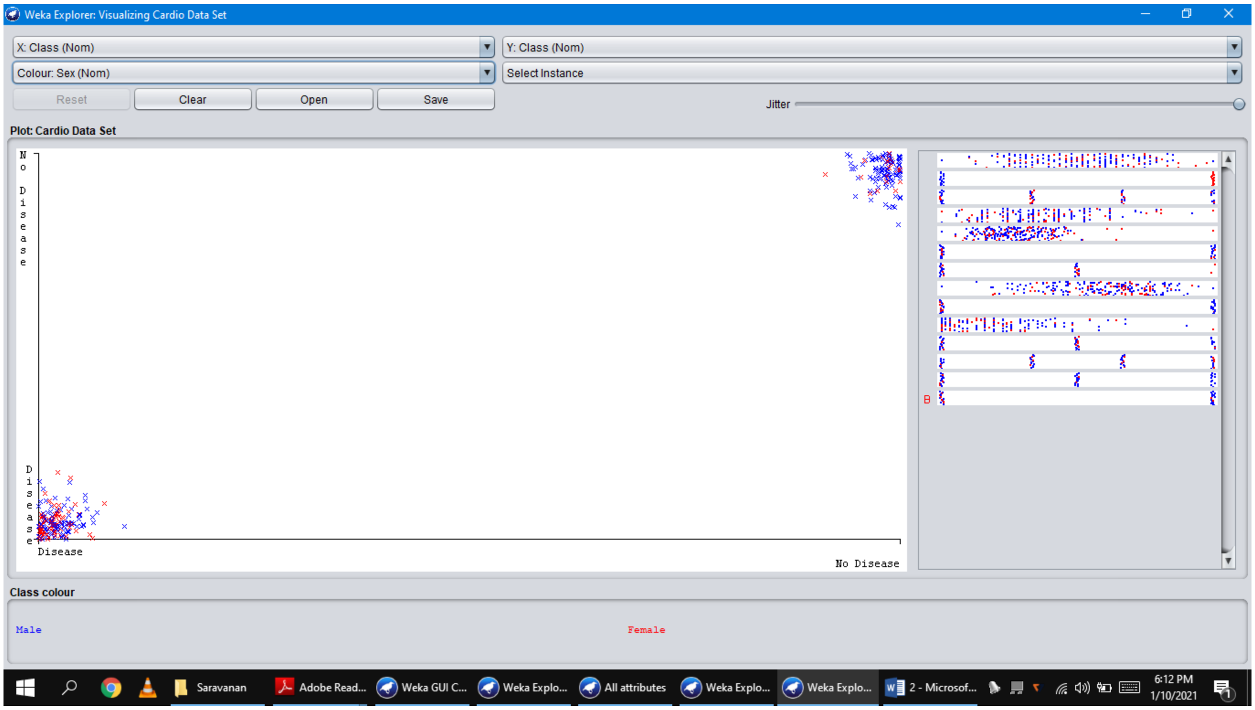

Figure 15: Visualization of Sex **Attribute**

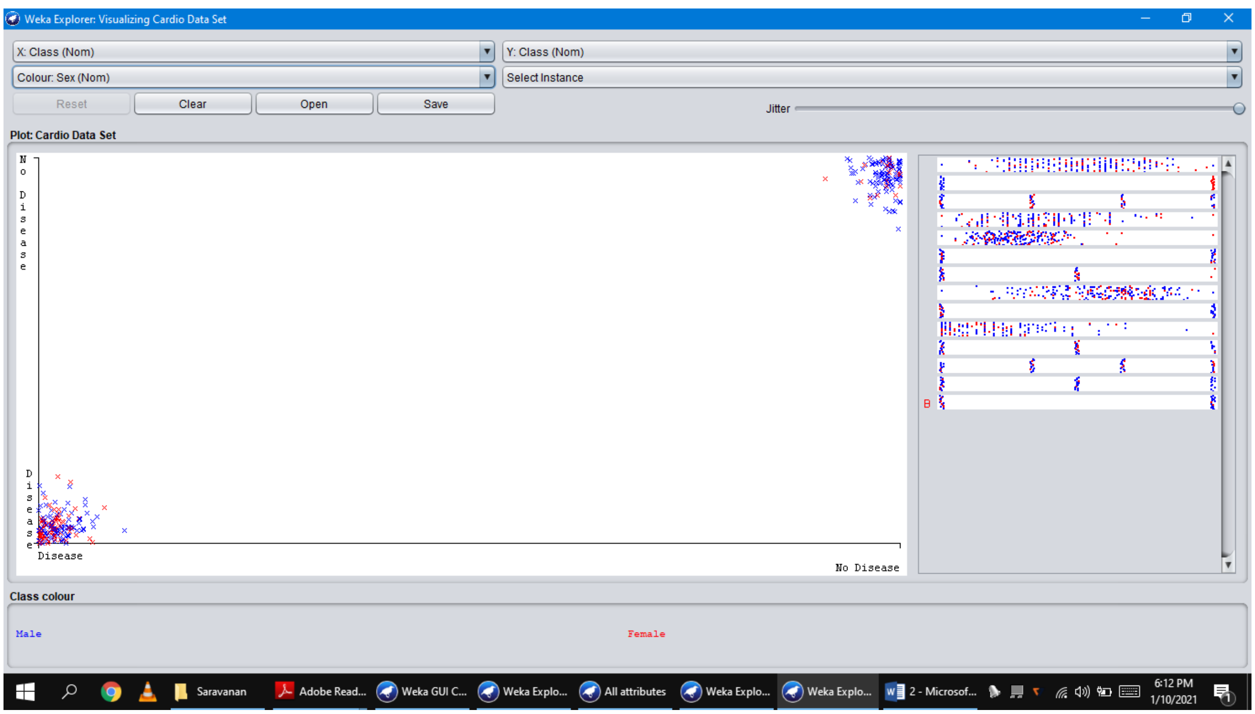

Figure 16: Visualization of Age **Attribute**

The diagrams shows that the data visualizations of all the attributes like there are 14 attributes lies on the disease and no disease classes.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 710.4670812208781

Initial starting points (random):

Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non Anginal Pain',Normal,'No Disease'
Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                  Cluster#
Attribute                                          Full Data               0                1
                                                     (270.0)         (115.0)          (155.0)
=============================================================================================
Age                                                  54.4333         56.8348          52.6516
Sex                                                     Male            Male             Male
Chest_Pain_Type                                 Asymptomatic    Asymptomatic Non Anginal Pain
Resting_Blood_Pressure                              131.3444        133.8783         129.4645
Serum_Cholestoral                                   249.6593        254.9913         245.7032
Fasting_Blood_Sugar                                    FALSE           FALSE            FALSE
Rest_ECG                    Left Ventricular Hypertrophy Left Ventricular Hypertrophy  Normal
Max_Heart_Rate                                      149.6778        136.9826         159.0968
Exercise_Induced_Angina                                   No             Yes               No
Oldpeak                                                 1.05          1.6157           0.6303
Slope                                              Upsloping            Flat        Upsloping
CA                                            Typical Angina  Atypical Angina  Typical Angina
Thal                                                  Normal Reversible defect         Normal
Class                                                Disease      No Disease          Disease


Time taken to build model (full training data) : 0.01 seconds
```

Figure 17: K Means cluster No=2

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 648.2251333482561

Initial starting points (random):

Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non Anginal Pain',Normal,'No Disease'
Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease
Cluster 2: 60,Male,Asymptomatic,117,230,TRUE,Normal,160,Yes,1.4,Upsloping,'Non Anginal Pain','Reversible defect ','No Disease'

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                              Cluster#
Attribute                                      Full Data             0                1                2
                                                 (270.0)        (49.0)          (130.0)           (91.0)
==========================================================================================================
Age                                              54.4333            59             51.4          56.3077
Sex                                                 Male        Female             Male             Male
Chest_Pain_Type                             Asymptomatic  Asymptomatic Non Anginal Pain     Asymptomatic
Resting_Blood_Pressure                          131.3444      134.6735         129.3308         132.4286
Serum_Cholestoral                               249.6593      271.6735         240.4308          250.989
Fasting_Blood_Sugar                                FALSE         FALSE            FALSE            FALSE
Rest_ECG                Left Ventricular Hypertrophy Left Ventricular Hypertrophy Normal Left Ventricular Hypertrophy
Max_Heart_Rate                                  149.6778      145.4694         161.9308         134.4396
Exercise_Induced_Angina                               No            No               No              Yes
Oldpeak                                             1.05        1.1816           0.5792           1.6516
Slope                                          Upsloping          Flat        Upsloping             Flat
CA                                        Typical Angina Typical Angina  Typical Angina  Atypical Angina
Thal                                              Normal        Normal           Normal Reversible defect
Class                                            Disease    No Disease          Disease       No Disease


Time taken to build model (full training data) : 0.01 seconds
```

Figure 18: K Means cluster=3

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

```
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 608.7121457525952

Initial starting points (random):

Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non Anginal Pain',Normal,'No Disease'
Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease
Cluster 2: 60,Male,Asymptomatic,117,230,TRUE,Normal,160,Yes,1.4,Upsloping,'Non Anginal Pain','Reversible defect ','No Disease'
Cluster 3: 64,Male,Asymptomatic,128,263,FALSE,Normal,105,Yes,0.2,Flat,'Atypical Angina','Reversible defect ',Disease

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                          Cluster#
Attribute                                 Full Data              0                   1                  2                  3
                                            (270.0)         (57.0)             (109.0)             (28.0)             (76.0)
==========================================================================================================================
Age                                         54.4333        58.5263             50.3303                 56            56.6711
Sex                                            Male         Female                Male               Male               Male
Chest_Pain_Type                        Asymptomatic   Asymptomatic    Non Anginal Pain       Asymptomatic       Asymptomatic
Resting_Blood_Pressure                     131.3444       134.0877            128.1193           132.0357           133.6579
Serum_Cholestoral                          249.6593        263.193            237.9817           261.0714           252.0526
Fasting_Blood_Sugar                           FALSE          FALSE               FALSE              FALSE              FALSE
Rest_ECG                Left Ventricular Hypertrophy Left Ventricular Hypertrophy  Normal Left Ventricular Hypertrophy Left Ventricular Hypertrophy
Max_Heart_Rate                             149.6778       146.9123            162.9633           155.8929           130.4079
Exercise_Induced_Angina                          No             No                  No                Yes                Yes
Oldpeak                                        1.05         1.0667               0.511             0.7679             1.9145
Slope                                     Upsloping           Flat           Upsloping          Upsloping               Flat
CA                                   Typical Angina Typical Angina      Typical Angina   Non Anginal Pain    Atypical Angina
Thal                                         Normal         Normal              Normal  Reversible defect  Reversible defect
Class                                       Disease        Disease             Disease         No Disease         No Disease
```

Figure 19: K Means cluster=4

```
ons: 5
m of squared errors: 581.9453678366942

points (random):

ale,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non Anginal Pain',Normal,'No Disease'
e,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease
e,Asymptomatic,117,230,TRUE,Normal,160,Yes,1.4,Upsloping,'Non Anginal Pain','Reversible defect ','No Disease'
e,Asymptomatic,128,263,FALSE,Normal,105,Yes,0.2,Flat,'Atypical Angina','Reversible defect ',Disease
ale,Asymptomatic,128,303,FALSE,'Left Ventricular Hypertrophy',159,No,0,Upsloping,'Atypical Angina',Normal,Disease

obally replaced with mean/mode

troids:
                                          Cluster#
                         Full Data               0                1                  2                 3                  4
                           (270.0)          (45.0)           (85.0)             (31.0)            (51.0)             (58.0)
==========================================================================================================================
                           54.4333         59.5778          47.8118            56.6129           55.2549            58.2586
                              Male            Male             Male               Male              Male             Female
                      Asymptomatic    Asymptomatic Non Anginal Pain       Asymptomatic      Asymptomatic   Non Anginal Pain
ssure                     131.3444        139.7556         125.7882           134.6774          129.3725           132.9138
                          249.6593        250.8444         231.1176           264.8387          246.3529           270.7069
ar                           FALSE           FALSE            FALSE              FALSE             FALSE              FALSE
      Left Ventricular Hypertrophy Left Ventricular Hypertrophy Normal Left Ventricular Hypertrophy Normal Left Ventricular Hypertrophy
                          149.6778        142.4667         164.8235           145.7742           129.098           153.2586
Angina                          No              No               No                Yes               Yes                 No
                              1.05          1.9333           0.4741              1.371            1.6431             0.5155
                         Upsloping            Flat        Upsloping          Upsloping              Flat          Upsloping
                    Typical Angina  Typical Angina   Typical Angina   Non Anginal Pain   Atypical Angina     Typical Angina
                            Normal          Normal           Normal  Reversible defect Reversible defect             Normal
                           Disease      No Disease          Disease         No Disease        No Disease            Disease
```

Figure 20: K Means cluster=5

The above pictures shown that the K Means clusters of all attributes (14 attributes) in the heart disease dataset for implementing deductive learning process.

Table 2: Metrics of K Means Clusters

S.No	Number of Clusters	Number of Iterations	Sum of Squared Errors (Within cluster)	Clustered Instances	Time Taken to Build the Model (In

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

								seconds)	
		A	**B**	**A**	**B**	**A**	**B**	**A**	**B**
1	2	3	4	710.46	466.36	0-115(43%) 1-155(57%)	0-42 (46%) 1-50 (54%)	0.01	0
2	3	4	4	648.26	426.63	0-49(18%) 1-130(48%) 2-91(34%)	0-23(25%) 1-46(50%) 2-23(25%)	0.01	0
3	4	5	8	608.71	398.56	0-57(21%) 1-109(40%) 2-28(10%) 3-76(28%)	0-22(24%) 1-12(13%) 2-24(26%) 3-34(37%)	0.01	0.03
4	5	6	5	581.95	379.77	0-45(17%) 1-85(31%) 2-31(11%) 3-51(19%) 4-58(21%)	0-21(23%) 1-13(14%) 2-18(20%) 3-30(33%) 4-10(11%)	0.01	0
5	6	7	9	572.62	355.02	0-27(10%) 1-83(31%) 2-40(15%) 3-55(20%) 4-17(6%) 5-48(18%)	0 -15 (16%) 1-17(18%) 2-14(15%) 3-20(22%) 4-12(13%) 5-14(15%)	0.02	0.01

- Cluster Model (Full Training Set) =A
- Cluster Model(66% Split)=B

* Implementing Euclidean distance (or similarity) function.

The above table represents that the various measurements producing while implementing full training and 66% training set of the heart disease dataset.

The below table represents that the centroid clusters of K means clusters for full and 66% training set in Weka 3.8.3 tool.

Table 3: Centroid clusters of K Means Clusters for Full / 66% Training set

Cluster Centroids / Clustering model (full training set)		
S.No	**Number of Clusters**	**Initial starting points (random)**
1	2	Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

		Hypertrophy',160,No,3.6,Downsloping,'Non AnginalPain',Normal,'No Disease' Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease
2	3	Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non AnginalPain',Normal,'No Disease' Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease Cluster 2: 60,Male,Asymptomatic,117,230,TRUE,Normal,160,Yes,1.4,Upsloping,'Non AnginalPain','Reversible defect ','No Disease'
3	4	Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non AnginalPain',Normal,'No Disease' Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease Cluster 2: 60,Male,Asymptomatic,117,230,TRUE,Normal,160,Yes,1.4,Upsloping,'Non AnginalPain','Reversible defect ','No Disease' Cluster 3: 64,Male,Asymptomatic,128,263,FALSE,Normal,105,Yes,0.2,Flat,'Atypical Angina','Reversible defect ',Disease
4	5	Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non AnginalPain',Normal,'No Disease' Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease Cluster 2: 60,Male,Asymptomatic,117,230,TRUE,Normal,160,Yes,1.4,Upsloping,'Non AnginalPain','Reversible defect ','No Disease' Cluster 3: 64,Male,Asymptomatic,128,263,FALSE,Normal,105,Yes,0.2,Flat,'Atypical Angina','Reversible defect ',Disease Cluster 4: 57,Female,Asymptomatic,128,303,FALSE,'Left Ventricular Hypertrophy',159,No,0,Upsloping,'Atypical Angina',Normal,Disease
5	6	Cluster 0: 62,Female,Asymptomatic,140,268,FALSE,'Left Ventricular Hypertrophy',160,No,3.6,Downsloping,'Non AnginalPain',Normal,'No Disease' Cluster 1: 42,Male,Asymptomatic,140,226,FALSE,Normal,178,No,0,Upsloping,'Typical Angina',Normal,Disease Cluster 2: 60,Male,Asymptomatic,117,230,TRUE,Normal,160,Yes,1.4,Upsloping,'Non AnginalPain','Reversible defect ','No Disease' Cluster 3: 64,Male,Asymptomatic,128,263,FALSE,Normal,105,Yes,0.2,Flat,'Atypical Angina','Reversible defect ',Disease Cluster 4: 57,Female,Asymptomatic,128,303,FALSE,'Left Ventricular Hypertrophy',159,No,0,Upsloping,'Atypical Angina',Normal,Disease

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

		Cluster 5: 50,Female,Asymptomatic,110,254,FALSE,'Left Ventricular Hypertrophy',159,No,0,Upsloping,'Typical Angina',Normal,Disease
Cluster Centroids / Clustering model (66% Slit)		
1	2	Cluster 0: 48,Male,'Non Anginal Pain',124,255,TRUE,Normal,175,No,0,Upsloping,'Non Anginal Pain',Normal,Disease Cluster 1: 38,Male,'Typical Angina',120,231,FALSE,Normal,182,Yes,3.8,Flat,'Typical Angina','Reversible defect ','No Disease'
2	3	Cluster 0: 48,Male,'Non Anginal Pain',124,255,TRUE,Normal,175,No,0,Upsloping,'Non Anginal Pain',Normal,Disease Cluster 1: 38,Male,'Typical Angina',120,231,FALSE,Normal,182,Yes,3.8,Flat,'Typical Angina','Reversible defect ','No Disease' Cluster 2: 44,Male,'Atypical Angina',120,263,FALSE,Normal,173,No,0,Upsloping,'Typical Angina','Reversible defect ',Disease
3	4	Cluster 0: 48,Male,'Non Anginal Pain',124,255,TRUE,Normal,175,No,0,Upsloping,'Non Anginal Pain',Normal,Disease Cluster 1: 38,Male,'Typical Angina',120,231,FALSE,Normal,182,Yes,3.8,Flat,'Typical Angina','Reversible defect ','No Disease' Cluster 2: 44,Male,'Atypical Angina',120,263,FALSE,Normal,173,No,0,Upsloping,'Typical Angina','Reversible defect ',Disease Cluster 3: 61,Male,Asymptomatic,120,260,FALSE,Normal,140,Yes,3.6,Flat,'Atypical Angina','Reversible defect ','No Disease'
4	5	Cluster 0: 48,Male,'Non Anginal Pain',124,255,TRUE,Normal,175,No,0,Upsloping,'Non Anginal Pain',Normal,Disease Cluster 1: 38,Male,'Typical Angina',120,231,FALSE,Normal,182,Yes,3.8,Flat,'Typical Angina','Reversible defect ','No Disease' Cluster 2: 44,Male,'Atypical Angina',120,263,FALSE,Normal,173,No,0,Upsloping,'Typical Angina','Reversible defect ',Disease Cluster 3: 61,Male,Asymptomatic,120,260,FALSE,Normal,140,Yes,3.6,Flat,'Atypical Angina','Reversible defect ','No Disease' Cluster 4: 58,Male,Asymptomatic,150,270,FALSE,'Left Ventricular Hypertrophy',111,Yes,0.8,Upsloping,'Typical Angina','Reversible defect ','No Disease'
5	6	Cluster 0: 48,Male,'Non Anginal Pain',124,255,TRUE,Normal,175,No,0,Upsloping,'Non Anginal Pain',Normal,Disease Cluster 1: 38,Male,'Typical Angina',120,231,FALSE,Normal,182,Yes,3.8,Flat,'Typical Angina','Reversible defect ','No Disease'

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

		Cluster 2: 44,Male,'Atypical Angina',120,263,FALSE,Normal,173,No,0,Upsloping,'Typical Angina','Reversible defect ',Disease Cluster 3: 61,Male,Asymptomatic,120,260,FALSE,Normal,140,Yes,3.6,Flat,'Atypical Angina','Reversible defect ','No Disease' Cluster 4: 58,Male,Asymptomatic,150,270,FALSE,'Left Ventricular Hypertrophy',111,Yes,0.8,Upsloping,'Typical Angina','Reversible defect ','No Disease' Cluster 5: 67,Male,Asymptomatic,120,237,FALSE,Normal,71,No,1,Flat,'Typical Angina',Normal,'No Disease'

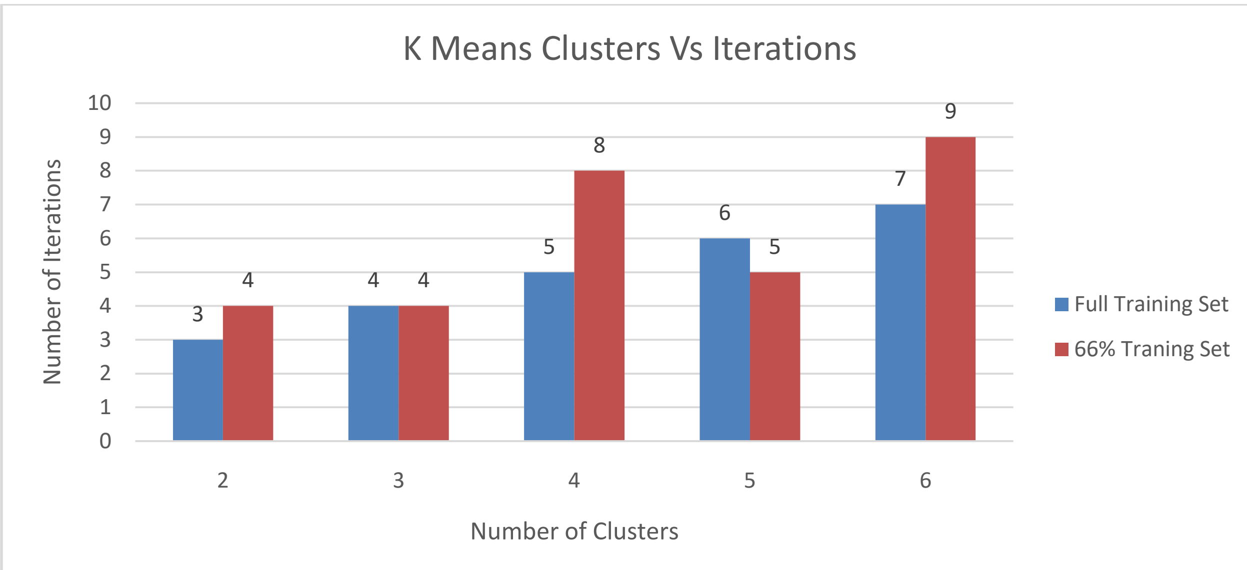

Figure 21: K Means cluster Vs Iterations

The above diagram clearly shows that number of cluster is 2, the model produces 3 iterations for full training set and 4 iterations for 66% training set, if the number of cluster is 3, the model produces 4 iterations for full training set and 66% training set, if the number of cluster is 4, the model produces 5 iterations for full training set and 8 iterations for 66% training set, If the number of cluster is 5, the model produces the 6 iterations for full training set and 5 iterations for 66% training set, If the number of cluster is 6, the model produces 7 iterations for full training set and 9 iterations for 66% training set.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

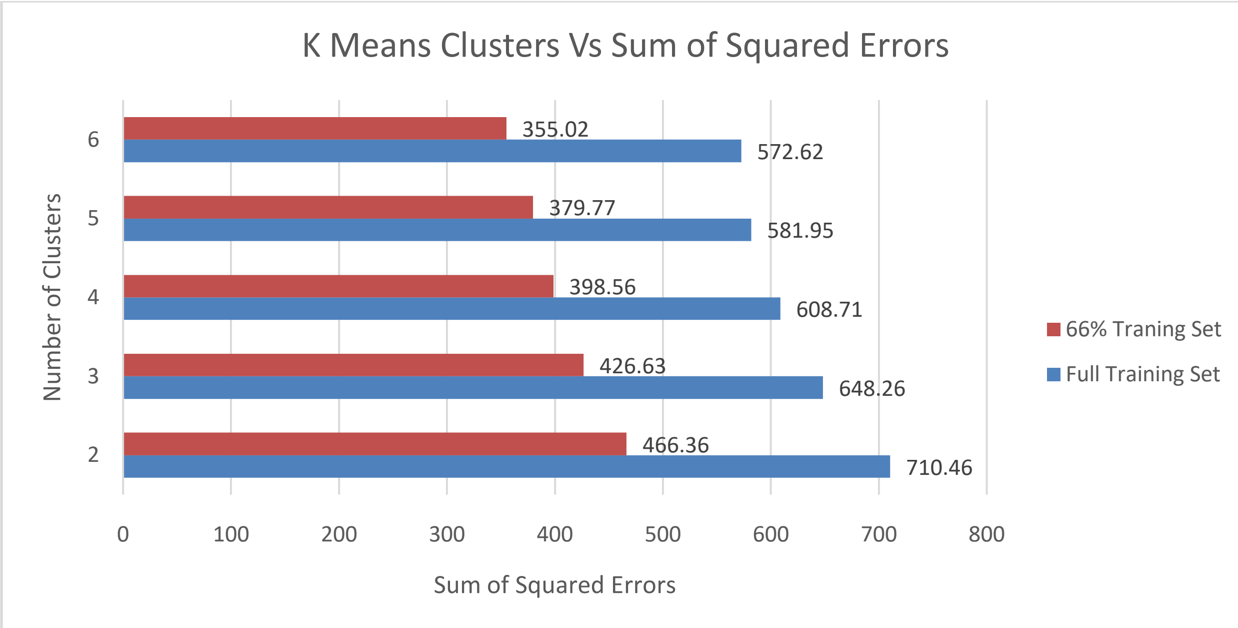

Figure 22: K Means cluster Vs SSE

The above diagram clearly shows that number of cluster is 2, this model has 710.46 sum of squared errors for full training set and 466.36 sum of squared errors for 66% training set, if the number of cluster is 3, this model has 648.26 sum of squared errors for full training set and 426.63 sum of squared errors for 66% training set, if the number of cluster is 4, this model has 608.71 sum of squared errors for full training set and 398.56 sum of squared errors for 66% training set, If the number of cluster is 5, this model has 581.95 sum of squared errors for full training set and 379.77 sum of squared errors for 66% training set ,If the number of cluster is 6 this model has 572.62 sum of squared errors for full training set and 355.02 sum of squared errors for 66% training set.

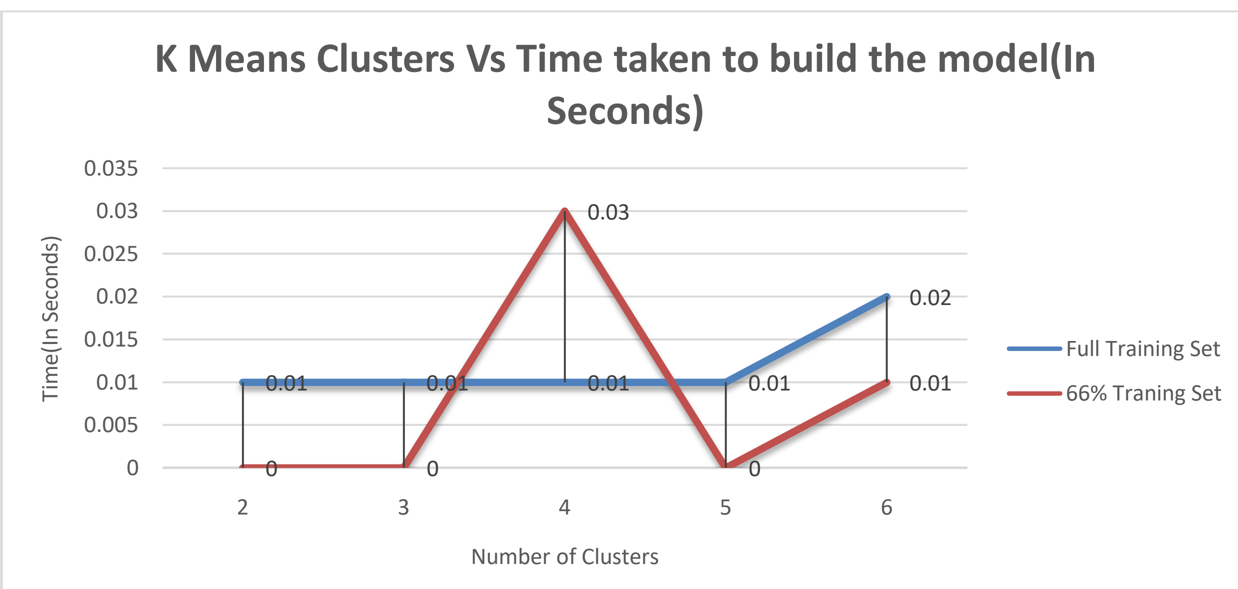

Figure 23: K Means cluster VsTime(Seconds)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

The above diagram clearly shows that number of cluster is 2, this model has taken the time to build the model is 0.01 seconds for full training set and zero second for 66% training set, if the number of cluster is 3, this model has taken the time to build the model is 0.01 seconds for full training set and zero second for 66% training set, if the number of cluster is 4, this model has taken the time to build the model is 0.01 seconds for full training set and 0.03 seconds for 66% training set, If the number of cluster is 5, this model has taken the time to build the model is zero second for full training set and 0.01 seconds for 66% training set ,If the number of cluster is 6,this model has taken the time to build the model is 0.02 seconds for full training set and 0.01 seconds for 66% training set.

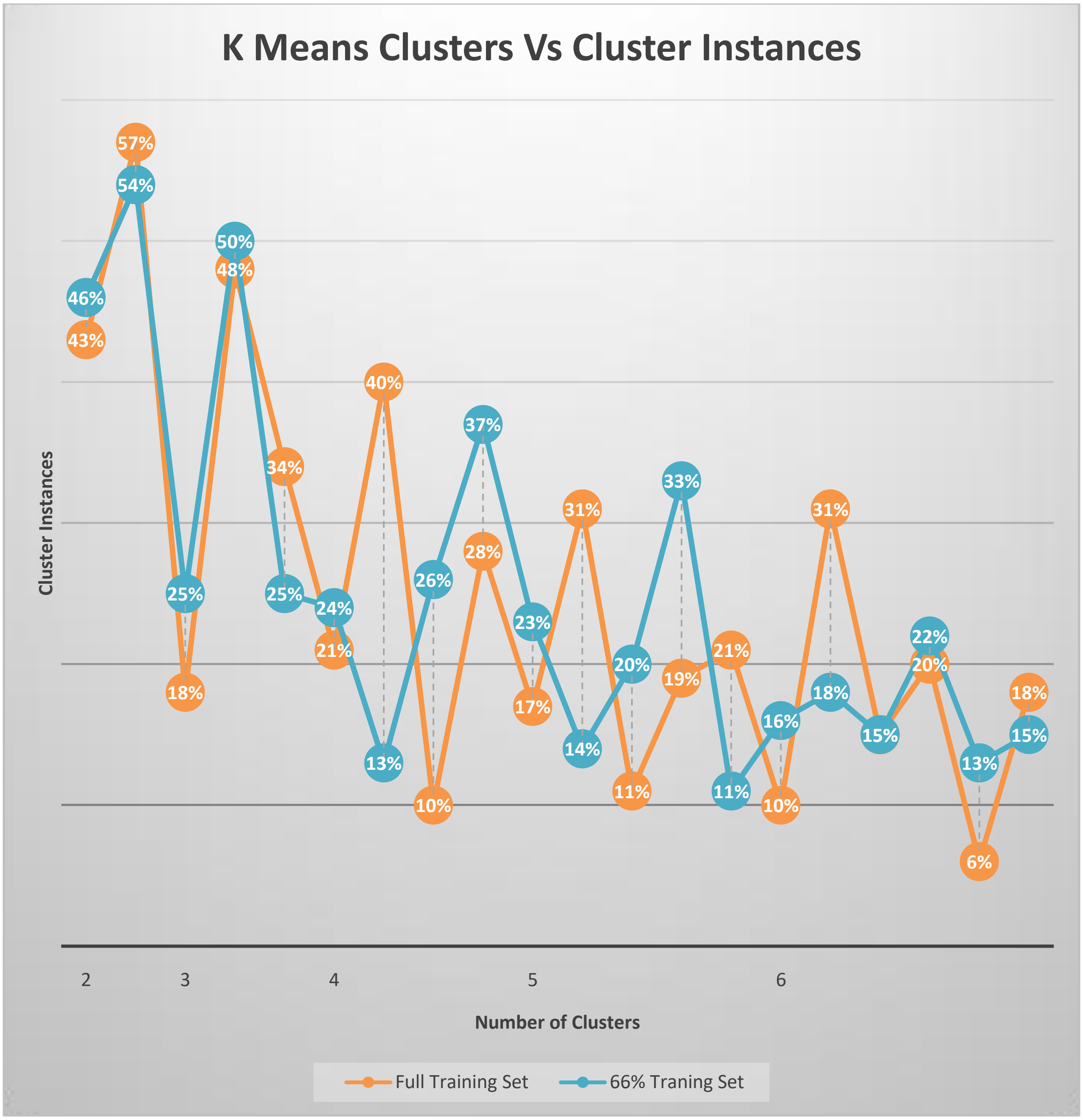

Figure 24: K Means cluster Vs Cluster Instances

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

The above diagram clearly shows that number of cluster is 2, this model has 43% & 57% of cluster instances for full training set and 46% & 54% of cluster instances for 66% training set, if the number of cluster is 3, this model has 18% 48% & 34% of cluster instances for full training set and 25%,50% & 25% of cluster instances for 66% training set, if the number of cluster is 4, this model has 21%,40%,10% & 28% of cluster instances for full training set and 24%,13%,26% and 37% of cluster instances for 66% training set, If the number of cluster is 5, this model has 17%,31%,11%,19% & 21% of cluster instances for full training set and 23%,14%,20%,33% &11% of cluster instances for 66% training set, If the number of cluster is 6, this model has 10%,31%,15%,20%,6% &18% of cluster instances for full training set and 16%,18%,15%,22%,13% &15% of cluster instances for 66% training set.Based on the time consumption the system recommends that cluster 2, 3 and 5 have zero second taken the time consumption for build the model in 66% training set. 0.01 seconds for cluster 6 and 0.03 seconds for cluster 4 in 66% training set models. Cluster 5 and 6 have low sum of squared errors for full training and 66% training set comparatively other models.

Conclusion

Finally this work concludes that when the proposed model has 6 clusters, it has more number of iteration to build the model like full training set has 7 iterations and 44% testing test has 9 iterations with 572.62 sum of squared error for full training set and 355.02 for 44% test set. It has taken the time to build the model 0.02 seconds for full training set and 0.01 second for 66% training set. This model produces the low sum of squared errors comparatively other models.

References

[1] G. Ayyappan ,K.Sivakumar, Heart Disease Data Set Classifications: Comparisons Of Correlation Co Efficient By Applying Various Parameters In Gaussian Processes, Indian Journal of Computer Science and Engineering (IJCSE) , Vol. 9 No. 5 Oct-Nov 2018, Page Number130-134, e-ISSN : 0976-5166, p-ISSN : 2231-3850.

[2] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., &Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.

[3] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database.

[4] S. Mohan, C. Thirumalai, G. Srivastava, 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. J. IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[5] Chandna, Deepali, 2014. Diagnosis of Heart Disease Using Data Mining Algorithm. International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1678-1680.

[6] Rani, K., 2011. Analysis Of Heart Diseases Dataset Using Neural Network Approach. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 4269 - 4289
Received 25 April 2021; Accepted 08 May 2021.

[7] Karthiga, A. Sankari, M. Safish Mary, M. Yogasins, 2017. Early Prediction of Heart Disease Using Decision Tree Algorithm. International Journal of Advanced Research in Basic Engineering Sciences and Technology 3.3 (2017).

[8] C. Sowmiya, P. Sumitra, 2017. Analytical study of heart disease diagnosis using classification techniques. IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303115.

[9] Bahadur, Shamsher, 2013. Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques. IOSR Journal of Agriculture and Veterinary Science. 4. 60-64. 10.9790/2380-0426164.

[10] G. Ayyappan ,K.Sivakumar, Heart Disease Data Set Classifications: Comparisons Of Correlation Co Efficient By Applying Various Parameters In Gaussian Processes, Indian Journal of Computer Science and Engineering (IJCSE) , Vol. 9 No. 5 Oct-Nov 2018, Page Number135-140, e-ISSN : 0976-5166, p-ISSN : 2231-3850.

[11] Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61.

[12] https://www.kaggle.com/mruanova/predict-heart-disease-using-random-forests#Random-Forest-Classifier

[13] https://www.kaggle.com/nyjoey/heart-disease

[14] https://towardsdatascience.com/exploratory-data-analysis-on-heart-disease-uci-data-set-ae129e47b323

[15] C. Sowmiya, P. Sumitra, 2017. Analytical study of heart disease diagnosis using classification techniques. IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303115.

[16] Parthiban, G., Srivatsa, Shesh, 2012. Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. International Journal of Applied Information Systems. 3. 25-30. 10.5120/ijais12-450593.

[17] Cömert, Z., A. F. Kocamaz, 2017. Comparison of machine learning techniques for fetal heart rate classification. Acta Phys. Pol. A 132.3 (2017): 451-454.

[18] Patel, Jaymin, Tejalupadhyay, Samir, Patel,Samir, 2016. Heart Disease Prediction using Machine learning and Data Mining Technique. International Journal of Computing Science and Communication10.090592/IJCSC.2016.018.

[19] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, J. Gutierrez, 2017. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.