Prediction of Diseases Using Data Mining Techniques and Patient History

[1]Jayshree Khade, [2]Akshay Jadhav, [3]Vivek Patil

[1,2,3]M.Tech(Project Management) Student, COEP, Pune, India.

ABSTRACT: Health care industry comprises of data which are frequently enormous and mixed in the fact that it contains diverse variables and missing values too for some attributes. These days, learning from such large data has turned into an important task. Recognition of illness is an essential however it is an unpredictable undertaking as it needs to consider a large number of parameters that should be performed minutely and the right computerization or automation is extremely attractive. Each individual can't be similarly capable as it requires a considerable measure of aptitude like as specialists. A computerized framework is required for improvement in medicinal care and it can likewise lessen the cost of treatment. In this project, we have developed a system that can productively analyze the infection with respect to the given parameter about patient's health. Data mining can be utilized to build models from social health care industry data like for example, diabetic patient dataset, heart disease dataset and so forth. Data mining in these medical data sets has successfully converted raw data into useful information. This information helps the medical experts in improving the diagnosis and treatment of diseases.

Keywords: Data mining, Decision tree, Naïve bayes classifier, Random forest, Heart disease, Diabetes.

I. INTRODUCTION

Data mining (DM) has a great potential for the health care industry to allow health care frameworks or systems to efficiently utilize information and examination to recognize wasteful aspects and best practices that enhance health and decrease costs. A large variety of gadgets are available in market for effective and efficient monitoring of health.

1.1 Introduction to health care system:

Health care is important to any society in order to improve the betterment of the society. Advanced or improved healthcare can help people live longer and to be more productive. It is imperative in different routes as it serves the general prosperity of individuals. It significantly takes a glance at different medical problems including preventions, vaccination and treatment in addition to other things. Our wellbeing is the most important component for living long life. In spite of the fact that fresh eating habits and exercise are a decent establishment for good wellbeing, keeping up wellbeing is a significantly more complex and possibly it is costly task. The principle objective of health care services is to improve the wellbeing by means of the treatment, finding and diagnosis of sickness, harm or whatever other mental and physical disabilities in individuals.

1.1.1 Motivation

In last few decades population is increased drastically. Henceforth there is absence of gifted specialists as the specialists and specialists accessible are not in extent with the present population. Additionally some time indications are dismissed. Also a lot of money is required if we wish to take treatment from a skilled doctor. To overcome these problems there is need of an automated disease prediction system which will solve these issues.

1.2 Use of health care data for prediction of]disease:

The medicinal service industry has produced a lot of information, driven by record keeping, asset and administrative necessities and patient care periodically of past years. While most information is kept in printed version shape, the present float is toward fast digitization of this information. Dealt with by obligatory needs and the possibility to upgrade the nature of medicinal services influences limiting the costs, these enormous amounts of information otherwise called “big data‟ hold the guarantee of supporting an extensive variety of medical purpose including among others clinical decision support[6]. Huge information in Health cares is overpowering a result of its volume as well as on account of the mixed qualities of information sorts and the speed at which it must be overseen. The aggregate of information identified with patient social insurance and prosperity make up “big data” in the health care services industry [7]. Huge information in Health cares is overpowering a result of its volume as well as on account of the mixed qualities of information sorts and the speed at which it must be overseen. The aggregate of information identified with patient social insurance and prosperity make up “big data” in the health care services industry [10].

II. PROCEDURE WITH CALCULATIONS

2.1 Introduction to data mining: DM is the procedure of extraction of helpful and useful data from large information. It permits clients who of it to information from different distinctive measurements, arrange it, and abridge the connections recognized between the information. There are various techniques are available. These techniques can be used for further processing of the collected data and thereby increasing their business revenue.

2.2 Data mining techniques: DM techniques are set of algorithms which are expected to locate the concealed knowledge from the information. Utilization of DM techniques will absolutely rely on upon the issue we would be solving. A portion of these algorithms or DM techniques are listed as following

A. Naïve Bayes
B. Decision tree
C. Random forest

A. Naïve Bayes classifier:

Naïve Bayes classification is based on Bayes theorem. Naïve Bayes is algorithm used for binary that is having two classes and having multi-class classification problems. This technique is easiest to understand than among all the techniques of data mining when it is described using categorical or binary input values. It has the name called Naïve Bayes because it consists of calculations of the probabilities for each hypothesis.

Representation that is used for naïve bayes model: Naïve Bayes is represented in the form of probabilities. There are lists of probabilities which are stored to file to be used for building naïve bayes model. This includes Class Probabilities These are the probabilities of each class that are available in the preparation dataset. Conditional Probabilities These are contingent probabilities of each info esteem given each class value.

Bayesian Theorem: In likelihood of hypothesis and insights, Bayes' hypothesis (then again Bayes' law or Bayes' administer) represents the likelihood of an event, in light of earlier information of conditions that may be identified with the event. Given training data X, posterior probability of a hypothesis H, $P\left(\frac{H}{X}\right)$ is as follows according to the Bayes theorem. $P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)\times P(H)}{P(X)}$

Algorithm: The algorithm of Naive Bayes which is based on Bayesian theorem is as follows: *Consider the data set. Each data sample from data set is represented by n dimensional feature vector, let's say X (x1, x2... xn) representing n measurements made on the sample from n attributes respectively A1, A2, An.*

Suppose that there are m classes, let's say C1, C2...Cm. Now given an unknown data sample X, the classifier will predict that X belongs to the class which is having the highest posterior probability, conditioned if and only if

$P\left(\frac{Ci}{X}\right) > P(\frac{Cj}{X})$ For all 1< = j< = m and j! = I

$$P\left(\frac{C}{X}\right) = \frac{(P\left(\frac{X}{C}\right) \times P(C)}{P(X)}$$

$P\left(\frac{C}{X}\right) = P\left(\frac{X1}{C}\right) \times P\left(\frac{X2}{C}\right) \ldots . \times P(\frac{Xn}{C}) \times P(C)$

Thus we maximized $P(\frac{Ci}{X})$

B. Decision tree

A decision tree is a standout amongst the most usually utilized information mining method since its model is anything but difficult to peruse and comprehend for its clients. In this system, the foundation of the decision tree is any straightforward condition or any question which is having various answers. Each answer of the condition or question then prompts an arrangement of taking after inquiries or conditions which thus help us to decide the information so that at last we can settle on our official choice in light of it to get came about esteem.

Algorithm of decision tree is as below

Let's say D be the training tuples of data partition for which the decision tree is to be constructed. Now follow below steps

Input: D is the data partition which is a set of training tuples and their associated class labels,attribute list, the set of candidate attributes;

Algorithm: *Create a node N. Then if tuples in D are all of the same class, C then return N as a leaf node labeled with the class C. If attribute list is empty then return N as a leaf node labeled with the majority class in D. Apply Attribute selection method (D, attribute list) to find the "best" splitting criteria*

label node N with splitting criterion. If splitting attribute is discrete-valued and multiway splits allowed then Attribute list ← attribute list − splitting attribute. For each outcome j of splitting criterion now let Dj be the set of data tuples in D satisfying outcome j. If Dj is empty then attach a leaf labeled with the majority class in D to node N else attach the node returned by Generate decision tree (Dj, attribute list) to node N then return N.

C. *Random Forest*

Random Forest calculation is one of the best among order calculations as it can characterize a lot of information with closest. Irregular Forest is a group learning technique. It can be additionally considered as a type of closest neighbor indicator. It build an expansive number of choice trees at the season of preparing model and after that it yields the class that is the normal of all the yield of produced classes worked by model person.

Algorithm: *Assume N be the number of cases in the training set. Then, sample of these N cases is taken randomly but with replacement. This sample will be treated as the training set for growing the tree. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on this m is used to split the node. The value of m is held constant while we grow the forest. Each tree is grown to the largest extent possible and there is no pruning. Predict new data by aggregating the predictions of the n tree trees (i.e., majority votes for classification, average for regression.*

2.3 Dataset details

Dataset has been collected through various web sources. It consist of a number attributes for each disease. Also for each disease the attributes vary as every disease has its specific symptoms. Each attribute has its own different values.

3. Data mining algorithm of Prediction of disease:

3.1 Literature Review:

Past reviews give an overview of present systems of learning revelation in databases utilizing information mining procedures that are being used in today's therapeutic research especially in heart and diabetes illness expectation.

3.1.1 Heart disease:

Four calculations like Naïve Bayes, Decision tree, K closest neighbor and SVM were carried out by RovinaDbritto *et al*. [1]. In this they talked about the significant characterization strategies utilized as a part of information digging for forecast of heart disease and with the assistance of precision examination they have demonstrated that SVM is superior to anything other two techniques when we have vast informational index of sections. Diagnosis of heart disease by using different data mining algorithms such as SVM, ANN and Decision Tree and RIPPER classifier, Naive Bayes has been carried out by Kumari [2]. K-means clustering strategy which is connected to discover groups in information which are additionally used to evacuate concealed structures identified with heart patients was utilized by Fatima *et al*. [3].

3.1.2 Diabetes:

To make compelling finding of any sickness, the disclosure of learning or data from medicinal databases is essential Yang Guo *et al*. [4] expressed it in very effective way. The dataset utilized was the Pima Indian Diabetic Database (PIDD). Preprocessing was done to enhance the nature of informational collection. Classifier was connected to the dataset which is changed to develop the Naïve Bayes show. At long last they utilized weka to do reenactment and the precision of their subsequent model was 72.3%. DM method through RapidMiner for diabetes information investigation and diabetes expectation demonstrate was utilized by Han *et al*. [8]. A choice tree was utilized for forecast of diabetes with 72 % of precision. ID3 Algorithm was additionally utilized for this reason which gave 80 % exact outcomes so conclusion is that ID3 gives more precise outcome. Utilizing unpleasant sets on the PIDD surprisingly He utilized the equivalent recurrence binning criteria for a similar reason was proposed by Breault [9].

A model in light of Neural Network and Fuzzy k-Nearest Neighbor Algorithm was proposed by Pradhan *et al*. [5]. They first pre-handled the information by disposing of the records containing the missing qualities from the PIDD.

3.2 Methodology:

Research methodology of this project is as following
A model is built by using the dataset which is called as training dataset. This model is then used for actual prediction. Then depending on the given values or inputs to the symptoms, probability of these values is calculated for each distinct value which is also called as class in the response attribute. Then the class having the highest probability will be the value for the predictor attribute for the given input.

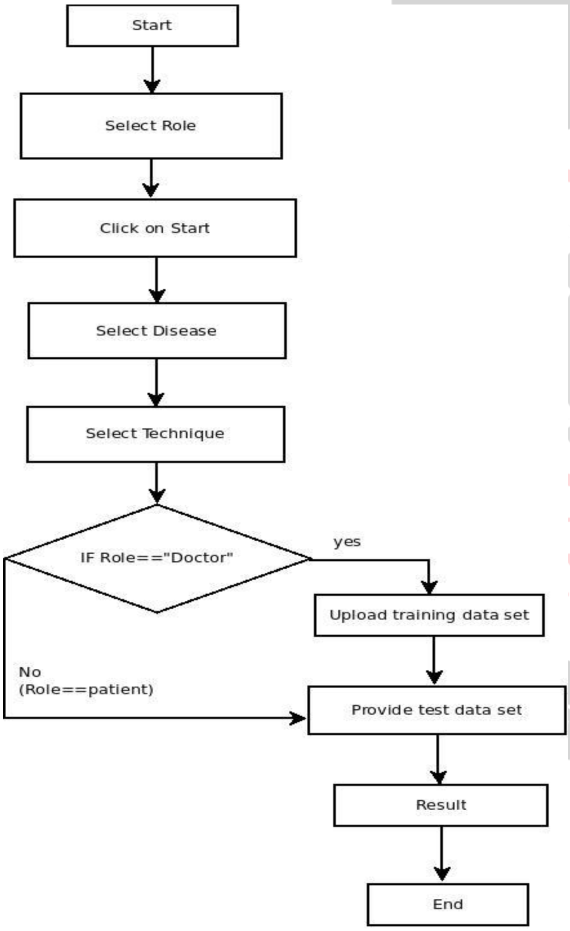

Figure 3.0 methodology.

3.3 Naïve Bayes classifier

Naïve bayes classifier has been implemented in R language and following are the steps for its implementation:
1. Firstly load the package named “e1071”.
2. Then load the dataset into one variable.
3. Split the dataset into two areas. One is of 80% of the first dataset and other is of 20% of the first dataset. 80% data is dealt with as preparing data and 20% data is dealt with as test data. This was accomplished for checking the precision of the model.
4. Then naïve bayes model is built by loaded data set using naïve bayes function of the loaded package
5. Data for which the prediction has to carry out was retrieved from the form. Then using predicts function of the same package the model gives the result of the prediction. The “e1071” package is used for constructing the naïve bayes model. It has various built in functions for constructing models for prediction like svm, plot.svm, naivebayes etc.

3.4 Decision tree: Decision tree has been implemented in R language and following are the steps for its implementation:
1. Firstly load the package named “tree”.
2. Then load the dataset into one variable.
3. Split the dataset into two areas. One is of 80% of the first dataset and other is of 20% of the first dataset. 80% data is dealt with as preparing data and 20% data is dealt with as test data. This was accomplished for checking the precision of the model.
4. Then decision tree model is built by loaded data set using tree function of the loaded package.
5. Data for which the prediction has to carry out was retrieved from the form. Then using predicts function of the same package the model gives the result of the prediction. The “tree” package is used for constructing the classification and regression tree. It has various built in functions for constructing tree that can be further used for prediction like tree, plot. tree etc.

3.5 Random forest: Random forest has been implemented in R language and following are the steps for its implementation:
1. Firstly load the package named “random Forest”.
2. Then load the dataset into one variable.
3. Split the dataset into two areas. One is of 80% of the first dataset and other is of 20% of the first dataset. 80% data is dealt with as preparing data and 20% data is dealt with as test data. This was accomplished for checking the precision of the model.
1. Then random forest model is built by loaded data set using random Forest function of the loaded package. 2. Data for which the prediction has to carry out was retrieved from the form. Then using predicts function of the same package the model gives the result of the prediction. The “random Forest” package is used for constructing the classification and regression tree. It has various built in functions for constructing tree that can be further used for prediction like random Forest and decision tree etc.

III. RESULT ANALYSIS

3.1 Results of actual application: The results obtained by implementing DM techniques like naïve bayes, random forest and decision tree on datasets of diabetes, heart disease in actual application are explained as following

3.1.1 Error matrix of each disease: Error matrix is a table that is used to represent the performance of a classification model which is based on a set of test data for which the true values are known. Horizontal attribute values indicate actual value for the response variable and vertical values indicates predicted value.

This is the result obtained in R language. “No” and “Yes” in the below tables represents the value of the response attribute. Also these values in the gray colored are the actual values in the dataset and the one which are with italic are the predicted values by the model built by the actual application.

3.1.1.1 Diabetes: The below table shows the error matrix which is obtained by applying the three above mentioned techniques on diabetes dataset. Horizontal row which is filled with gray color indicates the technique which is applied on the diabetes dataset.

Technique	Naïve Bayes		Random Forest		Decision Tree	
Attribute value	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
No	81	17	81	17	80	18
Yes	20	36	22	34	26	30

Table 3.1 Error matrix for diabetes

3.1.1.2 Heart disease:

Technique	Naïve Bayes		Random Forest		Decision Tree	
Attribute value	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
No	47	20	51	16	55	12
Yes	19	40	28	31	34	25

Table 3.2 Error matrix for heart disease

.**3.2 Accuracy table:** The dataset has been divided into training data set and testing data set. Training data set consists of 80% of whole data set and test data set contains 20% of the whole data set. Below is the results obtained from R.

Technique\Disease	Heart disease	Diabetes
Naive Bayes	**69%**	**79%**
Random forest	**65%**	**75%**
Decision Tree	**65%**	**72%**

3.3 Comparison between results of weka and R language: The comparison of results obtained in R language and weka tool for dataset of three diseases with techniques like naïve bayes, decision tree and random forest etc. is as follows.

3.3.1 Naïve Bayes Classifier: The comparison of the accuracy of naïve bayes classifier in R language and weka tool. The table shows results obtained in R language are better than the weka tool.

Tool\Disease	Heart Disease	Diabetes Disease
Weka	71%	76%
R	69%	79%

Table 3.5 Comparison for naïve byes classifier

3.3.2 Random Forest: Below is the comparison of the accuracy of random forest in R language and weka tool. The table shows results obtained in weka tool are better than the R language.

Tool\Disease	Heart disease	Diabetes Disease
Weka	96%	91%
R	65%	75%

Table 5.6 Comparison for random forest technique

3.3.3 Decision tree: Below is the comparison of the accuracy of random forest in R language and weka tool. The table shows results obtained in R language are better than the weka tool.

Tool/ Disease	Heart disease	Diabetes Disease
Weka	70%	74%
R	65%	72%

Table 5.7 Comparison for decision tree technique

IV. CONCLUSION

The prediction system has been developed using three different algorithms namely naive bayes, decision tree and random forest. This study has shown that naïve bayes gives more accuracy than other DM techniques. This system takes input for given symptoms and then builds model according to selected technique and for the selected disease and after then predicts whether a patient is likely to have a selected disease or patient is normal. The performance of the achieved results from the actual developed system was validated using weka DM tool.

The designed system takes commitment for various symptoms of a particular ailment and after that predicts whether the patient is having the ailment or not by building the model. This system has worked upon two contaminations specifically diabetes, and heart disease. This framework boosts the doctor for finding illnesses and furthermore supports the new doctor who doesn't have much understanding for conclusion of illness precisely and treated in like manner.

REFERENCES

[1] Aqueel Ahmed and Shaikh Abdul Hannan," Data Mining Techniques to Find Out Heart Diseases An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2012.

[2] Yang Guo , Guohua Bai , Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, "Using Bayes Network for Prediction of Type-2 Diabetes".

[3] Karegowda, A.G., Jayaram, M.A., Manjunath, A.S.,"Cascading K-means Clustering and KNearest Neighbor Classifier for Categorization of Diabetic Patients", in International Journal of Engineering and Advanced Technology (IJEAT) ISSN 2249 – 8958, Volume-1, Issue-3, February (2012).

[4] Han, J., Rodriguze, J.C., Beheshti, M., "Diabetes data analysis and prediction model discovery using RapidMiner", Second International Conference on Future Generation Communication and Networking.96-9 (2008).

[5]. Pradhan, M., Sahu, R.K." Predict the onset of diabetes disease using Artificial Neural Network". Intl J Comp Sci & Emerging Tech. 2 303-11 (2011).

[6] .Lemke F, Mueller J-A., "Medical data analysis using self-organizing data mining technologies", Systems Analysis Modeling Simulation. 2003; 43(10) 1399–408.

[7] Lamia AbedNoor Muhammed," Using Data Mining technique to diagnosis heart disease", IEEE, International conference on statistics in science, Business and Engineering, pp.1-3, 2012.

[8] Han, J., Rodriguze, J.C., Beheshti, M., "Diabetes data analysis and prediction model discovery using RapidMiner", Second International Conference on Future Generation Communication and Networking.96-9 (2008).

[9] Breault, Joseph L., Colin R. Goodall, and Peter J. Fos. "Data mining a diabetic data warehouse." *Artificial intelligence in medicine* 26, no. 1 (2002) 37-54.

[10] Duhamel A, Nuttens MC, Devos P, Picavet M, Beuscart R. A preprocessing method for improving data mining techniques. Application to a large medical diabetes database. Stud Health Technol Inform. 2003; 95 269–274.