International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Data Mining Algorithms on Prediction of Cardiovascular Diseases

Archana Singh, Seema Mahajan

Abstract— In the age of data generation known as Big Data, where data is produced in enormous amount, managing it has become a big challenge and along with this drawing information from the gathered data is equally important and challenging. Inferring relationships and predicting patterns from theses structured and unstructured data is now an area of research for researchers. And the data mining techniques have evolved as a tool for generating results and deducing conclusions. These mining algorithms find their applicability in almost every domain likewise understanding market segment, fraud detection, trend analysis, healthcare sector, education sector and many more. Looking at the wide range of applicability, in this paper, a brief overview of data mining algorithms is discussed. This discussion comprises of different data mining algorithms, their mathematical modelling, their evaluation methods, and their limitations. To support the fact a case study is conducted on a cardiovascular disease dataset and the measures of these mining techniques are compared.

Index terms: SVM, Naïve Bayes, Random Forest, ROC analysis, confusion matrix, visual matrices, performance matrix

I. INTRODUCTION

The terminology of data mining was visualized around 1990 in the field of the database. This innovative and magical term is also referred as information garnering, information sighting, and information abstraction. The he concept of finding insights from data has evolved over years and years and in earlier stages of classification, two main analysis techniques played an important role that is Bayes' theorem and regression. But the advancement in computerization and computing techniques the ability of collecting, storing and manipulating data has increased in dramatic manner [12]. In the era of digitization, the amount of data which is produced, in every domain is huge, vast and unmanageable. This has given birth to a new terminology known a big data. Therefore deriving meaningful conclusion from big data has made data mining interest for researchers.

Today where information/data is coming from multiple sources. It is heterogeneous and massive and is having dynamic nature. The data today is used in distributed and shared environment. And processing this massive data is a big challenge. The approach now needs to be changed from quantitative to qualitative, although we have strong parallel programming models available, like Map Reduce, which provides strong computational power, but still to increase the real time applicability these models needs to work with machine learning and mining systems/techniques [13]. These classification algorithms have to go through rigorous training of data to obtain insight of data and draw knowledge. This requires rigorous computation to access the data (big data) repeatedly. Henceforth in second section of this paper, major data mining algorithms are discussed.

As we know that data is growing with a drastic speed in every field, velocity and volume of data is of prime concern to almost all the sectors like education, business, scientific research etc. Likewise data generated healthcare sector is a prominent problem and lots of research is being done in the field of big data to manage these data. And as data mining is an inherent part of big data for predictions and classification, so is mining important in the healthcare sector.The medical history of the patients have inherent and hidden sequences which are mandatory for analysis in the detection of diseases. In this research paper focus is on heart diseases. According to one of the article published in Times of India, the death proportion as a result of Cardio capture has declined by a noteworthy 41% inside the North American country somewhere in the range of 1990 and 2016, though in Asian nation it ascended by around 34% from 155.07 to 209.1 deaths per one hundred thousand populace inside a similar sum said another global investigation published by Elsevier in the Journal of American school of cardiology. One of the report of the world health organization (WHO) says, the ongoing and advantageous diagnosis of heart maladies assumes a wonderful job in counteracting its improvement and lessening pertinent treatment cost. It is very much clear that there are certain measures which have high impact on heart diseases. And understanding this pattern can help suggest people about the changes they could make in order to prevent such cardiovascular problems [12]. Looking at the criticality of the problem researchers have experimented different mining algorithms to classify and diagnose cardiovascular diseases before they become fatal. Mining algorithms are blend of statistical analysis, database technology and machine learning to extract knowledge and establish relationships among large databases [1].

Manuscript published on 30 September 2019
* Correspondence Author
Archana Singh*, Research Scholar, Indus University, Ahmedabad – Gujarat, India, E-Mail: archanasingh.rs@indusuni.ac.in
Seema Mahajan, Indus University, Ahmedabad, Gujarat, India, E-Mail: ce.hod@indusuni.ac.in

Consequently, third section provides a literature survey showing the importance of mining algorithms in prediction of cardiovascular diseases, which is taken as the case study for this survey.

Finally, in the fourth section an attempt is made to showcase the impact of data mining techniques by performing the mining task on heart disease dataset (Kaggle's Dataset).

The data mining tool used here for analysis is "Orange Canvas". And lastly a comparison of various data mining algorithms experimented by some of the researchers working in the direction of heart diseases prediction is summed, followed by the conclusion.

II. DATA MINING ALGORITHMS

The algorithms available, are to be applied based upon the nature of the problem, descriptive or predictive. Here the concentration is on predictive data mining and some of the algorithms explored are, Support Vector Machine (SVM), Random Forest, KNN Model of classification and Naïve Bayes.

2.1 Support Vector Machine

This algorithm (SVM) designs the classifying parameters in high magnitude feature domain. It is a machine learning technology that divides the input space with a hyperplane and hence increases the marginal gap between the different classes, due to which this technique results into high predictive measures. SVM executes linear regression in a high dimensional feature space using an insensitive loss (ε). Evaluation accuracy of the machine is dependent on a favourable ambiance of constant C, intensive loss (ε) and the kernel parameters. It uses a technique called the kernel trick where it converts not separable parameters to separable parameters. It is most of the time used for non-linear separation .The figure 1 (Garcia-Gonzalo, 2016) represents SVM for linearly separable dataset.

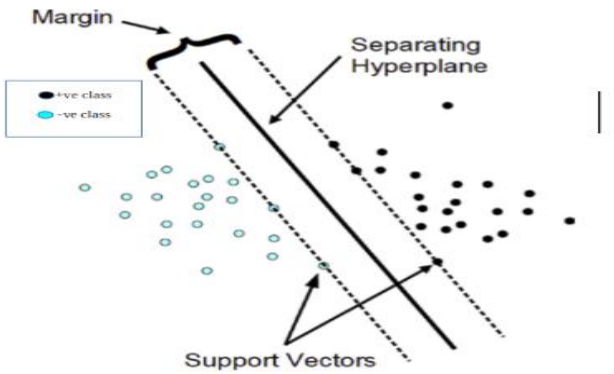

Figure 1. Rrepresentation of SVM for linearly separable dataset

SVM is a convex optimization issue, having a convex optimization goal and a set of limits that define a convex set as the feasible region. Convex set is a set of points in which a line joining any two points lies entirely within the set. SVM hypothesis with respect to machine learning model is as equation 1.

$$h_{w,b}(x) = g(w^T x + b)$$

Where g (z) = 1 if z ≥ 0, otherwise -1 (Eq.1)

Class Labels: -1 is used as label for negative class and +1 is used as label for positive class in SVM, $y \in \{-1,1\}$ Two major terminologies associated with SVM are Functional margin and Geometric margin. The hyper plane that separates the positive and negative examples is as equation 2.

$$\pi : \left(w^T x^i x + b = 0\right) \quad \text{(Eq.2)}$$

Equation of separating hyperplane; w is normal to the hyperplane, each example to be trained is denoted as x, and superscript (i) denotes ith training example and y superscripted with (i) denotes label corresponding to the ith training set. Functional margin for a hyperplane with respect to ith training example is defined as equation 3.

$$\gamma^{(i)} = y^{(i)}(w^T x^i x + b) \quad \text{(Eq.3)}$$

Geometric margin of a hyperplane with respect to ith training item set is defined as functional margin. SVM maximizes the margin (as shown in fig. 1) by deriving a suitable decision boundary separating hyperplane. So final optimization formulation is as represented by equation 4.

$$\min \frac{1}{2\,||w||^2}$$

$$Where\ y_i \left(w^T x^i x + b\right) \geq 1\ , \forall x \quad \text{(Eq.4)}$$

SVM aims at maximizing the geometric margin and returns the corresponding hyperplane. What it means is that out of all possible hyperplanes (each hyperplane has a geometric margin w.r.t. the point closest to it which is the least of all other geometric margins defined w.r.t. all other points), it chooses that hyperplane which has the maximum geometric margin.

2.2 Random Forest

It is a supervised algorithm for classification. It can be considered as superset of decision tree, where a number of decision trees are created to form a forest of trees and more the trees grow in number more robust is the system. This algorithm can be used for task like regression and classification both and this classifier can handle the missing values in most appropriate manner. Because of more number of trees system never over fit and this helps random forest to handle categorical values. Mathematically random forest can be defined as below by equation 5.

$$h(X|\theta_1), \ldots \ldots h(X|\theta_k) \quad \text{(Eq.5)}$$

On the basis of classification tree, the parameters (θ_k) are chosen randomly from a random vector (θ). F(x) represents the final classification function and is combination of family of classifiers $(\{h_k\ (x)\})$. Each tree in the forest represents the most popular class at input x, and the class with the maximum votes wins. Following is the pseudocode for Random Forest creation:

1. *Select any "k" features randomly out of total "m" features, k << m*
2. *Calculate node "d" using the best split point, among the selected "k" features.*
3. *Node "d" to be split into daughter nodes with the best split.*
4. *Until "n" number of nodes has been reached, Repeat 1 to 3 steps.*
5. *Repeat steps 1 to 4 for "n" number times in order to create "n" number of trees.*

2.3 KNN Model of Classification

The k-Nearest neighbor is a non – parametric strategy for classification, which is straightforward however proficient as a rule. For an information record t to be analyzed and classified , its k-closest neighbor are recovered and this structures an area of t. Greater part casting a ballot among the information records in the area is typically used to choose the classification for t with or without thought of separation based weighting. The estimation of k assumes a noteworthy job in grouping. The parameter k has an important role in classification, the success of classification is dependent on value of k, so an appropriate value of k must be chosen. And the most common way to choose the value of k is to execute the algorithm n times with varied values of k and the best performance is chosen [15].

A case is grouped by a lion's share vote of its neighbors, with the case being appointed to the class most normal among its K closest neighbors estimated by a separation work. On the off chance that K = 1, at that point the case is basically appointed to the class of its closest neighbor. Separation capacities utilizing condition 6 to 7.

A case is categorized by a most vote of its neighbors, with the case being given to the category commonest amongst its K nighest neighbors calculable by a distance perform. On the off chance that K = 1, at that point the case is basically appointed to the class of its closest neighbor. Separation capacities use equation 6 to 7.

$$uclidean \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \text{(Eq.6)}$$

$$Manhattan \sum_{i=}^{k} |x_i - y_i| \quad \text{(Eq.7)}$$

$$Minkowski \left[\sum_{i=1}^{k}(|x_i - y_i|)^q\right]^{1/q} \quad \text{(Eq.8)}$$

Above stated distance measures are valid only for continuous variables, for categorical variables the Hamming distance is used. When there is blend of numerical and categorical variable dataset, the numerical variables need to be normalized between 0 and 1, as shown by equation 9 for hamming distance. For example consider the following scenario, which shows that the category of X and Y is dependent on the distance value acquired as illustrated in table 1.

$$D_H = \sum_{i=1}^{k} |x_i - y_i| \quad \text{(Eq.9)}$$

x = y infers D = 0 and x ≠ y infers D = 1

Table 1. Results based on Euclidean distance calculation

X	Y	Euclidean Distance
Male	Male	0
Male	Female	1

2.4 Naïve Bayes Classifier

This is a classifier which working on the principle of probability, known as "Bayesian Theorem". This theorem works on the principle that existence of a specific feature in a cluster is irrespective to the existence of any other feature. To exemplify, a fruit can be classified as apple, if it is round, it is red and is approximately 3 inches in diameter. Although if we assume that all these three are dependent, or this happens on the availability of the other features, or these features distinctly contributes to the possibility that the fruit is an apple, through Bayesian theorem or Naïve Bayes [20]. This model can be easily build and it outperforms on large datasets and is also considered to be highly sophisticated classifier. This classifier works on the conditional probability and conditional probability can be derived through equation 10 and equation 11 while the conditional probability is as represented by equation 12.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{(Eq.10)}$$

$$P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)} \quad \text{(Eq.11)}$$

$$P(A/B) = \frac{P(A/B) * P(B)}{P(A)} \quad \text{(Eq.12)}$$

For an instance, from a population of 100 people in a school, they can be categorized as teachers and students or as a mass of females and males. And by applying conditional probability their classification can represented shown by table 2.

Table 2: Population distribution of a school

	Female	Male	Total
Teacher	8	12	20
Student	32	48	80
Total	40	60	100

The probability that a given person 'Teacher' with a condition that he is a 'Man' can be formulated as shown by equation 13:

$$P(Teacher/Male) = \frac{P(Teacher \cap Male)}{P(Male)} = \frac{12}{60} = 0.2 \quad \text{(Eq.13)}$$

III. LITERATURE REVIEW

A survey is done to study the impact of data mining and the effect of mining algorithms on healthcare data especially cardiovascular data in the direction of prediction of heart diseases. [10] Emphasizes on the amount of data which is produced in various sectors like meteorology, complex physics simulation, business, healthcare etc. And also states that the data produced is huge and rich and varied. Most important is the processing of this data and having an

infrastructure to support this flexible and dynamic data structure. The author concludes firstly, among the various tasks like locating, identifying, and understanding and citing data, data processing is considerably more challenging. Secondly, automatic way of large scale analysis has to be explored. Finally, a robust work is needed in data integration, mapping and transformation. And the procedures for probing and excavating big data are distinct from the statistical analysis done on small samples. Mining helps improvise the superiority and reliability of data, understand its meaning and provides efficient querying functions.

Chang, Chang-Lang and Chih-Hao [5], reasoned that clinical decisions are as often as possible made subject to the master's senses and experience rather than on the data-rich data concealed in the database. This preparation prompts unfortunate tendencies, botches and super helpful costs which impact the idea of organization provided for the patients. Besides, the suggestion that data exhibiting and examination instruments, for instance, information mining can deliver a learning-rich condition and can help to generally improve the idea of clinical decisions.

In their study Saima [8], expressed that data mining can improve fundamental basic decision making by discovering tests and examples in a great deal of complex data. This can enable protection to the organization for medicinal services to recognize misrepresentation and misuse cases, administration to settle on better choices in dealing with their clients and most altogether professionals to convey better benefits. The focus here is on the digitized data of healthcare which is huge and complex in structure, difficult to be handled by a traditional software system. The major challenges highlighted are, firstly, variability of structured and unstructured data likewise doctor's handwritten prescriptions, reports, diagnostic images (MRI magnetic resonance imaging), Computed Tomography (CT scan), and radiographic images (X-ray). Secondly, the presence of noisy and heterogeneous dataset produced frequently. And top of all, improvising medical measures such as superiority of service, patient data security, and reduction in healthcare cost. The summarization of data mining algorithms done by the author at a glance shows how these mining techniques can be of use in predicting diseased.

Chaithra N and Madhu [4], separated that the immense measures of information produced by medicinal services are mind-boggling and voluminous. The author states that cardiovascular diseases are one of the most common diseases now-a-days, And the cure and treatment of this is very high and not affordable by most of the patients. It is also found that the situation becomes fatal due to lack of information. So, mining is a great aid in predicting diseased prior in time. Here the author has applied three data mining algorithms J48 Decision Tree, Neural Network and Navie Bayes on a data set of 336 records. Finally concluded that Neural Network has outstripped the other mining algorithms by performing well on parameters like accuracy, sensitivity and specificity.

A survey conducted by Salma Banu and Suma Swamy [9], gives the possibility of various models accessible from 2004 to 2015 .and the distinctive information mining methods utilized. The precision got with these models is likewise referenced. It is seen that every one of the methods accessible have not utilized big data analytics. Utilization of big data analytics examination alongside information mining will give promising outcomes to get the best precision in planning the forecast model. The key target here is to recognize the key examples and highlights from the medicinal information of the patient by joining information mining strategies alongside huge information investigation to foresee the coronary illness before it causes to support the restorative specialists. The other target will be to diminish the informational indexes and increment the precision of the forecast model. A small survey is done in order to see how researcher have used data mining algorithms in prediction of cardiovascular diseases as illustrated in table 3.

Table 3: Summary of the performance of Data Mining Classifier in area of heart disease prediction.

Classifier Author	SVM	Naïve Bayes	Random Forest	K-NN	NN	J4 Decision Tree
Isra'a Ahmed Zriqat et al.,2016	76%	78.8%	93.4%	×	×	×
H. Benjamin Fredrick David et al.,2018	×	31.3%	75.6%	×	×	×
Chaithra N and Madhu B, 2018	×	74.4%	×	×	97.9 %	92.5%
Boshra Bahrami et al.,2015	82.7%	81.8%	×	82.7%		83.7%
Koppula Srinivas Rao et al.,2015	92.1%	96.5%	×	61.3%	×	×

IV. EXPERIMENTS AND RESULTS

In order to understand the importance and functioning of these data mining algorithms, a case study is done on a healthcare dataset. Here a large dataset with 76 attributes, responsible for heart diseases is considered. Out of 76 features, 14 features are of most use.

Rest of the research will focus on the result of the experiment and will also show the performance comparison of the three classification algorithms (SVM, Naïve Bayes and Random Forest) are used. The description of the data set is represented in Table 4. There are many risk components correlated with coronary heart disease and stroke. Some risk components, such as family history,

cannot be altered, while other risk components, like high blood pressure, can be altered with treatment. You will not necessarily establish cardiovascular disease if you have a risk component. In today's era where we live such a fast life as we can see the inclination of people towards junk food, the chances of acquiring heart diseases has increased a lot. And not only elderly people but now we cases of cardio arrest even in small children. Here in table 4, some of the most influential factors for heart diseases are mentioned and it is being tested that which of the factors are having maximum effect on predicting heart failure.

Table 4. Attributes of the data set used for this experiment

Sr. No	Attribute Name	Description	Type
1	Age	Age of the patient in years	Numeric
2	Sex	1 = male; 0 = female	Categorical
3	Cp	Chest pain type	Numeric
4	trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Numeric
5	chol	Serum cholesterol in mg/dl	Numeric
6	Fbs	Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Categorical
7	restecg	Resting electrocardiographic results	Numeric
8	thalach	Maximum heart rate achieved	Numeric
9	exang	Exercise induced angina (1 = yes; 0 = no)	Categorical
10	Oldpeak	ST depression induced by exercise relative to rest	Numeric
11	slope	The slope of the peak exercise ST segment	Numeric
12	Ca	Number of major vessels (0-3) colored by fluoroscopy	Numeric
13	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect	Numeric
14	target	1 or 0	Numeric

Three different experiments SVM, Random Forest and Naïve Bayes are conducted respectively in order to predict the heart diseases. There are many ways for evaluating the performance of classification algorithm, but one of the most efficient ways to evaluate through the number of instances of correctly and wrongly classified samples (data). At first glance, this gives an idea of how properly the classification is accomplished. In order to get a better insight of achieved classification, we must know which classes of data are frequently misplaced, and a convenient method to analyze the results of a classifier is the confusion matrix. The confusion matrix is applied when the ground reality known i.e. the classification task is known. This is a 2 × 2 matrix used to tabulate the results of 2 class supervised learning problems and is depicted by figure 2.

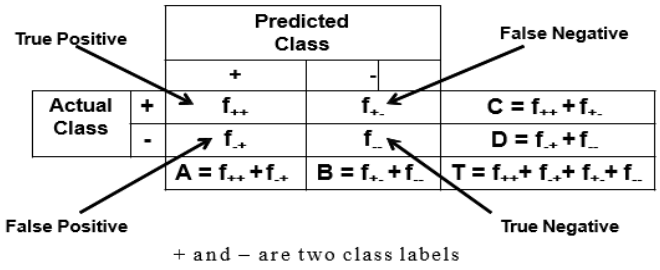

Figure 2: Elements of confusion matrix

Here the entries (i,j) represent the number of elements with class label i, but predicted to have class label j. True positive, true negative, false positive and false negative conditions are recorded. Where true positive (TP) is a number of correct positive predictions (Y), true negative (TN) is the number of correct negative predictions (N), false positive (FP) is a number of cases where the predicted category is (Y), false negative (FN) is a number of cases where the predicted category is (N). Below is the confusion matrix for the three data mining techniques applied.It is found that among the three machines used SVM is giving the best performance by giving 86.7% correct prediction. The detailed analysis of these algorithms is discussed in the following section.

Results of SVM : Figure 3, shows how SVM classifies the given healthcare dataset, the description of the attributes having effect on SVM machine is presented in Table 1. It is very clear that the hyper plane here or the support vectors maximize the margin between the diseased and the healthy ones, giving clear classification . Confusion matrices is calculated in order to measure the performance of SVM represented in Table 5.

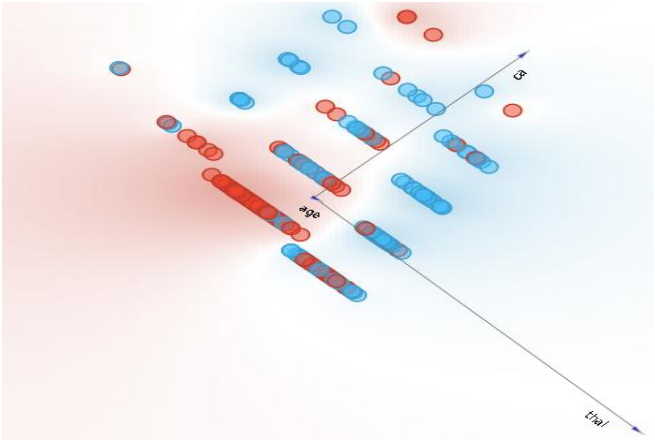

Figure 3: Linear projection of support vector machine

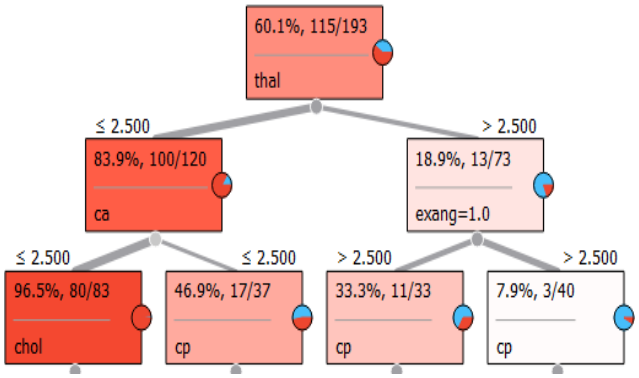

Figure 4: One portion of pythagonal forest of random forest tree generated from the dataset

Table 5. Confusion matrix generated by SVM

	Predicted0.0	Predicted 1.0	Predicted ∑
Actual 0.0	70.3 %	29.7 %	138
Actual 1.0	13.3 %	86.7 %	165
Actual ∑	119	184	303

Result of Naïve Bayes: With the help of conditional probability represented by equation 10, 11 and 12, the confusion matrices generated by Naïve Bayes algorithm is show in table 6.

Table 6. Confusion matrix generated by Naïve Bayes

	Predicted0.0	Predicted 1.0	Predicted∑
Actual 0.0	75.7 %	21.0 %	138
Actual 1.0	24.3 %	79.0 %	165
Actual∑	136	167	303

Result of Random forest: Rather than searching for the most meaningful component while splitting a node, it searches for the best component among a random subset of components. This results in a wide diversity that helps in yielding a better model. Figure 4, is the representation of one portion of pythagonal forest of Random forest tree generated from the dataset (Heart disease) used in this study of predicting heart disease. We see that most effective attribute here is thal-thallium exercise stress test (it is a nuclear scanning technique to determine if there is adequate blood flow to myocardium), secondly, the tree is affected by byca-coronary arteries (network of blood vessels that branch off the aorta to supply the heart muscles with oxygen rich blood) and the third most influencing attribute is exang-exercise induced angina. So, these attributes are going to have high impact in predicting the diseased. And it is followed by the confusion matrix generated for classification represented in table 7.

Table 7. Confusion matrix generated by Random Forest

	Predicted0.0	Predicted 1.0	Predicted ∑
Actual 0.0	76.3 %	22.1 %	138
Actual 1.0	23.7 %	77.9 %	165
Actual ∑	131	172	303

The tables above show the results of prediction for all the three classifiers respectively. 86.7%, 79.0% and 77.9% of records are correctly diagnosed that they are affected and predicted as having a disease. 13.3%, 24.3% and 23.7% records are wrongly classified as actually, they don't have but, they are classified diseased, this part of the analysis is a serious issue as, 29.7%, 21.0% and 22.1% patients are diagnosed diseased, but they are not. 70%, 75.7% and 76.3% records are correctly diagnosed as they don't have disease and predicted as not having the disease likewise for SVM, Naïve Bayes and Random Forest classifiers respectively.

Performance Matrix: Measuring the performance of data mining algorithm is an important aspect of machine learning. Valuation is basic for understanding the nature of the model, for refining parameters in the iterative procedure of learning and for choosing the most fitting method. Various execution frameworks can be gotten from the confusion matrix. Some of them are as stated below:

Accuracy: accuracy is the proportion of correct predictions, formulated as by equation 14.

$$a = \frac{f_{++} + f_{--}}{T} \quad \text{(Eq.13)}$$

Recall: recall is the proportion of "+" data points predicted as "+", formulated as by equation 15.

$$TRP = \frac{f_{++}}{f_{++} + f_{+-}} \quad \text{(Eq.14)}$$

Precision: precision is the proportion of data points predicted as "+" that are truly "-", formulated as by equation 16.

$$P^{+} = \frac{f_{++}}{f_{-+} + f_{++}} \quad \text{(Eq.15)}$$

Visual Matrices: ROC Analysis

It is observed that scalar matrices provide a week summary especially when we are dealing with non-parametric methods like neural networks and decision trees.

And also few matrices evolved from confusion matrix are sensitive to data variances like a class skew. Due to which ROC (Receiver Operating Curve) came into existence, to convey the information of confusion matrix visually, but in a more robust and spontaneous manner. This is 2-dimensional graph visually depicting the performance and performance trade-off of a classification technique. ROC gives rise to 2 other performance matrices. For constructing the ROC, true positive rate is plotted against the false positive rate using equation 17 and 18.

$$True\ Positive\ Rate = \frac{TP}{TP + FN} \quad \text{(Eq.16)}$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} \quad \text{(Eq.17)}$$

The ROC curve for the experiment carried out and the three information mining systems utilized appear in figure 4. The inclining line from the base left corner to the upper right corner signifies arbitrary classifier execution. To one side base of the irregular execution, the line is the preservationist execution area and the classifier in this locale submit a couple of false positive mistakes. The point in the base left corner is a conservative classifier technique classifying all instances as negative. It is very much clear from the ROC analysis curve represented by Figure 4, that all the three classifiers used here generate a curve above mean (0.5) and are gradually moving closer to the upper left part of the plot, which indicates a decent classification ratio. The True Negative rate of the SVM algorithm is 86.7%, True Negative rate of Naïve Bayes algorithm is 79.0% and the True Negative rate of Random Forest algorithm is 77.9%. All the algorithms have performed better in identifying negative cases as their True Negative rates are high, so it can be said that these prediction algorithms are good at identifying diseased cases, with SVM outperforming Naïve Bayes and Random Forest algorithms.

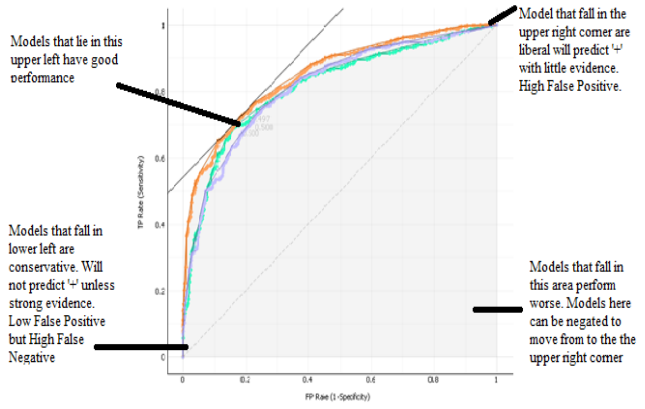

Figure 4: ROC Analysis

Table 8 shows the performance measure for the three mining algorithms used here.The performances matrix illustrates the accuracy, precision and recall for the three chosen predictive classifiers, and is observed that SVM provides the highest accuracy (88.8%) and precision (79.4%).

Table 8. Performance matrix (Comparison of the three mining algorithms)

Method	AUC	CA	Precision	Recall
SVM	0.888	0.792	0.794	0.792
Random Forest	0.829	0.772	0.772	0.772
Naive Bayes	0.862	0.776	0.775	0.776

IV. CONCLUSION

As data mining has become a strong tool in the area of classification and pattern recognition, an attempt is made to show its significance in the field of the health care sector and experiment is conducted to predict heart diseases. As per the case study three data mining algorithms, Support Vector Machine (SVM), Naïve Bayes and Random Forest are used respectively and is found that SVM outperforms the other two by giving 88% of accuracy and the other classifiers have also done a good job in identifying diseased cases. Finally, it can be concluded that these mining tools are of great importance to the health sector or any other domain, as they help classifying and providing better services. And it is suggested to attempt other data mining tools too.

REFERENCES

1. Ajad Patel, Sonali Gandhi, Swetha Shetty, Bhanu Tekwani, "Heart disease prediction using data mining", International Research Journal of Engineering and Technology (IRJET), Vol. 04, Issue 01, (Jan 2017).
2. Asem H. Shurrab, Ashraf Y. A. Maghari, "Blood diseases detection using data mining techniques", 8th International Conference on Information Technology (ICIT), (2017).
3. Orhan Yilmaz, Senol Celik, "Comparison of different data mining algorithms for prediction of body weight from several morphological measurements in dogs", The Journal of Animal & Plant Sciences, 27(1), pp. 57-64, (2017).
4. Chaithra N, Madhu B., "Classification models on cardiovascular disease prediction using data mining techniques", Journal of Cardiovascular Diseases & Diagnosis, (2018). DOI: 10.4172/2329-9517.1000348.
5. Chang, Chang-Lang, Chih-Hao, "Applying decision tree and neural network to increase quality of dermatologic diagnosis", International journal expert systems with applications, Vol. 36, Issue 2, (March 2009). pp. 4035-4041, doi10.1016/j.eswa.2008.03.007.
6. Sudhir M. Gorade, Ankit Deo, Preetesh Purohit, "Early Identification of Diseases Based on Responsible Attribute Using Data Mining", International Research Journal of Engineering and Technology (IRJET), Vol. 04, Issue 07, (2017).
7. P. Hariharan, K. Arulanandham. "Design a disease predication application using data mining techniques for effective query processing results," Advances in Computational Sciences and Technology, Vol. 10, Issue 3, pp. 353-361, (2017).
8. Saima Anwar Lashari1, Rosziati Ibrahim, Norhalina Senan, "Application of Data Mining Techniques for Medical Data Classification: A Review", MATEC Web of and Conferences MUCET (2017). https://doi.org/10.1051/matecconf/201815006003
9. Salma Banu N.K, Suma Swamy. "Prediction of heart disease at early stage using data mining and big data analytics: a survey," International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), (2016).
10. Sowmya R, Suneetha K R., "Data mining with big data," 11th International Conference on Intelligent Systems and Control (ISCO), (2017).
11. Swaroopa Shastri, Surekha, Sarita, "Data mining techniques to predict diabetes influenced kidney disease," international journal of scientific research in computer science, engineering and information technology, Vol. 2, Issue 4, (2017).
12. Ramin Assari, Parham Azimi, Mohammad Reza Taghva, "Heart disease diagnosis using data

mining techniques," International Journal of Economics & Management Sciences, Vol. 6, Issue 3, (2017). DOI: 10.4172/2162-6359.1000415
13. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, (JANUARY 2014).
14. Gerard Biau, "Analysis of a random forests model," Journal of Machine Learning Research, Vol 13, p. 1063-1095, (2012).
15. Gongde Guo, Hui Wang, David Bell, Yaxin Bi, Kieran Greer, "KNN model-based approach in Classification", International Conferences on the Move to Meaningful Internet Systems, pp 986-996, (2003).
16. Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh, "A comparative study for predicting heart diseases using data mining classification methods", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, (December 2016).
17. Koppula Srinivas Rao, Kandula Yellaswamy, Yerragopu Chandu, "A comparative study of heart disease prediction using classification techniques in data mining", International Journal of Applied Engineering Research, Vol. 10, Issues 92, (2015).
18. H. Benjamin Fredrick David, S. Antony Belcy, "Heart disease prediction using data mining techniques", Ictact Journal On Soft Computing, Vol. 09, issue 1, (October 2018).
19. Boshra Bahrami, Mirsaeid Hosseini Shirvani, "Prediction and diagnosis of heart disease by data mining techniques", Journal of Multidisciplinary Engineering Science and Technology (JMEST), vol. 2 Issue 2, (February 2015).
20. Sellappan Palaniappan, Rafiah Awang, "Intelligent heart disease prediction system using data mining techniques", IEEE/ACS International Conference on Computer Systems and Applications, (2008)
21. G´erard Biau, "Analysis of a random forests model", Journal of Machine Learning Research, vol 13, pp. 1063-1095, (2012).
22. García-Gonzalo, E., Fernández-Muñiz, Z., García Nieto, P.J., Bernardo Sánchez, A., Menéndez Fernández, M. "Hard-Rock stability analysis for span design in entry type excavations with learning classifiers", Materials, 2016, 9, 531

AUTHORS PROFILE

Ms. Archana Singh is an Assistant Professor at Gandhinagar Institute of Technology, Gandhinagar since February 2010. She has 13 years of academic experience and her area of interest includes big data, data mining, image processing.

Dr. Seema Mahajan is an Associate Professor and Head of the Department Computer Engineering at Indus University, Ahmedabad. She holds PhD and post-graduate degree in Computer Engineering. She has published many technical papers at refereed international journals and conferences. Her area of interest includes Data mining, Big Data, Machine learning and optimization Techniques.