Data Mining Approach to Detect Heart Diseases

Muhsin P.M[1], Bindu George[2]
Dept. of computer sciences Santhigiri college of computer sciences
Vazhithala P.O, Thodupuzha, Idukki
[1]*muhammedmuhsin707@gmail.com*
[2]*sr.tess@santhigiricollege.com*

Abstract: ***Living in the era of technology data is more important than anything else. The information that is stored digitally can be analysed to get a clear picture about the individuals likes his attitude etc. Which means that after systematic collection and analysis of collected information leads to knowledge discovery. This process is termed as data mining. This paper is a study on the different data mining methods can be used to analyse the collected cardiological information of each individual in order to find out which of these methods can indicate the possibility of a cardiac arrest with the highest precision.***

Keywords: Data mining, Heart diseases, k-mean clustering, Regression model, decision tree .

1. Introduction

Globally, heart diseases are the number one cause of death. About 80% of deaths occurred in low- and middle-income countries. If current trends are allowed to continue, by 2030 an estimated 23.6 million people will die from cardiovascular disease (mainly from heart attacks and strokes). The healthcare industry gathers enormous amounts of heart disease data which, unfortunately, are not "mined" to discover hidden information for effective decision making. The reduction of blood and oxygen supply to the heart leads to heart disease. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data.if the data is analysed properly then it will lead to fast and efficient decision-making which intern helps to save more lives with the help of medical support systems. [4]

Data mining is a technique that helps you to analyse the data to give the information needed with precision.Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals etc. The data generated by the health organizations is very vast and complex due to which it is difficult to analyse the data in order to make important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to generate a powerful tool for analysing and extracting important information from this complex data. The analysis of health data improves the healthcare by enhancing the performance of patient management tasks.[2]

In this paper we will be discussing three important data mining techniques.k-mean clustering, multiple regression model and naive base method as to find out which is that one method that gives us the accurate solution.

2. Literature Survey

Data Mining is one of the most vital and motivating area of research.Data mining has become very popular over the last two decades as a discipline in its own..Data Mining came into existence in the middle of 1990's and appeared as a powerful tool that is suitable for fetching previously unknown pattern and useful information from huge dataset. Various studies highlighted that Data Mining techniques help the data holder to analyse and discover unsuspected relationship among their data which in turn helpful for making decision.

Data mining has applications inother various fields like targeted advertisement, Data mining applications are also used in every field of business, government, and science just to name a few. Starting from text mining apart from healthcare sector.[3]

You might think the history of Data Mining started very recently as it is commonly considered with new technology. However data mining is a discipline with a long history. It starts with the early Data Mining methods Bayes' Theorem (1700`s) and Regression analysis (1800`s) which were mostly identifying patterns in dataicreasing power of technology and complexity of data sets has led Data Mining to evolve from static data delivery to more dynamic and proactive information deliveries; from tapes and disks to advanced algorithms and massive databases (see the table below). In the late 80`s Data Mining term began to be known and used within the research community by statisticians, data analysts, and the management information systems (MIS) communities. By the early 1990`s, data mining was recognized as a sub-process or a step within a larger process called Knowledge Discovery in Databases (KDD).Which actually is "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" as per Fayyad's definition of KDD.The popularity of data mining escalated notably in the 1990`s, with the help of dedicated conferences, in addition to the fast increase in technology, data storage capabilities and computers` processing speeds. It was also possible for organizations to keep data in computer readable form and processing of large volumes of data using desk top machines were not far from reality.By the end of 1990`s, data mining was already a well-known technique used by the organizations after the introduction of customer loyalty cards. This opened a big door allowing organizations to record customer purchases and data, the resulting data could be mined to identify customer purchasing patterns. The popularity of data mining has continued to grow rapidly over the last decade.[7]

3. Methods / Approach

We will be discussing three important methods k-mean clustering to start off with multiple regression model and finally decision trees about how these methods can be used for heart disease detection.

I have chosen heart disease detection as specialization area of the study as it is one the major reason for lose of life in the developing nations which is caused not by the treatment but just as a lot of time is spent on disease identification or prediction and identification of risk elements accurately in time. Data mining technology can be used to overcome this mess. How ever technology is meant to improve the quality of human life. Average life expectancy is a quality measure for the same. So, technology must be used to increase the life expectancy before any other advancement. Because in a world without humans technology becomes useless. As human diseases are most common cause of death and as it can be reduced in large scale through efficient prediction I wanted to study a method for efficient analysis for prediction.

We consider the value of the disease detection test and not the images. We consider a set of attributes for each individual cardiac patient like Blood pressure level, ECG level, Treadmill test level, age, gender etc.
In this research we are explaining the methods but not the implementation.

3.1. Clustering

Cluster analysis is a method of organizing data into representative groups based upon similar characteristics. Each member of the cluster has more in common with other members of the same cluster than with members of the other groups. The most representative point within the group is called the centroid. Usually, this is the mean of the values of the points of data in the cluster. Organize the data. If the data consists of a single variable, a histogram might be appropriate. If two variables are involved, graph the data on a coordinate plane. For example, if you were looking at the height and weight of school children in a classroom, plot the points of data for each child on a graph, with the weight being the horizontal axis and the height being the vertical axis. If more than two variables are involved, matrices may be needed to display the data.[1]

Group the data into clusters. Each cluster should consist of the points of data closest to it. In the height and weight example, group any points of data that appear to be close together. The number of clusters, and whether every point of data has to be in a cluster, may depend upon the purposes of the study.

For each cluster, add the values of all members. For example, if a cluster of data consisted of the points (80, 56), (75, 53), (60, 50), and (68,54), the sum of the values would be (283, 213). Divide the total by the number of members of the cluster. In the example above, 283 divided by four is 70.75, and 213 divided by four is 53.25, so the centroid of the cluster is (70.75, 53.25).[9]

3.1.1 k-mean clustering

K-means clustering is a type of unsupervised learning, which is used when you have un labelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided.Rather than defining groupsbefore looking at the data, clustering allows you to find and analyse the groups that have formed organically. The"Choosing K" section below describes how the number of groups can bedetermined. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

A. Clustering procedure

k-mean clustering algorithm is a two-step process firstly it is required to perform data processing then we can apply the k-means algorithm.

a. Data processing

Cleaning and filtering of the data might be necessarily carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns. The steps involved in the pre-processing of a dataset are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. To make data appropriate for the mining process it needs to be transformed. The raw data is changed into data sets with a few appropriate characteristics. Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm . This leads to removal of duplicate records and supplying the missing values in the heart disease data warehouse. In addition, it is also transformed to a new form which is appropriate for clustering. Clinical databases have accumulated large quantities of information about patients and their medical conditions.[5]

b. k-means algorithm

K-means algorithm is one of the partitioning based clustering algorithms .The general objective is to obtain the fixed number of partitions/clusters that minimize the sum of squared Euclidean distances between objects and cluster centroids.
Let X={xi| i=1,2,…………..,n} be a data set with n objects, k is the number of clusters, mj is the centroid of cluster cj where j=1,2,……….,k. Then the algorithm finds the distance between a data object and a centroid by using the following Euclidean distance formula.
The Euclidean distance between two points/objects/items in a dataset, defined by point X and point Y is defined by Equation below.

EUCLIDEAN DISTANCE(X,Y) = (|X1-Y1|2 + |X2-Y2|2 + ... + |XN-1-YN-1|2 + |XN-YN|2)1/2(1)

OR Euclidean distance formula=√Σ|xi-mj|2 (2)

where **X** represents is the first data point, **Y** is the second data point, **N** is the number of characteristics or attributes in data mining terminology.
Starting from an initial distribution of cluster centers in data space, each object is assigned to the cluster with closest center, after which each center itself is updated as the center of mass of all objects belonging to that particular cluster. The procedure is repeateduntil convergence.

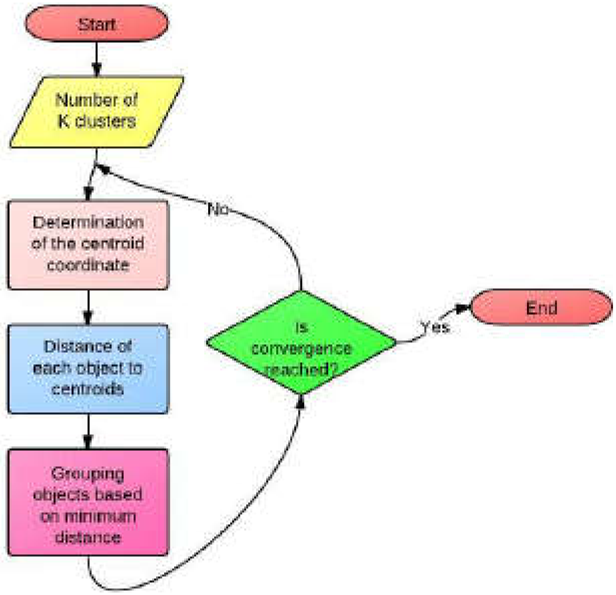

Start
Number of K clusters
Determination of the centroid coordinate
Distance of each object to centroids
Grouping objects based on minimum distance
No
Is convergence reached?
Yes
End

fig 1: Steps of k-means algorithm

Algorithm

```
INPUT: // Set of n items to cluster
D= {d1, d2, d3,………………, dn}
   // No. of cluster (temporary cluster)
      randomly chosen i.e. k
// So below, K is set of subset of D
as temporary cluster and C is set of
centroids of those clusters.
K= {k1, k2, k3,………………, kk
},
C= {c1, c2, c3,………………, ck}
Where k1= {d1}, k2=
{d2}, k3= {d3}…… kk= {dk}
And c1=d1, c2=d2,
c3=d3,……… ck=dk,
// here k<=n
Output: // // K is set of subset of D as final
cluster and C is set of centroids
of these cluster.
K= {k1, k2, k3,………………, kk
},
C= {c1, c2, c3,………………, ck}
Algorithm:
K-means (D, K, C)
1. Arbitrarily choose k objects from D as
the initial cluster centers.
2. Repeat
3. (re) assign each object to the cluster
to which the object is the most similar,
based on the mean value
of the objects in the cluster.
4. Update the cluster means, i.e.,
calculate the mean value of the objects
for each cluster.
      5. Untilno change.            [19]
```

3.2 Multiple regression model

Regression is a data mining (machine learning) technique which is used to fit an equation for the dataset.
Multiple Linear Regression is a statistical model that can be used to describe data and to explain the relationship between one dependent variable and two or more independent variables. Analysing the correlation and directionality of the data, fitting the line, and evaluating the validity and usefulness of the model are the different stages of multiple linear regression model .In this technique, a dependent variable is modelled as a function of several independent variables with corresponding multiple regression coefficients, along with the constant term. Multiple regressionsrequire two or more predictor variables, and that is why it is called multiple regression.[10]
The regression line represents the estimated disease chance for a given combination of the input factors. Scatter plot is defined by a linear equation
In regression model there is a continuous random variable called the dependent variable, Y, and a number of independent variables, x1, x2..., xp. Our purpose is to predict the value of the dependent variable (also referred toas the response variable) using a linear function of the independent variables. The values of the independent variables(also referred to as predictor variables, regressors or covariates) are known quantities for purposes of prediction, the model is

$Y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp+\varepsilon$ (3)

whereε, the "noise" variable, is a Normally distributed random variable withmean equal to zero and standard deviationσwhose value we do not know. Wealso do not know the values of the coefficientsβ0,β1,β2,...,βp. We estimateall these (p+2) unknown values from the available data.The data consist ofnrows of observations also called cases, which giveus valuesyi,xi1,xi2,...,xip;i=1,2,...,n.The estimates for theβcoefficientsare computed so as to minimize the sum of squares of differences between the fitted (predicted) values at the observed values in the data.
[6]

3.2.1 Mathematical formulation

AMultiple regression technique is an extension of alinear regression technique which involves more than one predictor variable. Itallows response variable Y to be modelled as a linear function of multidimensional feature vector.Multiple regressionmodel consists of random variable Y (called as a response variable) as a linear function of randomvariable X1 (called as a predictor variable) and X2 and thatis represented by the equationthat is we have

$Y= \alpha +\beta 1X1+\beta 2X2$ (3)

Where α ,β1and β2 are regression coefficients The regression coefficient α , β1 & β2are solved by the methodof least squares, whichminimize the error between the actual data & the estimate of the line.Basically multipleregressiongenerally explains the relationship between multiple independent or multiple predictor variables and one dependent or criterion variable. In multiple regression, a dependent variable is modelled as a function of several independent variables with corresponding multiple regression coefficients, along with the constant term. [6]

Algorithm for Multiple Regression

The Multiple regression technique works on the following algorithm.

Step 1: Take the values of variable Xi, Xb and Yi.

Step 2: Calculate the summation of the variable Xi,Xb,Yi.

Step 3: Calculate the product of summation terms.

$(\sum x1*x,\sum x1*x2,\sum x2* y,\sum x2*x2, \sum x1*y)$.

Step 4: Solve the equations

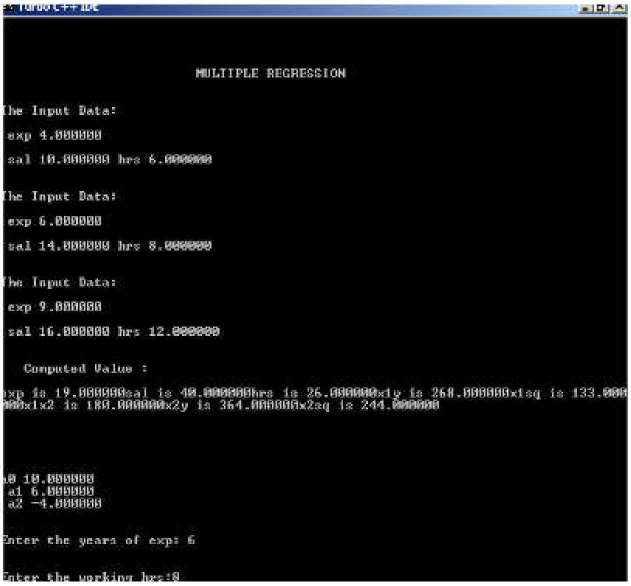
$\sum y=na0 +a1\sum x1 +a2\sum x2$ $\sum x1y=a0 x1 +a1\sum x1*x1 +a2\sum x1x2$ $\sum x2 y=a0 x2 +a1\sum x1*x2 +a2\sum x2*x2$.

Step 5: Now calculate the value of a0, a1, a2 which is calculated by the inverse of a matrix.

Step 6: Finally obtain the value of response variable Y by knowing the values of a0, a1, a2 int he equation Y= a0+a1X1+ a2X2 of β(calculated in step 4), average of Xi and average of Yi .

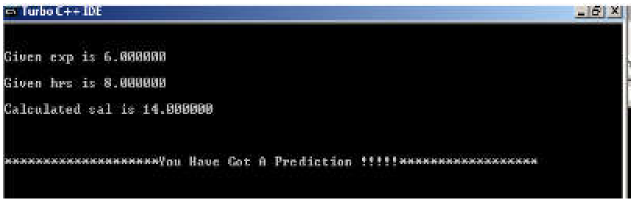
Step 7: Finally substitute the value of regression coefficients α and β in the equation Y= α + βX.

Sample input and output



```
                    MULTIPLE REGRESSION

The Input Data:
 exp 4.000000
 sal 10.000000 hrs 6.000000

The Input Data:
 exp 6.000000
 sal 14.000000 hrs 8.000000

The Input Data:
 exp 9.000000
 sal 16.000000 hrs 12.000000

   Computed Value :
exp is 19.000000sal is 40.000000hrs is 26.000000x1y is 268.000000x1sq is 133.000
000x1x2 is 180.000000x2y is 364.000000x2sq is 244.000000



a0 10.000000
 a1 6.000000
 a2 -4.000000

Enter the years of exp: 6

Enter the working hrs:8
```

Fig2:Sample input and output(a)



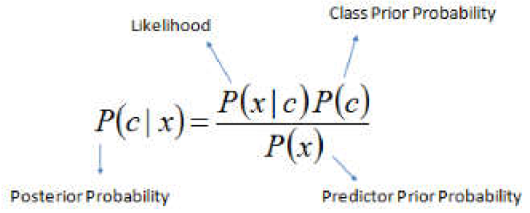
```
Given exp is 6.000000

Given hrs is 8.000000

Calculated sal is 14.000000


**********************You Have Got A Prediction !!!!!******************
```

Fig2:Sample input and output (b)

[20]

3.3 Naive Bayes

Native Bayes classifiers is a probabilistic classifier based on applying bayes theorem with strong(native) independence assumptions between the features. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful in the field of medical science for diagnosing heart patients. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.


Likelihood
Class Prior Probability
$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c\mid X)=P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Fig3:Equation to apply Naive base 1

P(c|x) is the posterior probability ofclass (target) given predictor (attribute).
P(c) is the prior probability ofclass.
P(x|c) is the likelihood which is the probability of predictor givenclass.
P(x) is the prior probability ofpredictor

Where C and X are two events (e.g. the probability that the train will arrive on time given that the weather is rainy). Such Naïve Bayes classifiers use the probability theory to find the most likely classification of an unseen (unclassified) instance . The algorithm performs positively with categorical data but poorly if we have numerical data in the training set. [21]

The Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. [14]

3.3.1 Bayes Rule

A conditional probability is the likelihood of some conclusion say C, given some evidence/observation, E, where a dependence relationship exists between Cand E. This probability is denoted as P(C|E) where

P (C|E) =P (E|C) P(C)
P(E) (4) [16]

3.3.2. Naive Bayesian Classification Algorithm

The naive Bayesian classifier, or simple Bayesian classifier, works as follows

1. Let D be a training set of tuples and their associated class labels as Ca and Cp. As usual, each record is represented by an n-dimensional attribute vector, X=(x1, x2...,xn-1,xn), depicting n measurements made on the tuple from n attributes, i.e.A1 to An.
2. Suppose that there are m number of classes for prediction, C1, C2... Cm. Given a record, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if

   P(Ci|X)>P(Cj|X) (5)
   for I<=j<=m
   and
   j ≠i

Thus we maximize P(Ci|X). The class Ci for which
P (Ci | X) is maximized is called the maximum
posteriori hypothesis. By Bayes' theorem.

P (Ci | X) =P (x | Ci) P(Ci)
P(x) (5)

4. Given data sets with many attributes, it would be extremely computationally expensive to compute P(X | Ci). To reduce computation in evaluating P(X | Ci), the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple
(i.e., that there are no dependence relationships among the attributes). Thus,

P(X | Ci)=π^{m}k=1=P(xk |Ci)
=P (x1|Ci) *P (x2|Ci) *... P (xm |Ci)

We can easily estimate the probabilities P (x1|Ci), P (x2|C i)... P (xm |Ci) from the database training tuples. Recall that here xk refers to the value ofattribute Ak for tupleX.For each attribute, we will see that whether the attributeis categorical or continuous-valued. For instance, to compute P (X| Ci), we consider the following:

(a) If Ak is categorical, then P (Xk |Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci, D| ,the number of tuples of class Ci in D.

5.In order to predict the class label of X, P(X |Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

P(X |Ci)P(Ci)>P(X |Cj)P(Cj) for 1 ≤j ≤m, j ≠i

In other words, the predicted class label is the class Ci for which P (X |Ci) P (Ci) is the maximum.[22]

4. Result/Discussion

clusters found by k-means are used to train a classification model. These clusters alone give a decent model with an accuracy of 78.33%.[11]

K means algorithm is relatively simple to implement. It scales to large data sets. Guarantees convergence. Can warm-start the positions of centroids .Easily adapts to new examples. Generalizes to clusters of different shapes and Generalizes to clusters of different shapes and sizes, such as elliptical clusters. sizes, such as elliptical clusters.

It has a few set backs too.

Clustering data of varying sizes and density.

k-means has trouble clustering data where clusters are of varying sizes and density. To cluster such data, you need to generalize k-means as described in the Advantages section.

Clustering outliers.

Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering.[18]

In regression model, the most commonly known evaluation metrics include:

1. **R-squared** (R2), which is the proportion of variation in the outcome that is explained by the predictor variables. In multiple regression models, R2 corresponds to the squared correlation between the observed outcome values and the predicted values by the model. The Higher the R-squared, the better the model.
2. **Root Mean Squared Error** (RMSE), which measures the average error performed by the model in predicting the outcome for an observation. Mathematically, the RMSE is the square root of the *mean squared error (MSE)*, which is the average squared difference between the observed actual outcome values and the values predicted by the model. So, MSE = mean((observeds - predicteds)^2) and RMSE = sqrt(MSE). The lower the RMSE, the better the model.
3. **Residual Standard Error** (RSE), also known as the *model sigma*, is a variant of the RMSE adjusted for the number of predictors in the model. The lower the RSE, the better the model. In practice, the difference between RMSE and RSE is very small, particularly for large multivariate data.
4. **Mean Absolute Error** (MAE), like the RMSE, the MAE measures the prediction error. Mathematically, it is the average absolute difference between observed and predicted outcomes, MAE = mean(abs(observeds - predicteds)). MAE is less sensitive to outliers compared to RMSE.[12]

Multiple linear regression allows the investigator to account for all of these potentially important factors in one model. The advantages of this approach are that this may lead to a more accurate and precise understanding of the association of each individual factor with the outcome.[17]

On the other hand, it has set backs as Linear Regression Only Looks at the Mean of the Dependent Variable. Linear regression looks at a relationship between the mean of the dependent variable and the independent variables.Linear Regression Is Sensitive to Outliers. Outliers are data that are surprising and the data should be independent.[15]

We prefer naive base algorithm in situations we expect more efficient output, as compared to other methods output When the data is high, we choose naive base algorithm over the others and also when the attributes are independent of each other.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, NaiveBayes is known to outperform even highly sophisticated classification methods.The Naive Bayes algorithm affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows. The build process for Naive Bayes is parallelized. (Scoring can be parallelized irrespective of the algorithm.)

On the other side naive Bayes is also known as a bad estimator, **so** the probability outputs are not to be taken too seriously. Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.[22]

It is simple and easy to implement. It doesn’t require as much training data. It handles both continuous and discrete data.

It is highly scalable with the number of predictors and data points.

It is fast and can be used to make real-time predictions

It is not sensitive to irrelevant feature

The main limitation of Naive Bayes is the assumption ofindependent predictor features**.** Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it’s almost impossible that we get a set of predictors that are completely independent or one another. If a categorical variable has a category in the test dataset, which was not observed in training dataset, then the model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use a smoothing technique. [23]

5. Conclusion

Heart diseases are one of the major causes of death around the world despite of all the advancements in healthcare it continues to be in the same way. On theother hand, a large cardiac related patient information is being generated which can be analysed to predict the disease early and prevent it by analysing the symptoms. But it is not so much common as the analysis is not that efficient. Here datamining comes to our rescuedatamining algorithms can be used for efficient analysis of data.

In this paper we have discussed k-means clustering algorithm, Multiple linier regression model and naive byes the three important methods in datamining for the prediction of heart diseases by analysing information on each factor that can lead to heart diseases. We also came the pros and cons of each of them and also, we have determined the accuracy or method to find accuracy of each of the mentioned approaches.

To summarise about choosing a method among these for heart disease prediction it can be explained simply as
If the data items or items or attributes are independent of each other go for naive bayes algorithm other wise leave this method as this method is native i.e it implicitly assumes that all features to be independent.

Among k-means clustering algorithm and multiple linier regression model which assumes that their is a relation existing between features or attributes go for Multiple linier regression model as k-means algorithm is less efficient when each clustering is of varying sizes and it also suffer more from outliers (values that dose not follow the common trend found by analysing each data object) than multiple linier regression model as k-means algorithm gives more importance to the mean of values instead of giving importance to each value set within a cluster.

6. References

[1]. Analysis And Study Of K-Means Clustering Algorithm
Sudhir Singh and Nasib Singh Gill
Dept of Computer Science & Applications
M. D. University, Rohtak, Haryan

[2]. Data Mining; A Conceptual Overview Joyce Jackson University of South Carolina, joyce.jackson@sc.edu

[3]. A survey on Data Mining approaches for Healthcare Divya Tomar and Sonali Agarwal Indian Institute of Information Technology,
Allahabad, India
divyatomar26@gmail.com,sonali@iiita.ac.in

[4].Data Mining Approach to Detect Heart Diseases
November 2013
Authors:
Vikas Chaurasia
Veer Bahadur Singh Purvanchal University
Saurabh Pal
Veer Bahadur Singh Paranuchal University

[5]. HEART DISEASE FORECASTING SYSTEM USING K-MEAN CLUSTERING ALGORITHM WITH PSO AND OTHER DATA MINING METHODS
Shilna S , Navya EK Mtech Student, Assistant Professor, Department of Computer Science, Malabar Institute of Technology Kannur, Kerala, India shilna94@gmail.com1 , navya.06rimaan@gmail.com2

[6]. Prediction of Heart Disease using Multiple Linear Regression Model
K.Polaraju, D.Durga Prasad
1M.Tech Scholar,2Assistant Professor
1,2Department of Computer Science & Engineering,
1,2Baba Institute of Technology and Engineering , Visakhapatnam, INDIA.

[7]. https://www.exastax.com/big-data/the-history-of-data-mining/

[8]. https://blogs.oracle.com/datascience/introduction-to-k-means-clustering.

[9]. https://sciencing.com/centroid-clustering-analysis-10070345.html

[10]. Multiple Linear Regression in Data Mining.
https://ocw.mit.edu/courses/sloan-school-of- management/15-062-data-mining-spring-2003/lecture-notes/lecture9.pdf

[11]. https://towardsdatascience.com/kmeans-clustering-for-classification-74b992405d0a

[12]. Regression Model Accuracy Metrics: R-square, AIC, BIC, Cp and more - Articles - STHDA

[13].https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn

[14]. Understanding the Mathematics Behind Naive Bayes | by Nikita Sharma | Heartbeat (fritz.ai)

[15]. https://sciencing.com/disadvantages-linear-regression-8562780.html

[16].https://onlinelibrary.wiley.com/doi/pdf/10.1197/j.aem.2003.09.006.

[17]. https://sciencing.com/disadvantages-linear-regression-8562780.html.

[18]. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

[19]. Analysis And Study Of K-Means Clustering Algorithm
Sudhir Singh and Nasib Singh Gill
Dept of Computer Science & Applications
M. D. University, Rohtak, Haryana

[20]. A Multiple Regression Technique in Data Mining
Swati Gupta Assistant Professor, Department of Computer Science Amity University Haryana, Gurgaon, India

[21].
https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[22]. Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing
Ms. Rupali R .Patil
Asst. Professor, Jawaharlal Nehru College of Engineering, (Affiliated to BAMU,Aurangabad),Maharashtra, India.

[23].https://www.i2tutorials.com/advantages-and-disadvantages-of-naive-bayes-classifier/

7. Authors Profile

Muhsin P.M currently pursuing bachelor of computer application degree at Santhigiri College of Computer Sciences,Vazhithala.

Bindu George received the MCA professional degree in computer science. Currently working as assistant professor in Santhigiri college of computer sciences ,Vazhithala.