www.ijrream.in Impact Factor: 2.9463

INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)

Design and Implementation of Heart Diseases Prediction using Naives Bayes Classifier

A. Richardwilliam[1], K. Sathya[2], S. Suguna, K. Durga[4], G. Jayabharathi[5]

[1]Assistant Professor & [2,3,4,5] UG Scholars – Dept. of Information Technology

Er. Perumal Manimekalai College of Engineering

ABSTRACT:

Data mining, a great developing technique that revolves around exploring and digging out significant information from massive collection of data which can be further beneficial in examining and drawing out patterns for making business related decisions. Talking about the Medical domain, implementation of data mining in this field can yield in discovering and withdrawing valuable patterns and information which can prove beneficial in performing clinical diagnosis. The research focuses on heart disease diagnosis by considering previous data and information. To achieve this SHDP (Smart Heart Disease Prediction) is built via Navies Bayesian in order to predict risk factors concerning heart disease. The speedy advancement of technology has led to remarkable rise in mobile health technology that being one of the web applications. The required data is assembled in a standardized form. For predicting the chances of heart disease in a patient, the following attributes are being fetched from the medical profiles, these include: age, BP, cholesterol, sex, blood sugar etc... The collected attributes acts as input for the Navies Bayesian classification for predicting heart disease. The dataset utilized is split into two sections, 80% data set is utilized for training and rest 20% is utilized for testing. The proposed approach includes following stages: dataset collection, user registration and login (Application based), classification via Navies Bayesian, prediction and secure data transfer by employing AES (Advanced Encryption Standard). Thereafter result is produced. The research elaborates and presents multiple knowledge abstraction techniques by making use of data mining methods which are adopted for heart disease prediction. The output reveals that the established diagnostic system effectively assists in predicting risk factors concerning heart diseases.

Keywords - Data Mining; Smart Heart Disease Prediction (SHDP); Web and Mobile Application; Navies Bayesian; Advanced Encryption Standard (AES); Data Collection; Classification; Prediction.

I. INTRODUCTION

Data mining process involves mining/extracting of very significant, hidden and valuable information from large databases [1]. Usually the Healthcare sector involves abundant of data related to patients, various diagnosis of the diseases etc… [2]. Nowadays the hospitals are adopting the culture of hospital IMS (information management systems) in order to handle their or patients data systematically and effectively. [3]. Large quantity of data is produced by such systems that is represented using charts, numbers, text and images. Though such sort of data is hardly employed for making any clinical decisions[4]. The current research emphasizes on heart disease diagnosis. Various techniques of data mining have been incorporated for diagnosing the disease thereby obtaining several probabilities [5]. Concerning the heart disease prediction numerous systems are being recommended which are being deployed by the means of various techniques and algorithms. Gaining quality service at affordable price remains the prime and challenging concern for the healthcare establishments. For offering quality services at par, there must be accurate diagnosis of the patients along with effective dosage of medicines. Low quality clinical diagnosis and treatment can yield in undesired and inadequate results. One solution for cost cutting by Healthcare establishments can be utilization of computer generated data or use of DSS (decision support systems). Usually the Healthcare sector involves abundant of data related to patients, various diagnosis of the diseases, resource management etc. This information or data must be further broken down by the Human services. Using computerized system, patients treatments records can be stored and using mining methods one can acquire significant information and queries concerning the hospital. Supervised and unsupervised learning are the 2 data mining methods. Supervised learning involves usage of training for learning model parameters where else no training set is required in unsupervised learning. Classification and prediction are the

basic approach of data mining. The Classification models helps in classifying distinct, disorganized data values on the other hand prediction model anticipated values that are continuous. There after making use of the analysis result for offering web / mobile application to the users. Following are the stages in the proposed approach: user registration and login based on Application, dataset collection, classification via Navies Bayesian, prediction and secure data transfer by the means of AES (Advanced Encryption Standard) and lastly output in PDF format. AES helps in transmitting user data to the database ina secured manner. From the security point of view, patient's personalized data is replaced with some mock values. The study considers and employs medicine datasets performances for predicting Heart disease in contrast to other Machine Learning techniques. The proposed technique assures to be extremely significant and effective in handling classification, resembling ML (Machine Learning) with respect to Naïve Bayesian model.

Following represent journal classification: Section 2 illustrate work of previous author. Section 3 put forth the proposed system of Heart disease classification and prediction and overview of various levels. Section 4, presents the experimental outcome. Lastly, Section 5 presents the conclusion and proposes research work for future.

II. RELATEDWORK

KaanUyar et.al, proposes some of the computational techniques for analyzing heart diseases be employing RFNN (recurrent fuzzy neural networks) and Genetic algorithm which must be assisted by medical experts for catering several parameters that may impact the decision making process. A total of 297 instances of patient data are taken into account, amidst which 45 are assigned for testing and 252 are employed for training. The testing yields an accuracy of 97.78%. With the help of heart disease testing dataset, investigations are carried out successfully. Following factors are calculated: accuracy, RMSE (root means square error), probability of the misclassification error, specificity, sensitivity, precision and F- score [6].

Anuradha Lamgund eet.al., recommends the genetic algorithm using back propagation technique approach to predict the heart disease effectively. The research utilizes numerous

input features for examining the system for heart disease prediction. Overall 13 medical attributes are utilized by the system like Gender, BP, cholesterol etc… for predicting the chances of a patient contracting heart disease [7].

Theresa Princy.R et.al., conducts a survey related to various classification techniques that can predict risk factors related to every individual considering the factors such as gender, age, BP, cholesterol, pulse rate. By means of various data mining classification techniques like NB-Naïve Bayes, Decision Tree Algorithm, KNN and NN-Neural Network etc., patients risk level can be classified. Since a lot of attributes are taken into account, high accuracy is achieved for the risk level [8].

S. Indhumathiet.al., recommends the Naïve Bayes algorithm for predicting high risk of heart disease in patients. For the training set, pre-processed data is considered. Classification and prediction forms the prime data mining phases. The classification phase involves pre-processing wherein the following tasks are performed: data cleaning, normalization, data reduction etc. The prediction phase involves classification and prediction of disease types. Hence training set includes disease type and the testing set is built using the questions. The output generated is forwarded to the doctor/specialist [9].

S. Dangare.et.al., presents the main three layers: input, hidden and the output layer. The input is fed to the input layer and the output layer projects the result acquired. Thereafter, comparison of both actual and the expected output is performed. Using the back propagation, error can be determined and weight amidst the output and prior hidden layers can be adjusted. After completion of back propagation, forward process commences and is carried forward till the error is reduced [10].

K. Pramanik.et.al., recommends a Hybrid Algorithm that being a blend of ID3 and KNN algorithm and are adopted for predicting heart disease. The data is pre-processed using the KNN algorithm hence it's also referred to as pre-processed algorithm. The pre-processed data forms the training set which is then classified in a form of tree structure. For predicting the heart disease, ID3 algorithm is implemented for the classifier. By using the KNN Algorithm, classification of in correct values is performed [11].

RishabhSaxena et.al, discusses that Heart diseases have become quiet rampant and common in the today's society. The HDD (Heart disease dataset) of Clevel and is taken in to account for exploring issues in terms of complexity and analyzing the patients effectively. CVD (Cardiovascular disease) stands for the scientific term symbolizing the heart diseases. The dataset fetched referring the medical test results of around 303 angiography patients (from Cleveland Clinic, Ohio), were imbibed on around 425 patients (from Hungarian Institute of Cardiology-Budapest, Hungary) having a frequency of 38% [12].

Ashok Kumar Dwivedi presents a significant model that identifies occurrence of heart disease in thousands of samples immediately. Herein capability of 6 machine learning techniques is being assessed for heart disease prediction. Using 8 different classification performance indices, performance of methods is evaluated. Also by using the receiver operative characteristic curve, the methods are evaluated. By deploying logistic regression, high classification accuracy of 85 % is achieved, yielding in 89% sensitivity and 81% specificity [13].

Animesh Hazra et.al, discusses that, day to day abundant amount of information is produced by the health care sector maximum of which remains unexplored and unused. But there does not exist adequate effective tools for extracting meaningful information from such data repository for carrying out detection of clinical diseases or any other task. The work targets towards summarizing few prevailing researches on heart disease prediction via data mining techniques, examining hybrid of different mining algorithms and deriving a conclusion against the best effective technique(s) [14].

Ashish Chhabbi et al. have performed study on several data mining techniques for withdrawing and exploring unknown patterns from the databases which can assist in attending complicated inquiries concerning the heart disease prediction. Collection of dataset is done using the UCI repository. Naive Bayes and improvised k-means algorithm have been deployed. According to the results generated, advanced k-means algorithm, yields in high accuracy in contrast to simple k-means (in which no: of clusters are already defined)[15].

Kamal Kant et al. recommends a heart disease prediction model by employing Naïve Bayes data mining technique. The technique is a statistical classifier that doesn't make the attributes dependent on each other. To decide upon the class, the posterior probability must be highly raised. Here, Naïve Bayes classifier yields great performance and is quiet efficient for predicting disease in statistical probability and real time expert system, then comes the Neural Network and Decision trees [16].

Sharan Monica L et.al carries out a survey on prevailing techniques of KDD (knowledge discovery in databases by adopting the following mining techniques namely - J48, NB Tree and simple CART for accurately predicting heart disease using least no: of attributes in the WEKA tool. J48 being an open source Java application of C4.5 that acquires information for making decisions. Naive Bayes (NB) classifier builds models mostly for continuous dataset using predictive proficiencies. Data relationships that are significant can be promptly projected using CART (Classification and Regression Trees). The above mentioned 3DT algorithms are deployed using WEKA. CART exhibited an accuracy of 92.2% and J48 was the fastest one, framed in just 0.08 sec[17].

Sumitra Sangwan et.al have built a hybrid algorithm that employs k-means and A-priori algorithm that are capable of extracting abundant data along with significant information. Initially, clustering is performed via k-means clustering

algorithm. Thereafter, frequent item-sets are determined using A-priori algorithm along with extraction of frequent term-sets for Boolean association rule. The approach of "bottom up" is imbibed where in frequent subsets are expanded with one single item at a time and testing of entire groups of candidates is performed against the data. With the output generated its elucidated that clustering followed by A-priori results in high performance for heart disease prediction [18].

Rishi Dubey et.al carries out the heart disease prediction by surveying various data mining techniques. It was exhibited from various researches that considering the 'accuracy' parameter, hybrid techniques surpasses a single classification technique. It was ascertained that for achieving prediction, neural network proves to be quiet effective. The system yields

in assuring outcome when trained well using genetic algorithms. Moreover, appropriate treatment could be selected for patients in the near future rather than merely anticipating the likeness of contracting heart disease in individuals [19].

Monika Gandhi et.al, explores and carries a study for determining how data mining techniques along with other techniques helps in drawing out unknown patterns from large databases enabling the healthcare establishments in decision making. The data mining classification techniques such as decision trees, Neural network and Naive Bayes are employed for data discovery, extraction and classification. [20].

Arun R et.al discusses about CVD or cardio vascular illness. Presence of numerous sort of symptoms and strong interventions becomes quiet complicated and astonishing. Secure execution and analysis exhibits Cloud computing feasibility via Naive Bayes and AES. By making use of social networking site, restrictive patient's information can be shared with patient's family members. The AES encryption algorithm safeguards sensitive but accessible patient's information [21].

III. PROPOSED WORK

A. Overview

As per today's advanced and hi-tech living style, majority of the people are contracting heart disease which gives a sudden jolt to an individual that at times one lacks time to get treated immediately. Hence its very much essential that timely and early diagnosis is performed which being quiet challenging concern for the medical association. Poor and in correct analysis carried out by the hospital can being down its reputation and working. The research focuses on to build cost cutting and effective approach by the means of data mining techniques so that DSS (decision support system) can be enhanced. Predicting heart disease with the help of numerous attributes/symptoms is quiet complicated. The present research utilizes Naives Bayesian - data mining classification technique for effectively enabling heart disease diagnosis and thereby offering appropriate treatment. Supervising different medical factors and post operational period stands very crucial. AES encrypts the patients'

records/data and save it in database. The results generated reveals that the diagnostic system built successfully predict the risk level associated with heart diseases.

B. Data Collection

Using the UCI dataset, Collection of medical data of patients with heart diseases is carried out. Throughout issues/matters are assumed for CAD and registered for angiography. Every patient's attributes are being assembled such as demographic, historic and laboratory features such as sex, age, hypertension, smoking history, diabetes mellitus, chest pain type, dyslipidemia, random blood sugar, low and high density lipoprotein, cholesterol, triglycerides, systolic and diastolic blood pressure, weight, height, BMI (body mass index), central obesity, waist circumference, ankle–brachial index, duration of exercise, METS obtained, rate pressure product, recovery duration with persistent ST changes, duke treadmill test and angiography result.

C. User Registration and Login

Data mining techniques have proven to be extremely advantageous in healthcare sector by assisting in diagnosis and identification of diseases effectively, protecting and enhancing patient's life span, helping the medical care takers in treatment plans and cutting down medical cost. First is the process of user registration wherein the user must fill up the registration form through a mobile application. After successful completion of user registration, using the systemIP (internet protocol) address the user can login anytime by using his/her own username and password. Every registered user's credentials are saved in the database. After this the complete symptoms list is given including the affected clinical features like age, sex, cholesterol, sugar, ECG, chest pain, Rest blood pressure, etc.

D. Classification

This classification algorithm basically employs conditional independence, this implies that value of an attribute for an available class is not dependent on other attribute values since the algorithm relies upon the Bayesian theorem.

E. Navies Bayesian based Classification

A Naive Bayesian (NB) classifier, also termed as "independent feature model" relies upon the Bayesian theorem and acts as a simple probabilistic classifier having powerful independence hypothesis. Generally, the NB classifier presumes that the existence/absence of a specific class feature is independent of the existence of the other class feature. NB classifiers usually perform in supervised learning. The classifier is based on conditional independence, this implies that value of a variable for an available class is independent of other existing variable value. In case of high dimensionality input, the classifier is highly appropriate. Using Naïve Bayesian, models having predictive potentials can be designed.

Algorithm

Step 1: Say D represents the training set and each record denoted by n-dimensional attribute vector, this means X=(x1, x2…, xn),predicting n measurements from n attributes (say A1 toAn.)

Step 2: Consider m no: of classes for prediction (sayC1,C2 Cm)

By Bayes' theorem:

$$P(Ci \mid X) = P(X \mid Ci) * P(Ci))$$

Step 3: Since P(X) being a constant for every class, hence P (X|Ci)* P(Ci) must be maximized.

Step 4: Thereafter class conditional independence is presumed.

Thus,

$$P(X| Ci) = P(x1| Ci)* P(x2|Ci). \quad P(xm|Ci)$$

Step 5: For predicting class of X, P(X|Ci)P(Ci) is computed for every class Ci.

Naive Bayes classifier predict that class label of X = Ci class if

$$P(X|Ci)P(Ci) > P(X|Cj)P(Cj)$$

for $1 \leq j \leq m, j \neq i$

A. Prediction

With the launch of automated medical diagnosis system, there is high development in the medical domain and at the same time cost consumption has reduced. There are numerous factors prevailing for heart attack diagnosis and mostly patient's test records are being referred and analyzed for carrying out the diagnosis. For enhancing the diagnos is process, experience and knowledge of various medical experts/doctors as well as patient's medical screening data is being collected in databases, resulting in an extremely significant system. With the blend of clinical decision support and computerized patient records, the medical faults can be reduced, patient's safety can be enhanced, variation in unwanted practices can be minimized there by improvising throughout patient's results. In addition, by making use of heart disease levels predictions, a prediction algorithm is being established.

B. Security for AES

AES is designed transmitting the data to the database in a secure manner. AES is a very popular and demanding encryption algorithm that is utilized very often. It uses bytes insteadofbitsforcarryingoutalltheoperations.Forinstancea plain text of 128 is assumed to be of 16 bytes framed or designedinamatrixoffourcolumnsandfourrowstocarryout any processing. For performing cryptography, this encryption algorithm is utilized by software and hardware both. Also no practicalcryptanalyticattackhasbeenreportedyetagainstAES algorithm. The result being generated in a PDF format. All the patient's details are encrypted using the above encryption algorithm.

IV. RESULT AND DISCUSSION

The section put forth a standard model for performing classification and prediction of composite web services. Data mining process involves mining/extracting of very significant, hidden and valuable information from large databases of health care sector. It's illustrated in the work that higher results are achieved when hybrid of data mining techniques are deployed instead of implementing one single mining technique on a data set. The partitioning of data

samples is done using tenfold, every fold has been imbibed in testing, and any folds left were utilized for training at time of cross validation.

Table 1 represents proposed classification techniques

S. No	No. of Techniques	Accuracy (%)	Time(s)
1	Sequential Minimal Optimization (SMO)	84.07	0.02
2	Bayes Net (BN)	81.11	0.02
3	Multi-Layer Perception (MLP)	77.4	0.75
4	Navies Bayesian (NB)	89.77	0.01

Table 2 depicts a comparison of AES (Advanced Encryption Standard) is contrast to PHEA (Parallel Homomorphic Encryption Algorithm). The AES algorithm offers effective security in comparison to rest others.

S. No	No. of Techniques	Security (%)
1	Advanced Encryption Standard (AES)	92.21
2	PHEA	98.2

V. CONCLUSION

Data collection is carried out using numerous sources that are primary factors responsible for any sort of heart disease and thereby using a structure the database is constructed. The research focuses on establishing SHDP (Smart Heart Disease Prediction that takes into consideration the approach of NB (Naive Bayesian) classification and AES (Advanced Encryption Standard) algorithm for resolving the issue of heart disease prediction. its revealed that in regard to accuracy, the prevailing technique surpasses the Naive Bayes by yielding an accuracy of 89.77%in spite of reducing the attributes. AES yields in high security performance evaluation in comparison to PHEA (Parallel Homomorphic Encryption Algorithm).

VI. FUTUREWORK

Application developers should work together with health care professionals and researchers to deliver disease apps which improve health care outcomes. An overall of research process to reduce delays would help to ensure application based heart disease prevention research is not entirely left behind by advances in technology.

REFERENCES

1. Purushottama.C, KanakSaxenab, RichaSharma(2016),"Efficient Heart Disease Prediction System", Elsevier, Procedia Computer Science, No. 85, pp. 962 –969.
2. Kipp W. Johnson, BS, Jessica Torres Soto, MS, Benjamin S. Glicksberg (2018), "Artificial Intelligence in Cardiology", Elsevier, Journal Of The American College Of Cardiology, Vol. 71, No. 23, pp. 2668 -2679.
3. ChalaBeyene, PoojaKamat (2018), "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques",International Journal of Pure and Applied Mathematics, Vol. 118, No. 8, pp. 165-174.
4. Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago(2016),"BuildingaCardiovascularDiseasePredictiveModelusing Structural Equation Model & Fuzzy Cognitive Map", IEEE, ICFS (FUZZ), pp.1377-1382.
5. Shalet K.S, V. Sabarinathan, V. Sugumaran, V. J. Sarath Kumar (2015), "Diagnosis of Heart Disease Using Decision Tree and SVM Classifier", International Journal of Applied Engineering Research, Vol. 10, No.68, pp.598-602.
6. KaanUyar,Ahmetİlhan(2017),"Diagnosisofheartdiseaseusinggenetic algorithm based trained recurrent fuzzy neural networks", Elsevier B.V, ICTASC, pp.588–593.
7. AbhishekRairikar, VedantKulkarni, VikasSabale, Harshavardhan Kale, AnuradhaLamgunde "Heart Disease Prediction Using Data Mining Techniques" © IEEE, ICCC, 2017, p.p.1-8.
8. Theresa Princy. R, J. Thomas "Human Heart Disease Prediction System using Data Mining Techniques", © IEEE, ICCPCT, 2016.