ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

Big Data Algorithm of Heart Disease Prediction Using Machine Learning

Sheema Shajan M S[1], Faesa K H[2], Ms. Neethu Kunjappan[3]

Student, Bachelor of Computer Applications SNGIST ASC, Manakkapady N. Parvoor, Kerala, India[1]

Student, Bachelor of Computer Applications SNGIST ASC, Manakkapady , N. Parvoor, Kerala, India[2]

Assistant Professor, Bachelor of Computer Applications SNGIST ASC, Manakkpady, N. Paravoor, Kerala, India[3]

ABSTRACT: Now days, health prediction in modern life becomes very much essential. Big data analysis plays a crucial role to predict future status of health and offers prominent health outcome to people. Today People's are work on computers hours together, they are finding no time to take care of their health. Due to the hectic schedules and consumption of junk food, their health is being affected. By which it leads to heart disease. So we are implementing a heart disease prediction system using big data technique. Here we are using ID3 algorithm, k-means clustering algorithm, Naive Bayes, Support vector machine (SVM), Logistic Regression. Naive Bayes algorithm is used to analyze on dataset based on risk factors. We also used decision trees like ID3 and combination of algorithms for the prediction of heart disease based on the above attributes. By using this algorithm, it helps in predicting the heart disease by using various attributes and it predicts the output as in the prediction form. For grouping of various attributes it uses k-means clustering and for predicting it uses ID3 algorithm. The survey concept is machine learning based on disease prediction from medical field and it uses the big data concept, which means that the machine learning is a data mining technic, but this technics applied in disease prediction to come some difficulties such as, incomplete data, not suitable in large or big hospital and some results are inaccurate. so this difficulties are come in the existing, then, the next level is called "Big Data". The term Big Data is becoming global today. The term Big data is a huge amount of variety of data, and then the data is increasing very rapidly according to the time.So there is a need to process the data and instead of just storing that data need to extract some meaningful information or knowledge from that the data is apply some clustering and classification techniques of data mining. There are various era available in the Big Data so that decided the medical industry first. And after that there are various diseases available to work on them or gain some knowledge or predict for help we decided the Heart disease. Heart disease is one of the disease due to that death will occurred mostly, and according to the world health organization the percentage is more for that. So Heart disease is decided for the big Data approach, and as Big Data is considered so used Hadoop Map reduce platform.

KEYWORDS: Big Data**,** ID3 algorithm, k-means clustering algorithm, Naive Bayes, Support vector machine (SVM), Logistic Regression.

I. INTRODUCTION

There has been a tremendous growth in the medical industry over the years, Due to the advancement in the technology and an increasing in health problems have been observed. Due to the busy schedules of the people has led to increased many health issues. In this paper we are discussing about the prediction of heart related problems. Today, Heart disease is one of the major causes of morbidity and mortality among the population of the world. The amount of data in the medical field is huge.

According to the article, heart disease leading to major cause of death for both women and men. The article states that: **Approximately, About 610,000 people die of heart disease in the United States every year. Heart disease leading to major cause of death for both women and men .due to heart disease, more than half of the death in 2009 were**

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

in men But after 2009,the death caused by heart disease become increased. The disease is widely spread to both men and women. Coronary Heart Disease (CHD) is the most common type of heart disease, killing over approximately 370,000 people annually. Every year about 978,000 Americans have a heart attack. Of these, 875,000 are a first heart attack and 510,000 happen in people who have already had a heart attack. [4]

This makes heart disease as a major concern to be dealt with. But it is very difficult to identify heart disease because of several contributory risk factors like, diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining, Big data and Machine Learning for predicting the disease. A major challenge facing by healthcare organizations like hospitals, medical centers, it is the provision of quality services at an affordable costs. Quality service implies diagnosing patients in a correctly manner and administering treatments that are effective. Hospitals also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information decision support systems. Many hospital information systems are designed to support inventory management, patient billing, and generation of simple statistics. Some hospitals use decision support systems, but they are largely on a limited manner.

II. INTRODUCTION TO HEART FAILURE

Heart failure (HF) could be a major and growing public unhealthiness that accounts for over one million hospital admissions each year. The identification of at-risk people could offer necessary opportunities to intervene early within the illness method. Of patients presenting with new-onset HF in medical specialty studies, four-hundredth to seventy one have failure with preserved instead of reduced ejection fraction (HFPEF versus HFREF). though the clinical course and survival when HF onset are determined for each HFPEF and HFREF. The clinical characteristics before the onset of HF haven't been assessed consistently in reference to these established HF classes. variations in risk factors preceding the onset of HFPEF versus HFREF might offer necessary insights into distinct pathophysiologic pathways that disagree between the two entities. what is more, for persons in danger for the event of HFREF, like those with well left bodily cavity (LV) disfunction, there's trial proof that early treatment will improve outcomes. In distinction, very little is thought concerning treatments that specifically profit people in danger for or with the diagnosing of HFPEF.

III. CLINICAL PERSPECTIVE

Few previous studies, to our information, have examined risk factors that specifically take issue within the prediction of incident HFPEF versus HFREF. Current Indian school of Cardiology/Indian Heart Association HF tips classify at-risk people as stage A HF, however this cluster of people remains ill-defined with relation to form of HF. The aim of our study was, first, to look at risk factors of incident HF, and, second, to distinction clinical risk issues for incident HF with preserved versus reduced ejection fraction during a massive community-based cohort with the idea that the identification of distinct risk factor clusters for HFPEF versus HFREF might have necessary implications for the look of clinical trials and for future therapies to stop or treat HF.

IV. STATISTICAL ANALYSIS

Baseline clinical characteristics were summarized on an individual basis for participants UN agency did and didn't develop HF and were any dampened by participants with HFPEF and HFREF. For HF subtype analyses, participants UN agency developed HF that was unclassified with relevancy ejection fraction were enclosed within the cluster while not HF. we tend to examined accumulative incidence of HFPEF and HFREF employing a Kaplan-Meier-like technique whereas accounting for competitive risks (death different HF type). Analyses were conducted on an individual basis in men and girls and additionally in subgroups with or while not previous MI.
Proportional hazards regression was wont to model associations between clinical characteristics and incident HF.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

V. TERM BIG DATA

Big knowledge may be a phrase wont to mean a vast volume of each structured and unstructured knowledge that's therefore massive it's troublesome to method victimization ancient info and computer code techniques. In most enterprise situations the degree of knowledge is just too massive or it moves too quick or it exceeds current process capability.

VI. INTELLIGENT CHOICES

Big knowledge has the potential to assist corporations improve operations and build quicker, a lot of intelligent choices. The info is collected from variety of sources together with emails, mobile devices, applications, databases, servers and different suggests that. This data, once captured, formatted, manipulated, hold on and so analysed, will facilitate a corporation to achieve helpful insight to extend revenues, get or retain customers and improve operations.

Study of different heart disease prediction Algorithms

- **K MEANS ALGORITHM**

In k means that rule the term "k-means" was 1st employed by James MacQueen in 1967 [1], tho' the concept goes back to Victor-Marie Hugo Steinhaus in 1956.
Thequality rule was 1st planned by Stuart Harold Clayton Lloyd of Bell Labs in 1957 as a way for pulse-code modulation, tho' itwasn't revealed asajournalarticle till 1982.Kmeans bunch maybea straightforward unattended learnin g rule that's accustomed solve bunch issues.
It follows a straightforward procedure of classifying a given knowledge set into variety of clusters, outlined by the letter "k," that is mounted beforehand. The clusters area unit then positioned as points and every one observations or knowledge points area unit related to the closest cluster, computed, adjusted so the method starts over victimization the new changes till a desired result's reached. K-means is one in all the best unattended learning algorithms, and it's acknowledged for terribly speed and simple.
In k means algorithm is used to determine the heart disease prediction, that is it is used In our system K-means algorithm plays an important role so as to get the acceptable number of knowledge groups. Using this algorithm along with Euclidean distance centroids are calculated for various patient attribute. Mean value is taken under consideration for sample data and henceforth it's judgemental to predicate the patient status.

- **PROPOSED APPROACH**

Proposed system is the medical sector application. This system helps to predict of heart condition , supported manually inserted inputs. Inputs are like patient's age, gender, chest pain, the resting blood pressure in mmHg, serum cholesterol in mg/d, hereditary, fasting blood pressure in mg/dl, that and smoking. The database contains all the knowledge of the user i.e. the user details and the admin uploaded files, these will be compared to produce an output. Output is consisting either patient has heart condition or not.

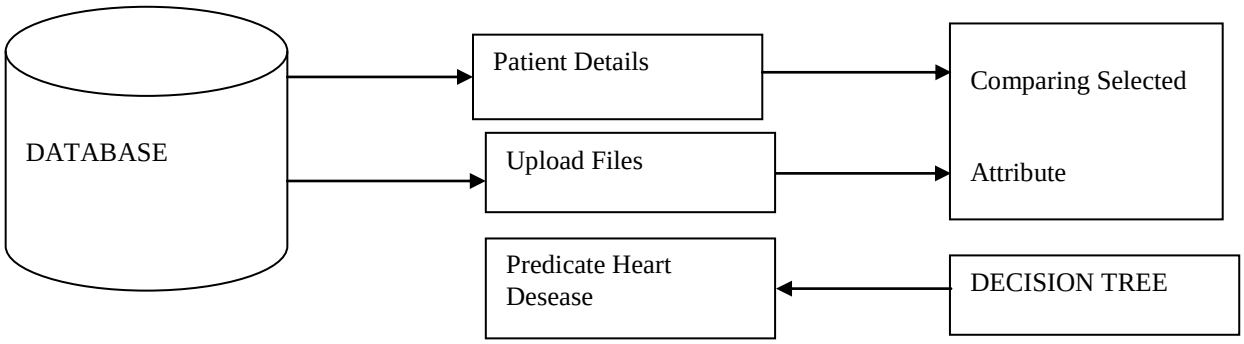
ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020



A. Decision Tree

Decision tree is a simpler classifier which is easy to implement and which does not require any domain knowledge. The main advantage is decision tree method can be applied to huge data which is to interpret. Decision tree is in the form of a tree like structure which consists of arcs, nodes, and branches. The arcs connect from one node to another. The branch has attributes, an internal node has a test on which the attribute is used for, and leaf node consists of the classes which are predicted from decision tree.to make an appropriate decision the traversing starts from root node to leaf node.

B. ID3 (Iterative Dichotomiser) ALGORITHM

ID3 algorithm is very easy and call tree learning algorithmic rule .It was developed by Ross Quinlan [2], in the year of 1983.ID3 algorithm is mainly used to construct the choice tree by using a top-down, here we have to use the greedy search for check the given set of attributes at every tree node. This algorithm also a non-progressive algorithmic rule, it means that it derives its categories from hard and fast set of coaching instances. In a decision tree consist of nodes and arcs that connect to nodes. If it form a choice, one start at the basis node, and ask to confirm that arc to follow, until one is reached at leaf node and the call is formed.

- **Naïve Bayes Algorithm**

Naïve Bayes algorithmic rule could be a smart tool in diagnosis. For example given an inventory of attributes named symptoms, it will speculate chance of a malady. Naïve Thomas Bayes take into account conditionally freelance attributes. The categoryifier processes every attribute's chance in a very class. The very best posterior chance category are thought of as results of the classification. Naïve Thomas Bayes is easy, economical and smart performance in classification. It additionally provides smart accuracy for general purpose analysis. Because of its smart accuracy it will be employed in diagnosis. [5]

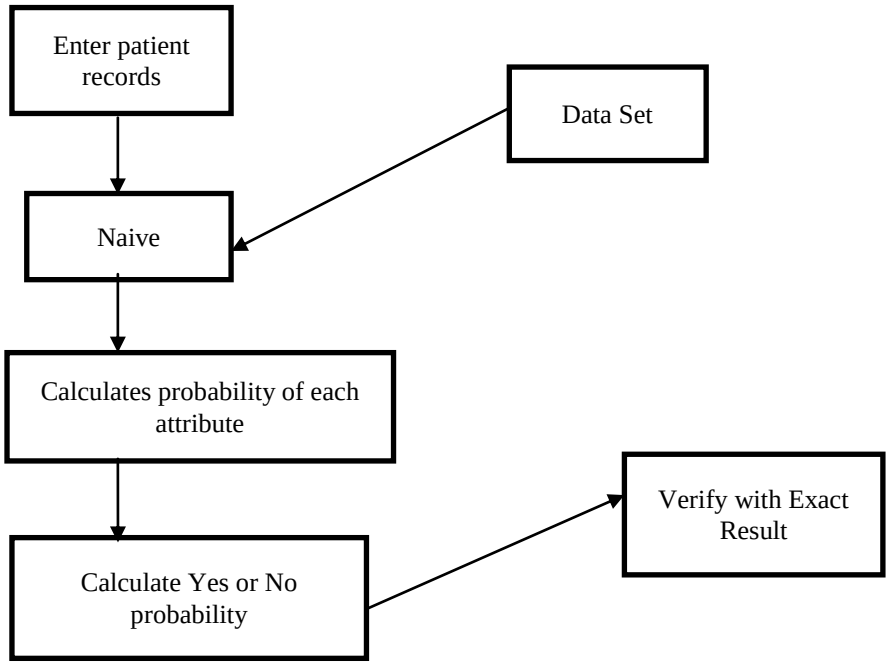
ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020



Mistreatment medical profiles like age, sex, vital sign and glucose, chest pain, graph graph etc. It will predict the chance of patients obtaining a cardiopathy. it'll be enforced in Python as Associate in Nursing application that takes medical test's parameter as Associate in Nursing input. It may be used as a coaching tool to coach nurses and medical students to diagnose patients with cardiopathy.

- **Support Vector Machine (SVM)**

A support vector machine may be a classification kind of methodology wont to scrutimize knowledge and acknowledge patters.In An exceedingly Regression and classification analysis. Support vector machine (SVM) is employed once your knowledge is assessed as 2 categories.

AN SVM acknowledges and separates similar knowledge by finding the most effective hyper plane that separates all knowledge points of 1 category from those of the opposite category. Model becomes higher once margins square measure larger between categories. A margin shouldn't have points in its interior half. The support vectors square measure the information coordinates that square measure on the boundary of the margin. Mathematical functions square measure concerned in SVM style that is often wont to model planet issues. Its performance amplify with range of attributes. [3]

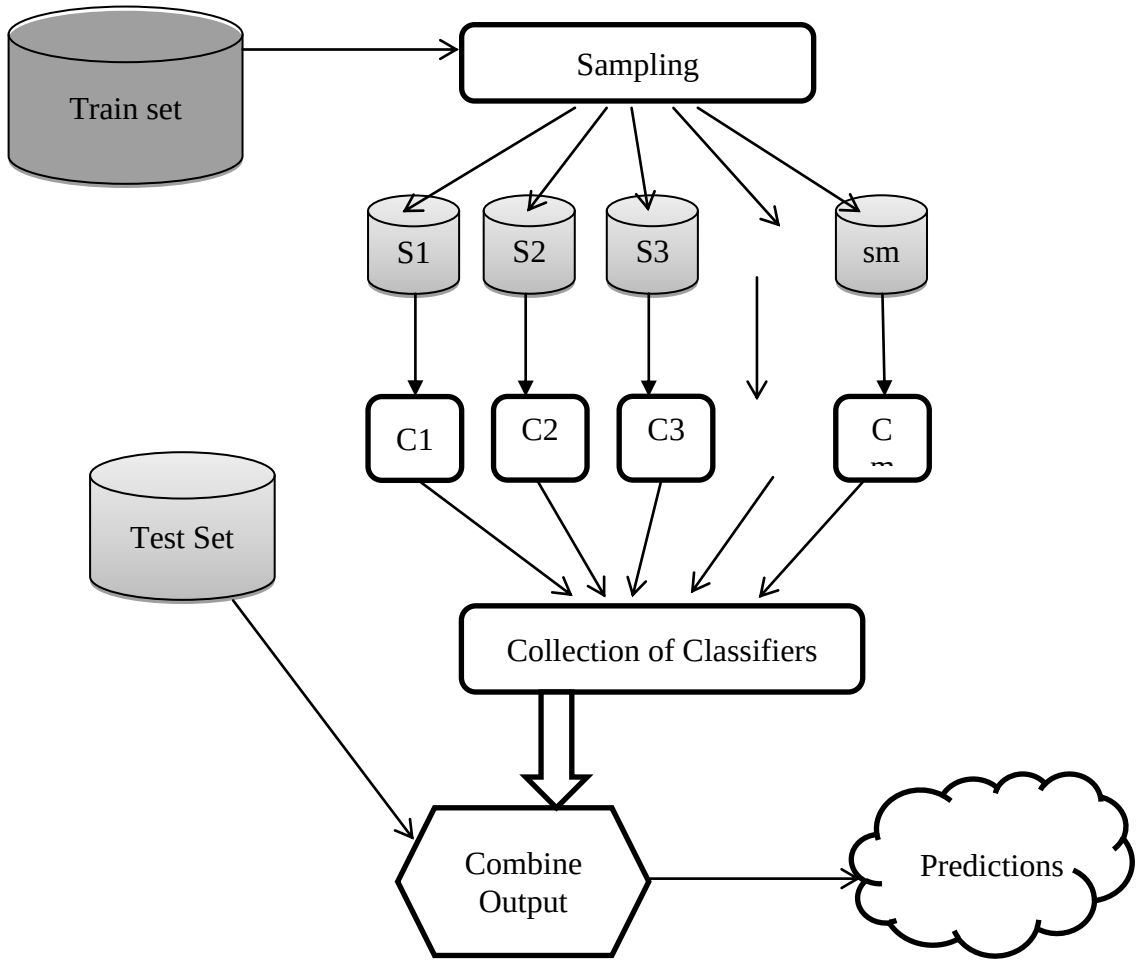
ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020



Flow Chart of SVM

Before moving more, let's discuss the options of SVM:

1. SVM could be a supervised learning formula. this implies that SVM trains on a group of labelled knowledge. SVM studies the labelled coaching knowledge then classifies any new input file reckoning on what it learned within the coaching section.
2. A main advantage of SVM is that it is used for each classification and regression issues. although SVM is principally glorious for classification, the Foreign Intelligence Service (Support Vector Regression) is employed for regression issues.
3. SVM is used for classifying non-linear knowledge by mistreatment the kernel trick. The kernel trick means that reworking knowledge into another dimension that features a clear dividing margin between categories of information. once that you'll simply draw a hyperplane between the assorted categories of information.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

- **Logistic Regression**

It's a kind of technique| statistical procedure} analysis method used for approximation and prediction of results of a specific dependent attributes. Dependent suggests that it'll take just a few set of values as an example binary values like true or false, sensible or dangerous, on or off likewise. Supplying regression is particularly used for prediction besides that it can even be used in calculative the prospect of success. Essentially supplying Regression involves fitting associate equivalentuation of the form to the data:

$Y = ß0 + ß1x1 + ß2x2 + \ldots + ßnxn$ – eq. 1

The Regression coefficients are sometimes calculable exploitation most chance estimation.
The utmost chance quantitative relation helps to see the applied mathematics significance of freelance variables on the dependent variables. The likelihood-ratio take a look at assesses the contribution of individual predictors (independent variables).
Then the chance (p) of every case is calculated exploitation odds quantitative relation,
P/(1-P) = eY – equivalent. a pair of From this p–value is observed. this provides the chance or probability for the individual to possess coronary cardiovascular disease. [3]

VII. MACHINE LEARNING AND PREDICTIVE MODELS

Machine learning uses applied mathematics ideas to modify machines (computers) to "learn" while not specific programming. A supplying approach fits best once the task that the machine is learning relies on 2 values, or a binary classification. exploitation the instance on top of, your pc may use this kind of study to create determinations regarding promoting your supply and take actions all by itself. And, as additional knowledge is provided, it may find out how to try to this higher over time.

VIII. CONCLUSION

The amount of Heart diseases will exceed the management line and reach to most purpose. malady| heart condition| cardiopathy| cardiovascular disease} square measure difficult and each and every year lots of folks square measure dying with this disease By exploitation this all systems one in all the foremost drawbacks of those works is principally focus solely to the appliance of classify techniques and algorithms for cardiovascular disease prediction, by of these learning numerous information cleanup and mining techniques that prepare and build a dataset acceptable for data processing. in order that I will use this Machine Learning in this logistical regression algorithms by predicting if patient has cardiovascular disease or not. Any nonmedical worker will use this code and predict the center illness and scale back the time quality of the doctors.

IX. FUTURE WORK

Today's, world most of the info is processed, the info is distributed and it's not utilizing properly. By Analysing the offered information we will additionally use for unknown patterns. the first motive of this analysis is that the prediction of heart diseases with high rate of accuracy. For predicting the center illness we will use logistical regression formula, navie bayes, sklearn in machine learning. the long run scope of the paper is that the prediction of heart diseases by exploitation advanced techniques and algorithms in less time quality.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

X.ACKNOWLEDGEMENT

This research was supported by our college SNGIST Arts and Science College, Manakkapady, N.Paravur. We thank our teachers who provided insight and expertise that greatly assisted the research, although they'll not accept as true with all of The Interpretations of This paper.
We thank our department, Department of Computer Applications, SNGIST ASC, for his or her assistance with particular Methodology and for comments that greatly improved the manuscript. We would also wish to show our gratitude to the opposite teachers and friends for sharing their pearls of wisdom with us during the course of this research, and that we thank reviewers for his or her so-called insights. We are also immensely grateful to our guides and mentors for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons.

REFERENCES

[1] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1.University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

[2] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106.

[3] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", IJCA, Vol.68- No.16 April 2013.

[4] https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

[5] Mrs.G.Subbalakshmi, "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering.