Big Data Analysis of Heart Disease

Authors: Dr.Sahar AbdulRahman Ismail, Alanoud BinMohareb, Dalal Alobaili, Ghada Alarwan, Nora Alabdalwahab, Rawan Albishr,Walaa Alzahrani

Princess Nora University, Computer Science Department
Riyadh, Kingdom of Saudi Arabia

***Abstract*— In late years the total of public biomedical data has increased dramatically. The increasing movement to electronic data platforms for the management of health information is creating an untapped resource with the power to change health care. These Big Data open up possibilities for advanced quality health care, better and faster clinical research. A key problem with Big Data is understanding the information rapidly, so by applying a visual analytics approach, the initial overwhelming scale of Big Data becomes a valuable asset. Interactive visualization permits everyone to understand large and multi-source data. In addition, predicting and diagnosing of heart disease become a challenging factor faced by doctors and hospitals, and it is the main reason of deaths in many countries. Thus, this project aims to develop a visualization method for heart disease to help in predicting and diagnosing, by applying some analysis algorithms on the heart disease dataset and display the result to the user in understandable graphical representation.**

Keywords*— *Biomedical Data Mining, Biomedical Data Visualization, Health Care Big Data Mining

I. INTRODUCTION

Medical big data have several distinctive features that are different from big data from other disciplines. Medical big data are frequently hard to access and most investigators in the medical arena are hesitant to practice open data science for reasons such as the risk of data misuse by other parties and lack of data-sharing incentives [1]. Medical big data are often collected based on protocols (i.e., fixed forms) and are relatively structured, partially due to the extraction process that simplify raw data [2]. Big data in healthcare is more complex, Considering the large and ever-growing repositories of biomedical data, there is a demand for systems and tools to aid in finding useful information in biomedical publications, text databases, image databases, electronic health care records, clinical notes, and other sources, including full text, to support clinical decisions [3].

.

II. PROBLEM STATEMENT AND SIGNIFICANE

In healthcare the amount of data that's being created and stored on a global level is almost inconceivable, and it just keeps growing. However, only a small percentage of data are actually analyzed. Another problem is the lack of varieties to present the collected data in a way that the common user understands, the complexity of presenting the data could be a big wall not allowing any nonprofessional user to proceed through Heart diseases are ongoing, generally incurable illnesses or conditions, such as heart attacks, Coronary Artery Disease and Heart Muscle Disorders. Heart disease is one of the most dangerous diseases which can cause death for both men and women. More than half of the deaths due to heart disease in 2009 were in men [4]. Heart attacks have several major warning signs and symptoms:

- Chest pain or discomfort.
- Upper body pain or discomfort in the arms, back, neck, jaw, or upper stomach.
- Shortness of breath.
- Nausea, light-headedness, or cold sweats [4].

.

III. PROPOSED SOLUTION

We attempt to build a computer software to process the big data in healthcare systems for heart diseases and mine useful information from it, then show it in a visualized presentation, using some big data Frameworks like Hadoop and some of its ecosystems and data mining algorithms to extract information from the medical records that can potentially give healthcare professionals, patients, and researchers useful knowledge to improve healthcare and health related decisions. Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

IV. BACKGROUND INFORMATION

A. Big Data:

Many applications now need to store and process data in time. In year 2000, volume of data stored in the world is of size 800,000 petabytes. It is expected to reach 35 zettabytes by the year 2020. However, the forecast will change with growing use of digital devices. We are storing data of several domains ranging from agriculture, environment, house holdings, governance, health, security, finance, meteorological and many more like. Just storing such data is of no use unless data are processed and decisions are made on the basis of such data. But in reality, making use of such large data is a challenge for its typical characteristics. More, the issues are with data capture, data storage, data analysis and data visualization [5].

Big data looks for techniques not only for storage, but also to extract information hidden within. This becomes difficult for the very characteristics of big data. The typical characteristics that hold it different than traditional database systems include volume, variety, velocity and value. The term volume is a misnomer for its vagueness in quantifying the size that fits to label as big data. Data that is not only huge but, expanding and holding patterns to show the order exist in the data, is generally qualifying volume of big data. A Variety of big data are due to its sources of data generation that include sensors, smart phones or social networks. The types of data emanate from these sources include video, image, text, audio, and data logs, in either structured or unstructured format. Historical database dealing with data of the past has been studied earlier, but big data are now considering data emerging ahead along the timeline and the emergence is rapid so, Velocity of data generation is of prime concern. For example, in every second large amount of data are being generated by social networks over the internet. So, in addition to volume, velocity is also a dimension of such data. Value of big data refers to the process of extracting hidden information from emerging data [5] (see Figure 1).

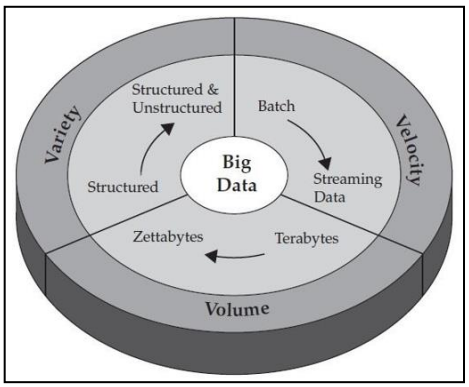


Fig. 1. V's of BigData – Variety Velocity Volume. [6]

Classification of big data from different perspectives is presented in Figure 2. The perspectives considered are data sources, content format, data stores, data staging, and data processing. The sources, generating data could be web and social media on it, different sensors reading values of parameters that changes as time passes on, internet of things, various machinery that throw data on changing sub-floor situations and transactions that are carried out in various domains such as enterprises and organizations for governance and commercial purposes. Data staging is about preprocessing of data that is required for processing for information extraction. From a data storage perspective, here the concern is about the way data stored for fast access. Data processing presents a systemic approach required to process big data [4].

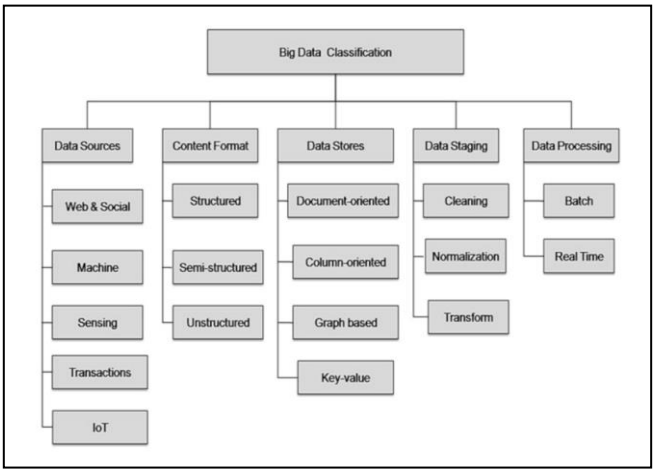


Fig. 2. Big data classification [4].

B. Big Data Processing Steps

Big data service process has few steps starting from Data acquisition, Data staging, Data analysis and application analytics processing and visualization. Figure 3 presents a framework for big data processing that models at a higher level, the working of such a system. Databases and internet-based application which store organizational data can be sources of data. When getting the data, unwanted and incomplete data will be removed in a stage that called data staging.

Then, it transforms data structure into a form that is required for analysis. In the process, it is most important to do data normalization so that data redundancy is avoided. Normalized data then are stored for processing. Big users from different domains such as social computing, bioscience, business domains and environment in space science looks forward information from gathering data. Analytics corresponding to an application are used for the purpose. These analytics being invoked in turn take the help of data analysis techniques to scoop out information hiding in big data. Data analysis techniques include machine learning, data mining and parallel algorithms for fast computation. Visualization the most important step in big data processing. Incoming data, information while in processing and result outcome are often required to visualize for understanding because structure often holds information in its folds; this is truer in genomics study [4].

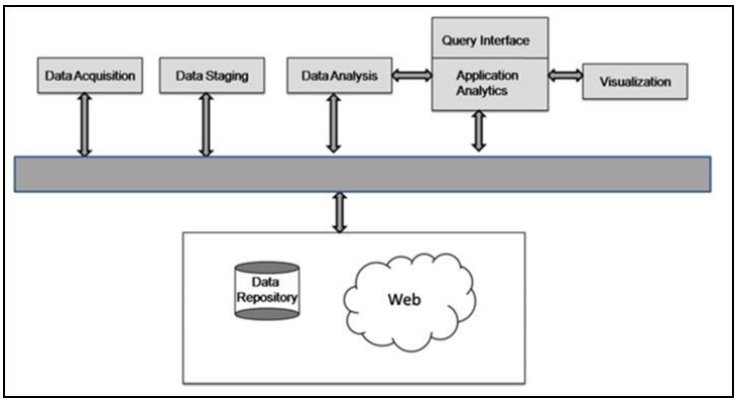


Fig. 3. Big Data Processing

C. *Advanced Analytical Theory and Methods*

1) *K-Means:*

Given a collection of objects, each with n measurable attributes, k-means is an analytical technique that, for a chosen value of k, identifies k clusters of objects based on the objects' proximity to the center of the k groups. The center is deter-mined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes. This section describes the algorithm to determine the k means as well as how best to apply this technique to several use cases. Whereas, medical Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters [7].

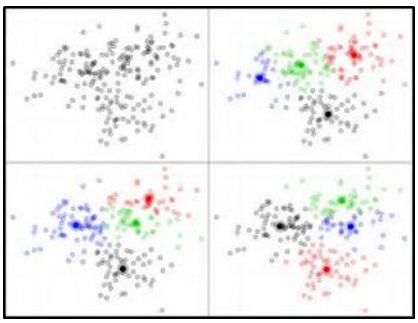


Fig. 4. Clustering sample

2) *Association Rules:*

The Apriori algorithm takes a bottom-up iterative approach to uncovering the frequent item sets by first determining all the possible. The first step of the A priori algorithm is to identify the frequent item sets by starting with each item in the transactions that meets the predefined minimum support threshold 8. These item-sets are 1-itemset denoted as. Each 1-itemset contains only one item. Next, the algorithm grows the item sets by joining onto itself to form new, grown 2-itemsets denoted as and determines the support of each 2-item set in .Those item sets that do not meet the minimum support threshold 8 are pruned away. The growing and pruning process is repeated until no item sets meet the minimum sup-port threshold. Permissively, a threshold N can be set up to specify the maximum number of items while the itemset can reach or the maximum number of iterations of the algorithm. Once completed, the output of the Apriori algorithm is the collection of all the frequent k-item sets. Algorithm, which is one of the earliest and the most fundamental algorithms for generating association rules. The A priori algorithm reduces the computational workload by only examining item sets that meet the specified minimum threshold. However, depending on the size of the dataset, the Apriori algorithm can be expensively computational cost. For each level of support, the algorithm requires a scan of the entire database to obtain the result. Accordingly, as the database expands, it takes more time to compute in each run [5].

Some approaches to improve Apriori's efficiency:

- **Partitioning:** Any item set that is potentially frequent in a transaction data-base must be frequent in at least one of the partitions of the transaction data-base.
- **Sampling:** It extracts a subset of the data with a lower support threshold then it use the subset to perform association rule mining.
- **Transaction reduction:** A transaction which not containing frequent k-item sets is useless in subsequent scans and it can be ignored.
- **Hash-based item set counting:** The k-itemset cannot be frequent If the corresponding hashing bucket count of a k-itemset is below a certain threshold.
- **Dynamic itemset counting:** When adding new candidate item sets, all of their subsets will be estimated to be frequent [5].

3) *Naïve Bayes Classifier*

This classifier is a powerful probabilistic representation, and its use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute Ai given the class label C. Classification will be done by applying Bayes rule to compute the probability of C given the particular instances of A1.....An and then predicting the class with the highest posterior probability. The main goal of classification is to predict the correct value of a designated discrete class variable given a vector of predictors or attributes. In particular, the Naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. Although the naive Bayesian (NB) algorithm is simple, it is very effective in many real-world datasets because it can give better predictive accuracy

than well-known methods like C4.5 and BP, and is extremely efficient in that it learns in a linear fashion using ensemble mechanisms, such as bagging and boosting, to combine classifier predictions. However, when attributes are redundant and not normally distributed, the predictive accuracy is reduced [9].

4) Support Vector Machine

Support vector machines exist in different forms, linear and non-linear. A sup-port vector machine is a supervised classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset per-fectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line". The best line is found by maximizing the distance to the nearest points of both classes in the training set. The maximization of this distance can be converted to an equivalent minimization problem, which is easier to solve. The data points on the maximal margin lines are called the support vectors. Most often datasets are not nicely distributed such that the classes can be separated by a line or higher order function. Real datasets contain random errors or noise which cre-ates a less clean dataset. Although it is possible to create a model that perfectly separates the data, it is not desirable, because such models are over-fitting of the training data. Overfitting is caused by incorporating the random errors or noise in the model. Therefore the model is not generic, and makes significantly more errors on other datasets. Creating simpler models keep the model from over-fitting. The complexity of the model must be balanced between fitting on the training data and being generic. This can be achieved by allowing models which can make errors. SVM can make some errors to avoid over-fitting. It tries to minimize the number of errors that will be made. Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to better overall empirical performance. Comparing the naive Bayes and the SVM classifier, the SVM has been applied the most [9].

5) Descicion Tree

The decision tree partitions its input space of a dataset into mutually exclusive regions, each of which is assigned a label, a value or an action to characterize its data points. The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made. The decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision-making unit that evaluates a decision function to determines which child node will be visiting next. The external node, on the other hand, has no child nodes and is associated with a label or value that [9].

6) K-nearst Neighbour

This classifier is considered as a statistical learning algorithm and it is very simple to implement and leaves itself open in a wide variety of variations. In brief, the training portion of nearest-neighbour does little more than store the data points presented to it. If asked to make a prediction about an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point according to some distance metric. The distance metric used in nearest neighbour methods for numerical at-tributes can be simple Euclidean distance [9].

D. Methodology

In general, the proposed method that is shown in Figure 5 will be used. The da-ta set from internal sources as electronic healthcare records distributed by the sources mentioned before. The data set was corrupted and have some messing values inside, so we cleaned it from that missing values and rearrange the data to be read-able in a proper format to import it to the Weka Tool. Where insert a heart disease records in Weka and return a potential of heart disease presence using some classification algorithms, such as Naïve base, K-means and Decision Tree [Figure 5]. Then get the resulting clusters and the potential of heart disease values and visualize it in as scaled shapes which users can choose options on it.

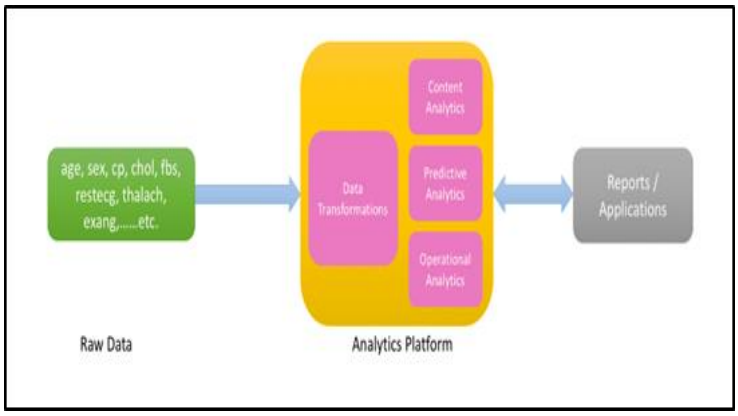


Fig. 5. Proposed Methodology

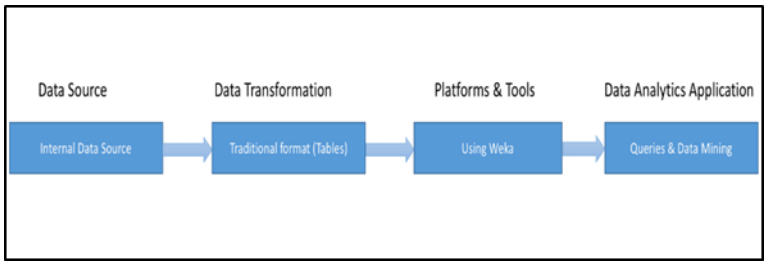


Fig. 6. The applied architecture of data analytics.

1) *WEKA Tool*

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used to it and existing algorithms can also be extended with this tool [9].

2) *Dataset Description*

This database contains 76 attributes, but all of published experiments refer to using a subset of only 14 of them in particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

A dummy value has been used insted of the name and social security number of the patients.

TABLE I. FEATURES IN THE DATASET

Feature No.	***Feature Name***	***Description***
1.	age	Patient's age in years.
2.	sex	(1=male; 0=female).
3.	cp	Chest's pain type.
4.	trestbps	Resting blood pressure (in mm Hg on admission to the hospital) .
5.	chol	Serum cholestoral in mg/dl.
6.	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false).
7.	restesg	Resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8.	thalach	Resting heart rate.
9.	exang	Exercise induced angina (1 = yes; 0 = no).
10.	oldpeak	ST depression induced by exercise relative to rest.
11.	slope	The slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping.
12.	ca	Number of major vessels (0-3) colored by flourosopy.
13.	thal	3 = normal; 6 = fixed defect; 7 = reversable defect.
14.	num	Diagnosis of heart disease (angiographic disease status) -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels).

[a.] Sample of a Table footnote. *(Table footnote)*

[b.]

Fig. 1. Example of a figure caption. *(figure caption)*

E. *Results and Discussions*

Four algorithms that are used to classify the dataset as indicated in Table 2.

TABLE II. WEKA CLASSIFIERS NAMES.

Generic Name.	***WEKA Name***	***The output***
Bayesian Network	Naïve Bayes (NB)	Figure 7
Support Vector Machine	SMO	Figure 8
C4.5 Decision Tree	J48	Figure 9
K-Nearest Neighbour	1Bk	Figure 10

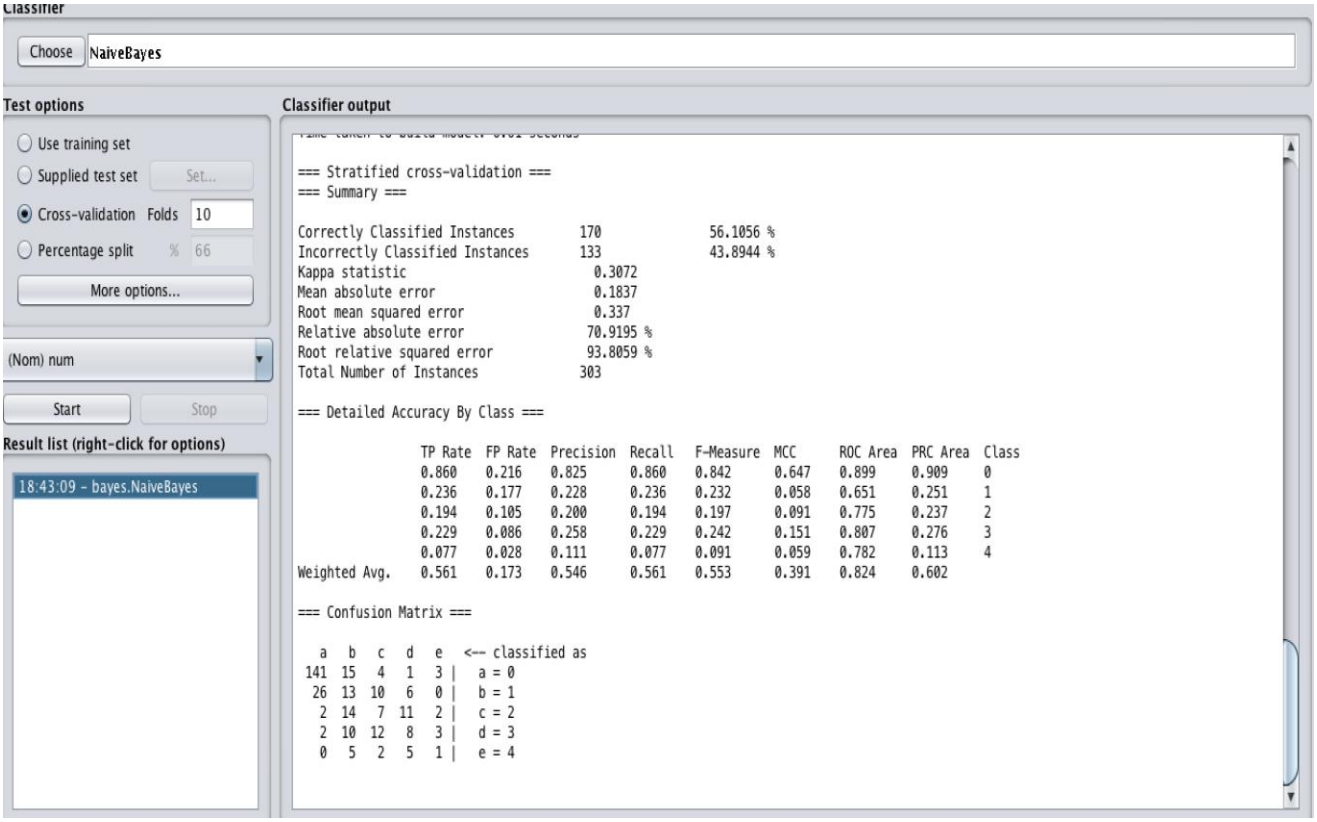


Fig. 7. Naïve bayes classifier

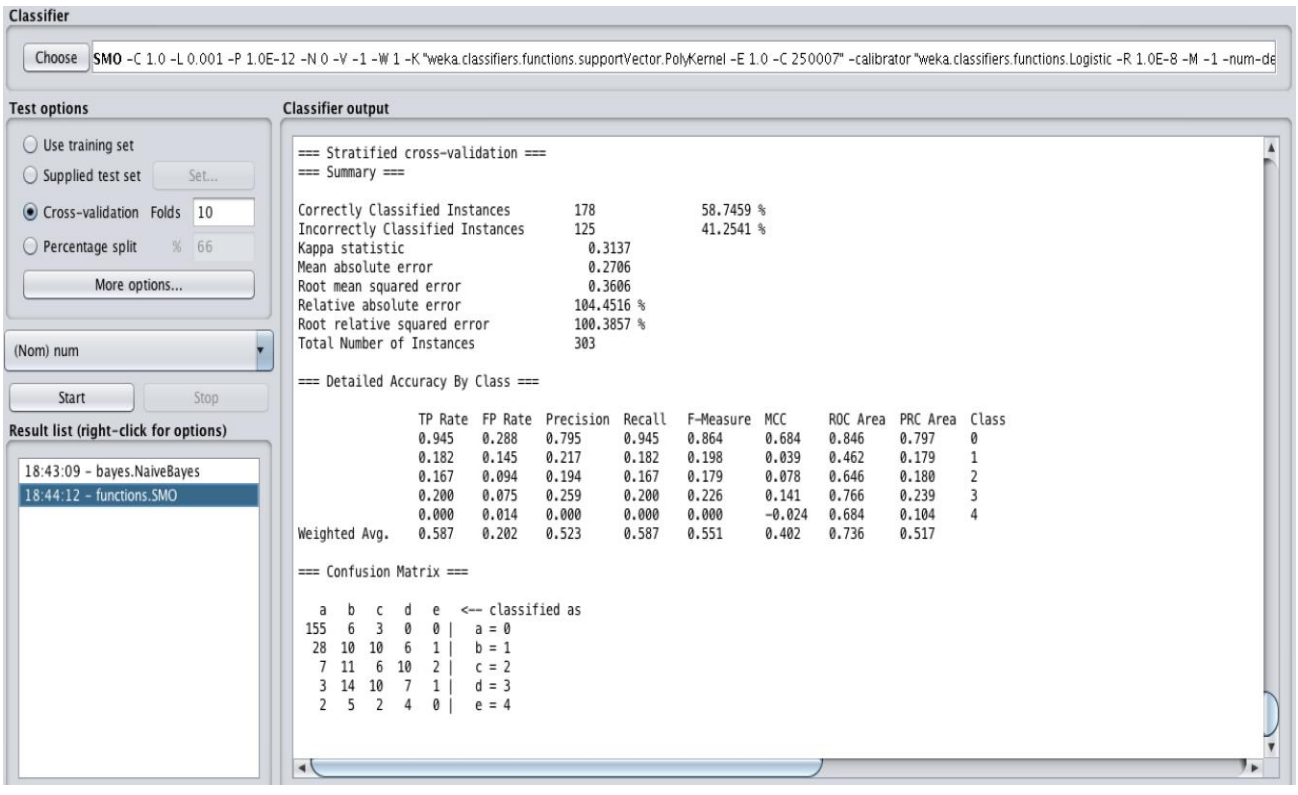


Fig. 8. Support vector machine

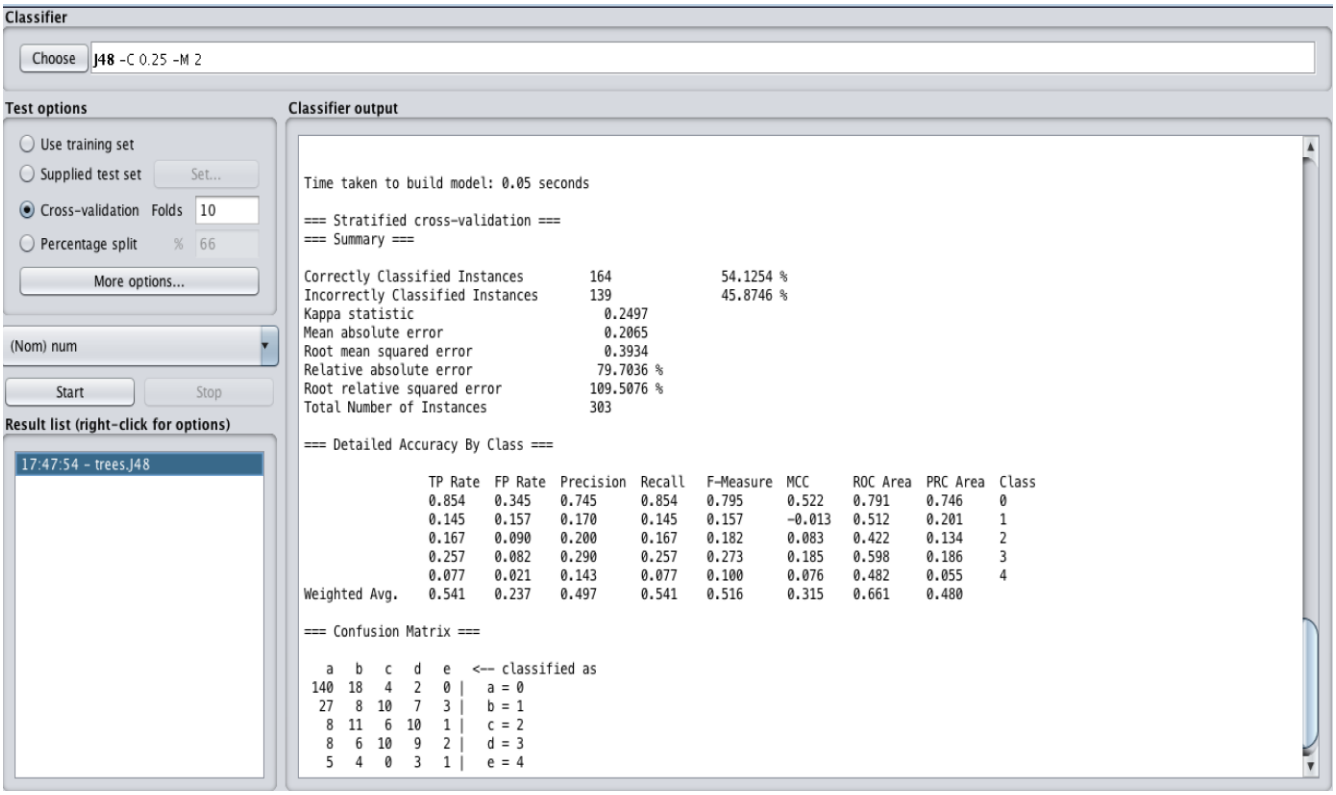


Fig. 9. Decision Tree

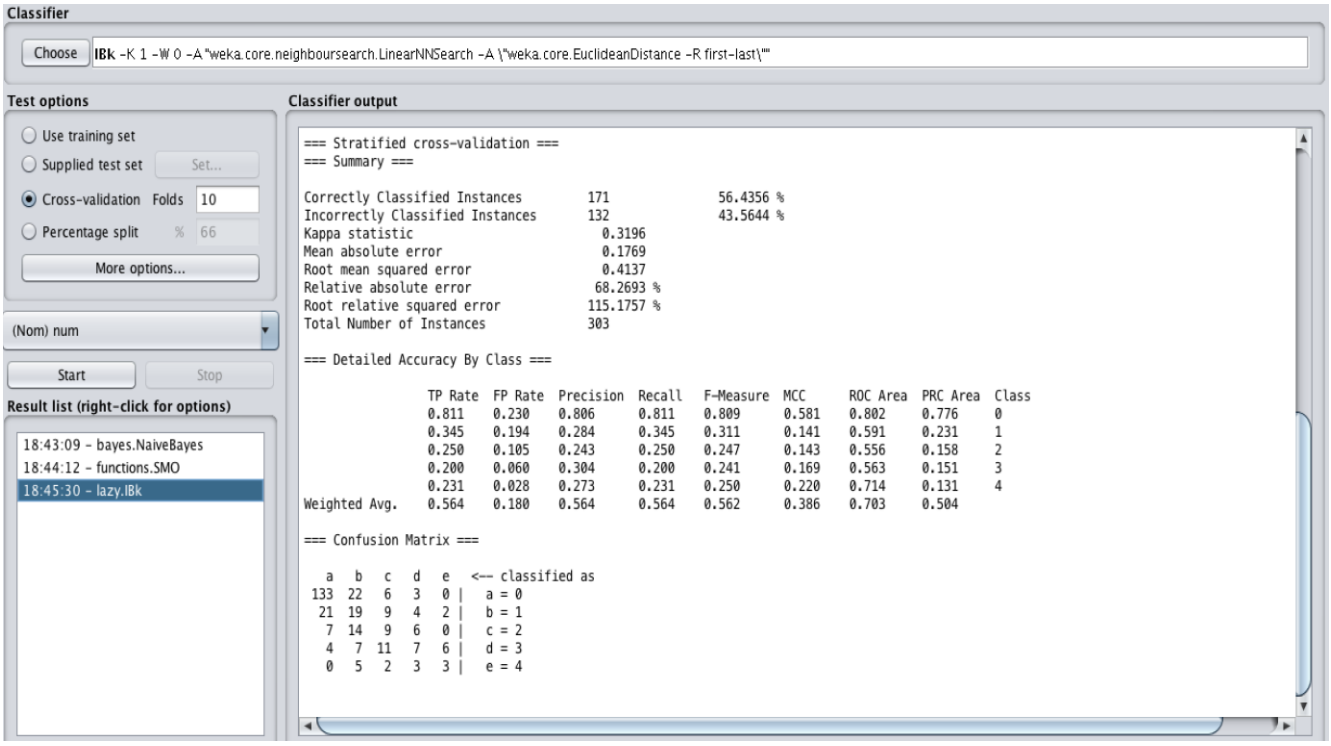


Fig. 10. K-Nearest Neighbour

1) Performance Comparasion

The performance comparison between the algorithms used an indicated in Table 3 and Figure 11.

TABLE III. PERFORMANCE COMPARISON.

Algorithm Classification	***Correct Classification Rate***	***Mis-Classification***
Naïve base classifier	56.1056	43.8944
Support Vector Machine	58.7459	41.2541
Decision Tree	54.1254	45.8746
K-Nearest Neighbour	56.4356	43.5644

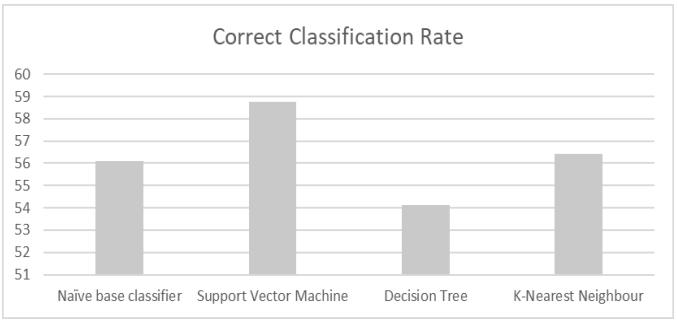


Fig. 11. Correct Classification Rate

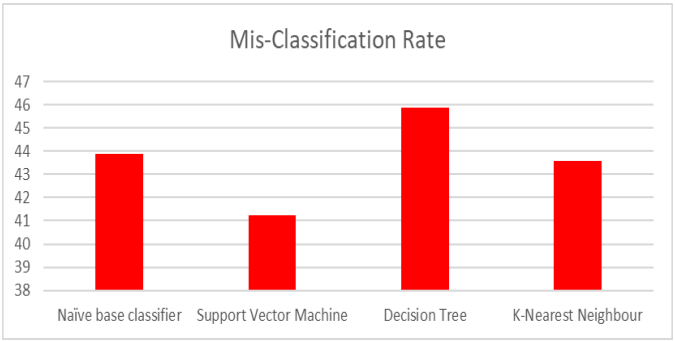


Fig. 12. Mis-Classification Rate

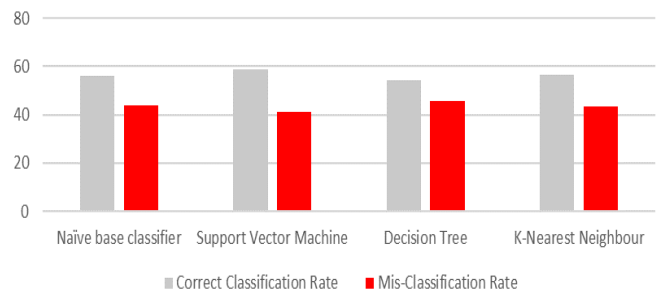
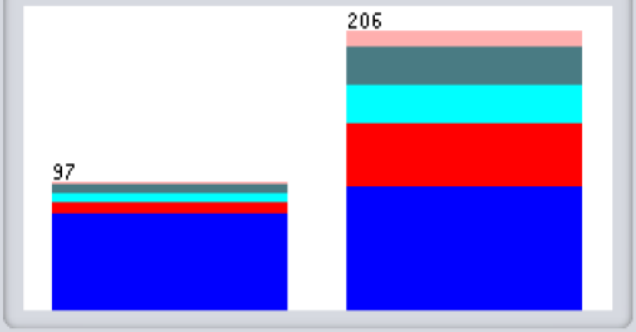
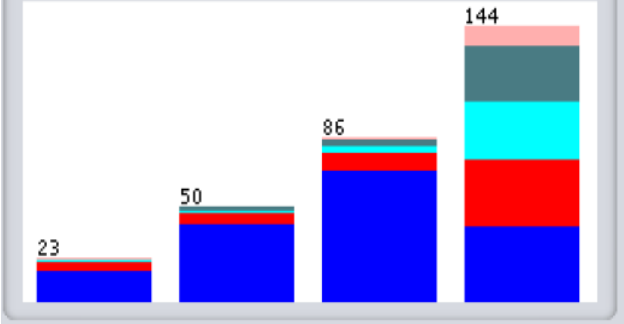


Fig. 13. Correct Classification VS. Misclassification rate

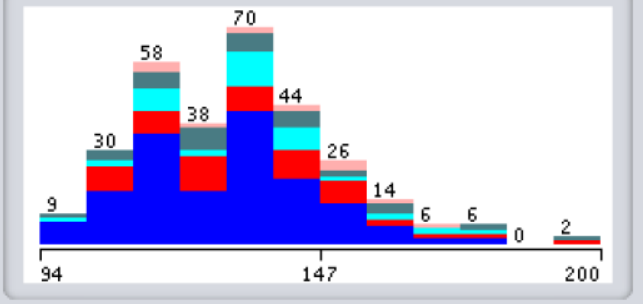
sex



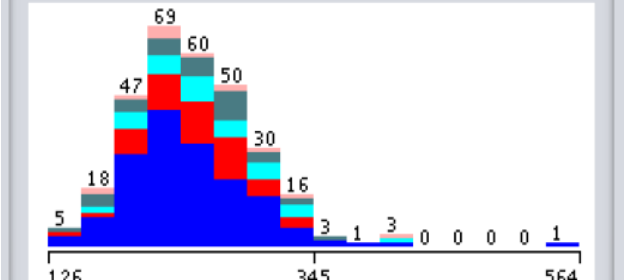
cp



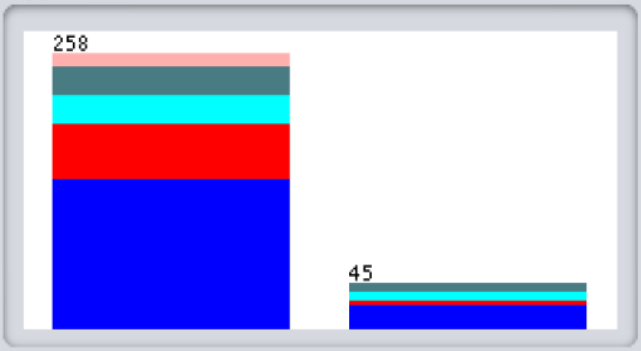
trestbps



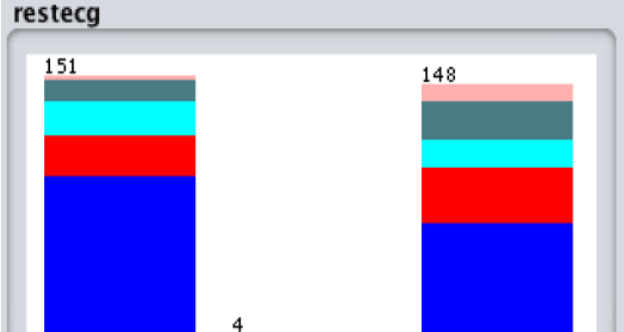
chol



fbs



restecg



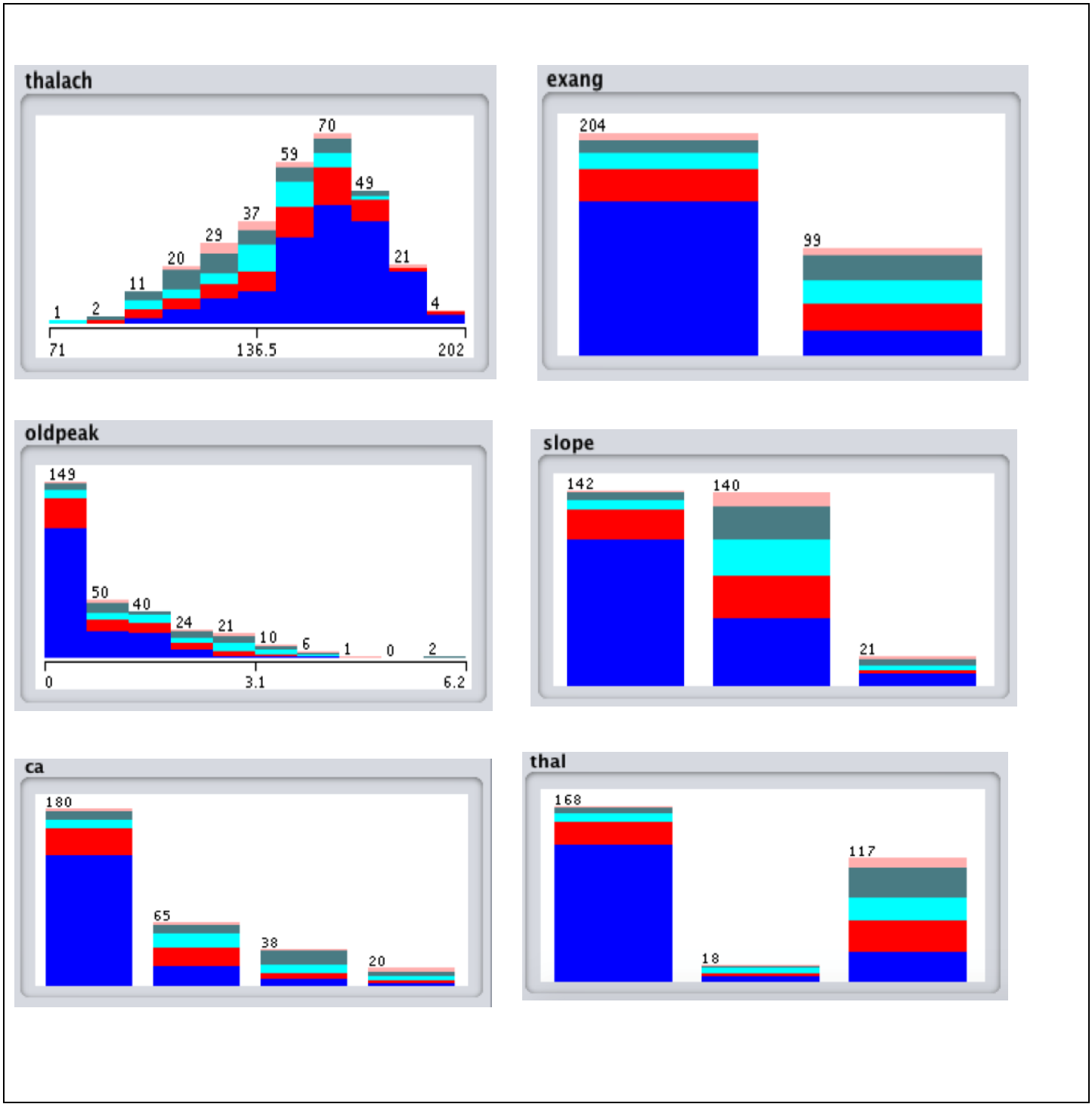


Fig. 14. Visualized Data on WEKA

2) Comparasion with other systems

System #1:

The goal of this study is to predict a patient's 30-day risk of hospital readmission and the associated cost of that readmission. They also look at the problem of estimating the cost of the hospital readmission and show how state of the art machine learning approaches outperform the statistical methods which were previously used in the cost prediction domain [8].

System #2:

Prediction and diagnosing of heart disease become a challenging factor faced by doctors and hospitals both in India and abroad. Machine learning and four popular classifiers are selected and applied to the same dataset to analyze and visualize an accurate result. After that, the doctors can take the decisions [9].

Strengths:

- It uses 4 classifiers and each one has a different way to analyze so, the results will be accurate results, which can make it very helpful.
- Graphical representation of the results.

System #3:

The huge amount of data in Medical makes the classification produce less accurate results and the heart disease is the leading cause of death in INDIA. So, there is a need to define a decision support system that helps clinicians decide to take precautionary steps. They propose a new algorithm which combines KNN with genetic algorithm for effective classification [10].

CONCLUSION

This paper shows the steps of analyze and visualize big data to predicting and diagnosing heart disease that will help hospitals and doctors. Some popular algo-rithms, such as K-means, C4.5, Support vector machines and Naïve Bayes, are used by WEKA tool to initial analyze and visualize the dataset. Then, We com-pared the correct classification and Mis-classification rate and show the compari-son in graphical representation.

REFERENCES

[1] Scruggs SB, Watson K, Su AI, Hermjakob H, Yates JR, 3rd, Lindsey ML, Ping P. Harnessing the heart of big data. Circ Res. 2015;116:1115–1119. doi: 10.1161/CIRCRESAHA.115.306013. [PMC free article][PubMed] [Cross Ref].

[2] Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. JMIR Med Inform. 2014;2:e1. doi: 10.2196/medinform.2913. [PMC free article] [PubMed] [Cross Ref].

[3] Partnership to Fight Chronic Disease. "The Growing Crisis of Chronic Disease in the United States", Retrieved from http://www.fightchronicdisease.org/sites/default/files/docs/GrowingCrisisofChronicDiseaseintheUSfactsheet_81009.pdf

[4] CDC NCHS. Underlying Cause of Death 1999-2013, https://www.cdc.gov/heartdisease/facts.htm , last accessed 2017/11/14.

[5] CDC NCHS. Underlying Cause of Death 1999-2013, https://www.cdc.gov/heartdisease/facts.htm , last accessed 2017/11/14.

[6] Senthilvel, G. (2014). DotNet Programming using Cassandra. Retrieved from https://www.codeproject.com/Articles/758803/DotNet-Programming-using-Cassandra

[7] John, W.,: *Data Science & Big Data Analytics:* Discovering, Analyzing, Visualizing and Presenting Data. Indianapolis, Indiana (2015).

[8] Aftab, H.: Predictive Analytics and Decision Support for Heart Failure patients. Pro Quest, Washington (2016).

[9] Snajay, S.: Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms. International Journal Of Engineering And Computer Science 6(6), 21632-21631 (2017).

[10] M.Akhil, J., B.L, D., Priti, C.: Classification of Heart Disease Using K-Nearest Neighbour and Genetic Algorithm. International Conference on Computational Intelligence: Modeling Techniques and Applications, pp. 85-94. Published by Elsevier Ltd, India (2013)..