Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 173 (2020) 130–139

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on *Smart Sustainable Intelligent Computing and Applications* under ICITETM2020

Heart Disease Prediction using Exploratory Data Analysis

R. Indrakumari[a], T. Poongodi[b], Soumya Ranjan Jena[c]

[a,c]*Assistant Professor,* [b]*Associate Professor*
School of Computing Science & Engineering, Galgotias University, Greater Noida, U.P., India

Abstract

Healthcare industries generate enormous amount of data, so called big data that accommodates hidden knowledge or pattern for decision making. The huge volume of data is used to make decision which is more accurate than intuition. Exploratory Data Analysis (EDA) detects mistakes, finds appropriate data, checks assumptions and determines the correlation among the explanatory variables. In the context, EDA is considered as analysing data that excludes inferences and statistical modelling. Analytics is an essential technique for any profession as it forecast the future and hidden pattern. Data analytics is considered as a cost effective technology in the recent past and it plays an essential role in healthcare which includes new research findings, emergency situations and outbreaks of disease. The use of analytics in healthcare improves care by facilitating preventive care and EDA is a vital step while analysing data. In this paper, the risk factors that causes heart disease is considered and predicted using K-means algorithm and the analysis is carried out using a publicly available data for heart disease. The dataset holds 209 records with 8 attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain. To predict the heart disease, K-means clustering algorithm is used along with data analytics and visualization tool. The paper discusses the pre-processing methods, classifier performances and evaluation metrics. In the result section, the visualized data shows that the prediction is accurate.

Keywords: Tableau; Exploratory Data Analysis; K-means Clustering Algorithm.

1. Introduction

A study in 2016 found that human beings are collectively generated data more than ten exabytes, or $5x10^{18}$ bytes from various sources (Lyman and Varian 2003). Exploratory Data Analysis (EDA) is a method to analyze data

10.1016/j.procs.2020.06.017

using advanced techniques to expose hidden structure, enhances the insight into a given dataset, identifies the anomalies and builds parsimonious models to test the underlying assumptions. Exploratory Data Analysis (EDA) is classified into Graphical or non-graphical and Univariate or multivariate Univariate data consider one data column at a time while multivariate method considers more than two variables while analyzing. The diagnostic methods of diseases are of two types namely, Invasive and Non-invasive

Invasive diagnostic method includes incise procedures in which instruments are used to cut the skin, mucus membrane and connective tissues. In contrast, non-invasive methods are used to diagnose diseases without opening the skin. Some of the machine learning algorithms based on non-invasive methods are Support Vector Machine (SVM), K- means clustering, K-Nearest Neighbour (KNN), Artificial Neural Network (ANN), Naive Bayes, Logistic Regression and rough set [15].

Predicting and diagnosing heart disease is the biggest challenge in the medical industry and it is based on factors like physical examination, symptoms and signs of the patient [1-3]. Factors which influence heart diseases are cholesterol level of the body, smoking habit, and obesity, family history of diseases, blood pressure and working environment. Machine learning algorithms play a vital and accurate role in predicting heart disease [4]. The advancement of technologies allows machine language to pair with big data tools to handle unstructured and exponentially growing data [5]. In the paper, K means clustering method is proposed in big data environment and the visualization is made with the tableau dashboard.

2. Heart Diseases

Heart disease is perceived as the deadliest disease in the human life across the world. In particular, in this type of disease the heart is not capable in pushing the required quantity of blood to the remaining organs of the human body in order to accomplish the regular functionalities [6]. Some of the symptoms of heart disease include physical body weakness, improper breathing, swollen feet, etc. The techniques are essential to identify the complicated heart diseases which results in high risk in turn affect the human life [7]. Presently, diagnosis and treatment process are highly challenging due to inadequacy of physicians and diagnostic apparatus that affect the treatment of heart patients [8]. Early diagnosis of heart disease is significant to minimize the heart related issues and to protect it from serious risks [9]. The invasive techniques are implemented to diagnose heart diseases based on medical history, symptom analysis report by experts, and physical laboratory report. Moreover, it causes delay and imprecise diagnosis due to human intervention. It is time consuming, computationally intensive and expensive at the time of assessment [10].

Heart disease can be predicted based on various symptoms such as age, gender, pulse rate etc. Data analysis in healthcare assists in predicting diseases, improving diagnosis, analyzing symptoms, providing appropriate medicines, improving the quality of care, minimizing cost, extending the life span and reduces the death rate of heart patients. ECG (Electro Cardio Gram) helps in screening irregular heart beat and stroke with the embedded sensors by resting it on a chest in order to track the patient's heart beat. Heart disease prediction is being done with the detailed clinical data that could assist experts to make decision. Human life is highly dependent on proper functioning of blood vessels in the heart. The improper blood circulation causes heart inactiveness, kidney failure, imbalanced condition of brain, and even immediate death also. Some of the risk factors that can cause heart diseases are obesity, smoking, diabetes, blood pressure, cholesterol, lack of physical activities and unhealthy diet.

Acute Myocardial Infarction (AMI) is the cardiovascular disease that happens due to interruption in the blood flow or circulation in the heart muscle, causes heart muscle to become necrotic (damage or die) [11]. The primary reason for this disease is the blockage means that the blood flow to the heart muscle become obstructed or reduced. If the blood flow is reduced or obstructed, the functioning of red blood cells that carries enough oxygen helps in sustaining consciousness and human life have a severe impact. Without oxygen supply for 6 to 8 minutes, heart muscle may get arrest that in turn resulted in patient's death. The significant cause of the cardiovascular disease is 'plaque' means a hard substance formed in the coronary arteries which is made up of cholesterol (fat), causes the blood flow to be reduced or obstructed. Sometimes, it can be formed in the arteries known as atherosclerosis and

investigating the cause of it are determined as a chronic inflammation. The increase in the amount of white blood cells causes inflammation and other subsequent disorders such as stroke or reinfarction [12]. Generally, there are two stages of wound healing in terms of monocytes and macrophages, namely, inflammatory and reparative stages. However, the two stages are compulsory for proper wound healing and if the inflammation is continued too long, then it leads to heart failure.

An unusual type of heart disease is the acute spasm or contraction in the coronary arteries. The spasms become visible in arteries suddenly with no symptom of atherosclerosis [13]. It blocks the blood flow that causes oxygen deprivation in the heart. Male genders are more likely to experience heart attack than females. Moreover, women can experience pain more than an hour and the duration to experience the pain of men is normally less than an hour. The cardiovascular disease has an impact in the complete physiological system, not only in the heart; changes occur everywhere that too in the remote organs such as bone marrow and spleen [14].

3. Materials and Methods

The analysis is carried out using a publicly available data for heart disease. The dataset holds 209 records with 8 attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain. The dataset is analysed with visualization tool tableau and K means clustering.

3.1 Dataset

The dataset to define the proposed algorithm is the Cleveland heart disease raw dataset with 76 features of 303 patients. During the pre-processing method, some samples are removed to eradicate error due to inconsistency of data. The prediction of heart disease is made with 209 samples with seven independent features like age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain and the habitual of physical exercise. Age is considered as the main risk factor for heart diseases as coronary fatty streaks develops in the adolescence stage. Male are at higher risk of coronary diseases than females, hence the data set considered here is for only male. Angina is the discomfort caused when the muscles of heart is not supplied with sufficient oxygen rich blood. High blood pressure is one of the major causes of heart disease as it damages arteries. Blood pressure combined with diabetes can increase the risk even more. Heart rate with high blood pressure increases the risk of heart diseases. Heart beat rate is directly proportional to the risk of coronary disease. The symptom of heart disease includes feeling gripping and tight usually on the chest but spread to shoulders up to the stomach. The types of angina are atypical angina, typical angina, asymptomatic and non-anginal pain.

Table 1: Features information and description of Cleveland heart disease dataset (**Source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease**)

Sl.No	Feature Name	Feature code	Description	Domain of Value
1	Age	Age	Age of the person in years	28 <age< 66
2	Type of chest pain	chest_pain	1. atypical angina 2. typical angina 3. asymptomatic 4. non-anginal pain	1 2 3 4
3	Resting blood pressure	rest_bpress	mm Hg	92 to 200
4	Fasting blood sugar	blood_sugar	Fasting blood sugar >120 mg/dl	t = true f = false

5	Resting electrocardiographic results	rest_electro	1. left_vent_hyper 2. normal 3. st_t_wave_abnormality	
6.	Maximum heart rate achieved	max_heart_rate		82 to 188
7.	Exercise-induced angina	exercise_angina	1. Yes 2. No	

3.2 K-means Clustering Algorithm

K-means clustering is an unsupervised class of machine learning algorithm. Usually, unsupervised algorithms project the desired output without referring any value. In K-means clustering algorithm, the data are clustered in such a way that it has highest intra-class similarity and minimal inter-class similarity. This algorithm lessens the sum of squares distance from the centroid within the cluster. The algorithm divides the data into k clusters with a centroid. K-means iteratively finds the centre that reduces the distance among individual points in a cluster and the cluster centre. The following flow chart shows the working of k-means clustering algorithm.

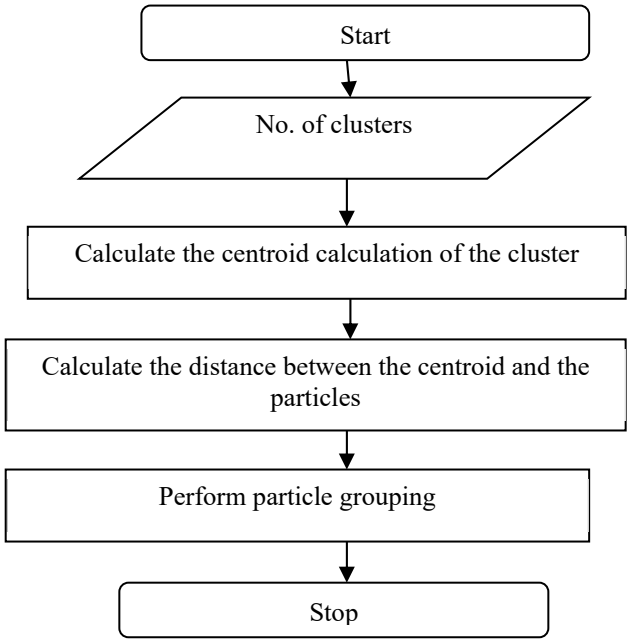

Fig.1 Workflow of K-means algorithm

3.3 Tableau

Tableau is one of the business intelligence software used to analyse data and visualize the insights in the form of graph and charts. User can develop and share an interactive dashboard which shows the hidden pattern, trends, density and variation of data. Tableau uses centroid-based k-means clustering algorithm that divides the data into K-number of clusters. Dashboards are created with the data set after applying K-means algorithm. It provides visual appealing clusters in order to predict the occurrence of heart disease from the given dataset.

4. Results and Discussion

The result of the data analysis to identify the necessary hidden patterns for predicting heart diseases are presented in this section. Here the variables considered to predict the heart disease are age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain and exercise angina. The heart disease dataset is effectively pre-processed by eliminating unrelated records and given values to missing tuples. The pre-processed heart disease data set [10] is then composed by K-means algorithm. Here, four types of heart diseases are discussed namely asymptomatic pain, atypical angina pain, non-anginal pain and non-anginal pain. The results are computed using all the four types of chest pain with other deciding variables.

Data Analysis

Histogram given in Fig.2 depicts the distribution of ages and the risk of heart disease for the targeted class. It is observed that target class with the age ranging from 50 to 55 is having high risk of heart disease as the development of coronary fatty streaks starts in this age range.

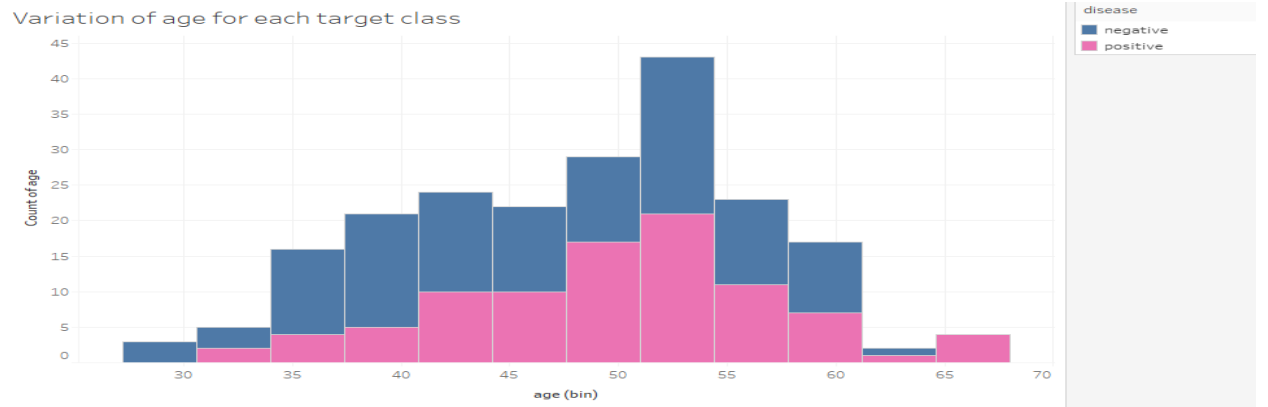

Fig.2 Histogram of variation of age for each target class

Fig.3 shows the consequences of diabetic target class with the maximum heart rate. It is observed from the colour code that the target class with diabetics' population is represented by blue colour and the red colour indicates the population without diabetics. Target class with diabetics and acceptable heart rate are showing negative symptom.

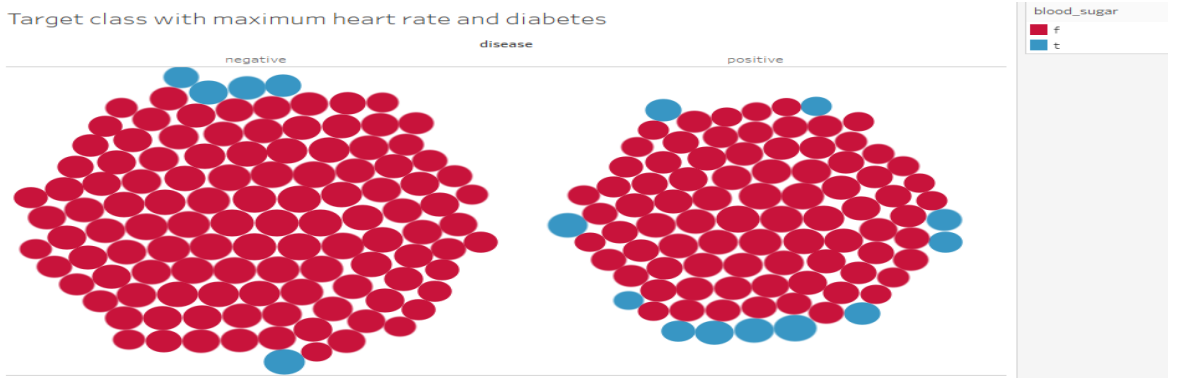

Fig.3 Target class with maximum heart rate and diabetics

Fig. 4 shows the impact of blood pressure and sugar in heart disease. It is inferred that population with diabetics and high blood pressure is expected to get heart disease.

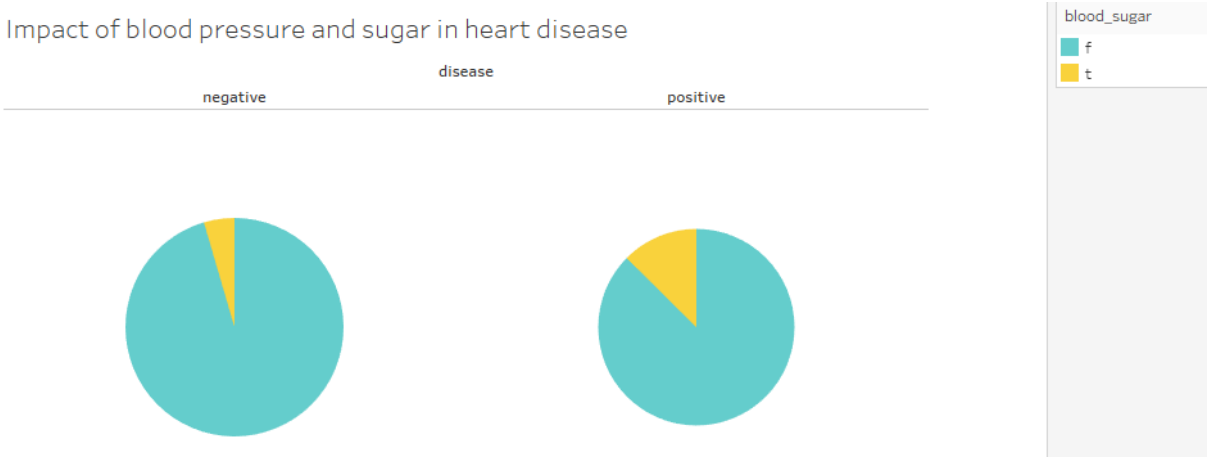

Fig.4 Impact of blood pressure and sugar in heart disease

Fig.5 shows the user defined filter to predict heart disease and it is applied on the type of chest pain, range of blood pressure and maximum heart rate. The filters applied on dimension category is called categorical filter and the filter applied to measures are called quantitative filter. Here the chest pain type is categorical filter and the blood pressure, heart rate comes under quantitative filter. With the help of slider the user can change the measurement and type to predict the heart disease.

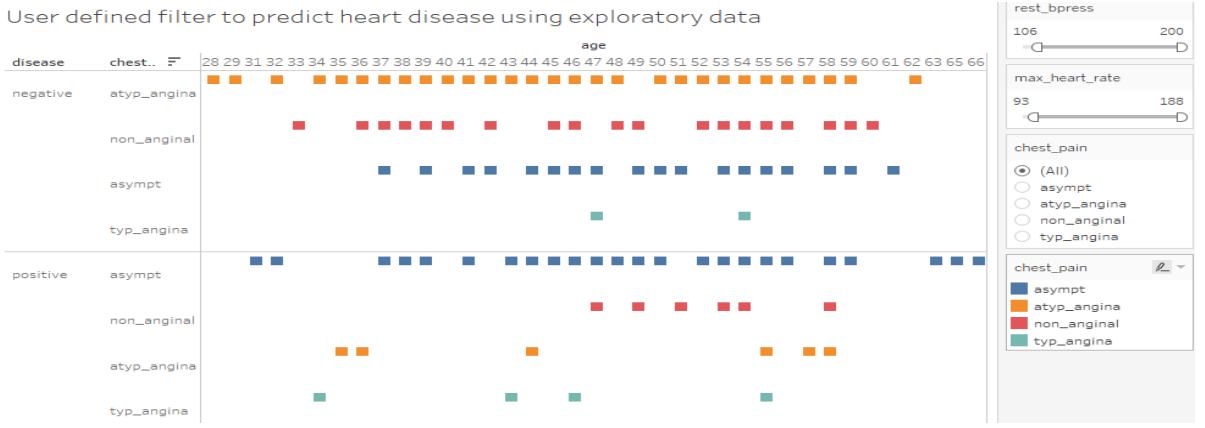

Fig.5 User defined filter to predict heart disease using exploratory data

K-means Clustering

K-means clustering algorithm is selected because of its efficiency, simplicity, capacity to produce even sized population and scalability in handling the web dataset to produce accurate output. K-means algorithm have minimum sum of squares to categorize clusters of data points. Here the dataset has 209 observations of 7 variables. The initial center of cluster is computed with the following steps.

i) Identify random K clusters
ii) Iteratively find the significant clusters
iii) If the distance between the observation and its nearest cluster center is higher than the distance among other closest cluster centers then the observation is replaced with nearest center by calculating Euclidean distance among the cluster and the observation.
iv) Within cluster sum of squares is calculated as:

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

where S_k is the set of observations in the kth cluster and $\bar{x}_{kj}$ is the jth variable of the cluster center for the kth cluster.

The iteration will get stop if the difference between the sum of squares in two successive iterations is minimal and this is called Final Cluster Centers.

The variables considered to predict the heart disease are age, maximum heart rate, chest pain type and disease. Here, four types of chest pains are considered and the results are discussed individually.

Chest Pain Type: Asymptomatic

The Fig.6 shows the plot of Age vs. Max Heart Rate broken down by Disease. Colour shows details about disease. The screen shot of the clustering are described below.

Summary of Diagnostics

No. of Clusters	:	2
No. of Points	:	102
Between-group Sum of Squares	:	20.285
Within-group Sum of Squares	:	9.5649
Total Sum of Squares	:	29.85

Table 2: Chest Pain Type: Asymptomatic

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster1	75	49.853	124.03	Positive
Cluster2	27	48.556	136.59	Negative

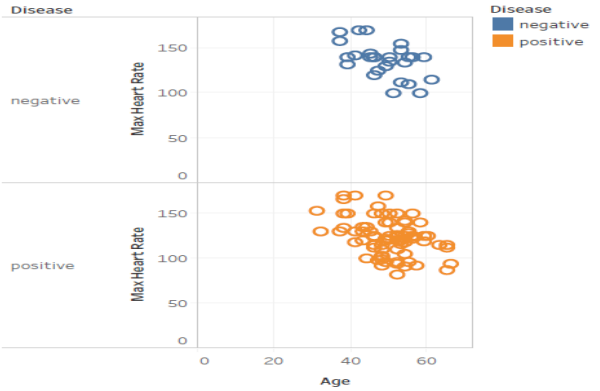

Fig.6 Age vs. Max Heart Rate broken down by Disease with asympt chest pain type

Chest Pain Type: Atypical Angina

Summary of Diagnostics

No. of Clusters	:	2
No. of Points	:	65
Between-group Sum of Squares	:	5.5109
Within-group Sum of Squares	:	8.3246

Total Sum of Squares : 13.835

Table 3: Chest Pain Type: Atypical Angina

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster1	59	45.492	147.47	Positive
Cluster2	6	47.5	139.5	Negative

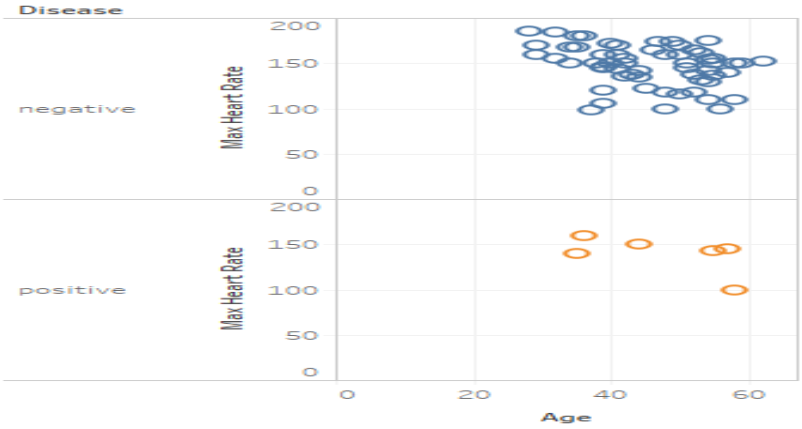

Fig.7 Age vs. Max Heart Rate broken down by Disease with atypical angina chest pain type

Chest Pain Type: Non-Angina

Summary of Diagnostics
No. of Clusters : 3
No. of Points : 36
Between-group Sum of Squares : 8.89
Within-group Sum of Squares : 2.251
Total Sum of Squares : 11.141

Table 4: Chest Pain Type: Non Angina

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster 1	15	39.533	162.8	Negative
Cluster 2	14	54.571	133.43	Negative
Cluster 3	7	52.857	140.29	Positive

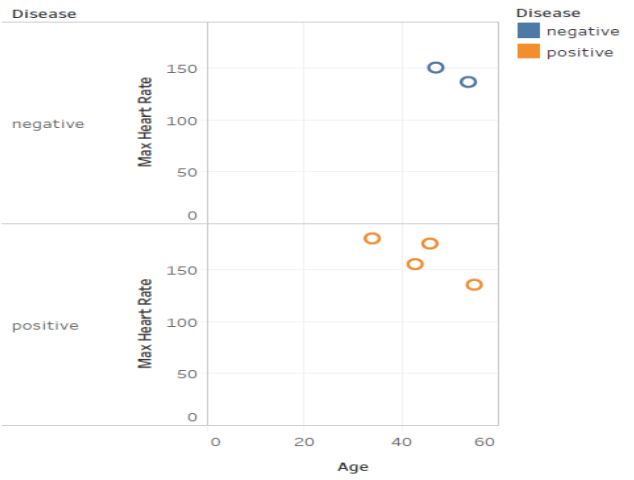

Fig. 8 Age vs. Max Heart Rate broken down by Disease with non_angina chest pain type

Chest Pain Type: Typical_Anginal Pain

Summary of Diagnostics

No. of Clusters	:	3
No. of Points	:	6
Between-group Sum of Squares	:	2.3779
Within-group Sum of Squares	:	0.52542
Total Sum of Squares	:	2.9033

Table 5: Chest Pain Type: Typical_Anginal Pain

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster 1	2	40.0	177.5	Positive
Cluster 2	2	49.0	145.5	Positive
Cluster 3	2	50.5	143.5	Negative

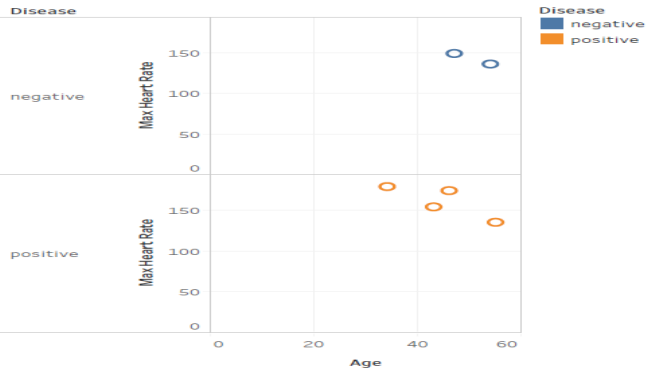

Fig. 9 Age vs. Max Heart Rate broken down by Disease with typical_angina chest pain type

From the above clusters it is inferred that age, maximum heart rate and the chest pain type plays a vital role in predicting the heart disease.

5. Future Advances and Conclusion

Heart stroke and vascular disease are the major cause of disability and premature death. Chest pain is the key to recognize the heart disease. In this work, the heart diseases are predicted by considering major factors with four types of chest pain. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Here the datasets are clustered and based upon the clusters the happening of chest pain is predicted. The role of exploratory data using tableau provided a visual appealing and accurate clustering experience.

References

[1] V. Manikantan & S.Latha,"Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering, Volume-2, Issue-2, pp.5-10, 2013.

[2] Dr.A.V.Senthil Kumar, "Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering", International Journal of Advanced Trends in Computer Science and Engineering, Volume-4, No.6, pp.07-18, 2015.

[3] Uma.K, M.Hanumathappa, "Heart Disease Prediction Using Classification Techniques with Feature Selection Method", Adarsh Journal of Information Technology, Volume-5, Issue-2, pp.22-29, 2016

[4] Himanshu Sharma, M.A.Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms:A Survey",International Journal on Recent and Innovation Trends in Computing and Communication,Volume5,Issue-8,pp.99-104, 2017.

[5] S.Suguna, Sakthi Sakunthala.N ,S.Sanjana, S.S.Sanjhana, "A Survey on Prediction of Heart Disease using Big data Algorithms", International Journal of Advanced Research in Computer Engineering & Technology,Volume-6,Issue-3,pp.371-378,2017.

[6] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," Nature Reviews Cardiology, vol. 8, no. 1, pp. 30–41, 2011.

[7] J.Mourão-Miranda,A.L.W.Bokde,C.Born,H.Hampel,and M. Stetter, "Classifying brain states and determining the discriminatingactivationpatterns:supportvectormachineon functionalMRIdata,"NeuroImage,vol.28,no.4,pp.980–995, 2005.

[8] S.Ghwanmeh,A.Mohammad,andA.Al-Ibrahim,"Innovative artificial neural networks-based decision support system for heartdiseasesdiagnosis,"JournalofIntelligentLearningSystems and Applications, vol. 5, no. 3, pp. 176–183, 2013.

[9] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal of Computer Science Issues, vol. 8, no. 2, pp. 150–154, 2011.

[10] K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," International Journal of Computer Applications, vol. 19, no. 6, pp. 6–12, 2011.

[11] Al Mamoon I, Sani AS, Islam AM, Yee OC, Kobayashi F, Komaki S (2013) A proposal of body implementable early heart attack detection system, 1-4.

[12] Patterson K (2016) Matthias Nahrendorf. Circ Res 119: 790-793.

[13] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[14] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world congress on engineering and computer science (Vol. 2, pp. 22-24).

[15] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early heart disease prediction using data mining techniques," in *Proceedings of Computer Science & Information Technology (CCSIT-2014)*, vol. 24, pp. 53–59, Sydney, NSW, Australia, 2014.