DATA MINING AND REPORTING IN HEALTHCARE

Divya Gandhi [1], Pooja Asher [2], Harshada Chaudhari [3]

[1,2,3] *Department of Information Technology, Sardar Patel Institute of Technology, Mumbai,(India)*

ABSTRACT

The amount of raw data stored in corporate databases is exploding. Trillions of bytes of healthcare data is available and needs to be processed and stored. In healthcare data mining is becoming increasingly popular. The use of data mining in medical diagnosis is becoming very common and has been used widely in diagnosis of cancer, tumors, hepatitis, diabetes and cardiovascular diseases etc. The main aim of this paper is to build, an predictive system which analyses certain parameters and predicts whether a person is at risk for diabetes and cardiovascular diseases using naïve bayes algorithm. The system helps to control and diagnose of disease, by regularly monitoring the risk factors for each person.

Keywords: BMI, Data Mining, Diabetes, Naive Bayes

I. INTRODUCTION

Data mining has been used intensively and extensively by many organizations. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods.[10]

A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends[11]. The algorithm uses the results of the analysis to find the optimal parameters for creating a data mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics. Choosing the best algorithm to use for a specific analytical task can be a challenge. Different types of data mining algorithms:

Classification algorithms, Regression algorithms, Segmentation algorithms, Association algorithms, Sequence analysis algorithms .For predicting diseases there are different algorithms like decision tree, naive bayes, Multilayer perception, Multiclass Classifier. Each algorithm has accuracy and error rate. Among these algorithms the proposed system uses naïve Bayesian algorithm.

Healthcare has tremendous amount of data which is present in unorganized form. It contains large amount of sensitive data which can help to integrate, store and analyze patients data. The incorporation of technologies in health care can improve quality of analysis, reduce errors, streamline processes, and improve performance. Data-mining technology is used in various fields in the health care industry, primarily for performance assessment and quality improvement.[12]

The paper is organized as follows. In Section II, the different data mining techniques are introduced. Section III describes the proposed system and Section VI contains information about the prototype. Finally, conclusions are drawn in Section V.

II. DATA MINING ALGORITHMS

1. Association Algorithm: It finds correlations between different attributes in a dataset.

Apriori algorithm: The Apriori Algorithm is an frequent item set algorithm. The algorithm analyzes a data set to determine which combinations of items occur together frequently. Frequent Itemsets are the sets of items that have minimum support. i.e if{AB} is a frequent itemset,both {A} and {B} should be frequent itemset.[8]

2. Classification Algorithm: It predicts one or more discrete variables, based on the other attributes in the dataset.

2.1 .Naive Bayes: Naïve Bayes (NB) based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions .Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. Given a set of variables, X = {x1,x2,x3,xn} we want to construct the posterior probability for the Cj event on a set of possible outcome Cj among a set of possible outcomes of C={c1,c2,c3..,cn}. [3]

$$P(C|x1,...,xd) = \frac{P(C)*P(x1,....,xd|C)}{P(x1,....,xd)} \quad \text{.....(1)}$$

2.2 .ID3:Iterative Dichotomiser 3 is a decision tree learning algorithm which is used for the classification of the objects with the iterative inductive approach. In this algorithm the top to down approach is used. The top node is called as the root node and others are the leaf nodes. Each node requires some test on the attributes which decide the level of the leaf nodes. The tree calculates entropy and information gain at each level.[9]

2.3. J48: J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc .[6] This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

2.4 . SVM(Support Vector Machine):Support vector machines are a moderately type of learning algorithm, originally introduced. Naturally, SVM aim at pointed for the hyper plane that most excellent separates the classes of data.[5] SVMs have confirmed the capability not only to accurately separate entities into correct classes, but also to identify instance whose establish classification is not supported by data.

Table 1: Survey of accuracy measures for each data mining algorithms

Algorithm	Accuracy	Reference
Naive bayes	81.48%	[3]
ID3	81.11%	[2]

SVM	74.1%	[5]
J48	78.11%	[6]

III. PROPOSED SYSTEM

The new system uses Data Mining for analysis of past data and to predict the occurrence of Diabetes or Cardiovascular diseases in a person. It is a text-based prediction system for medical records. The system takes various factors, listed in Table 2, as inputs from a patient and uses Naive Bayes Algorithm to calculate probability of risk for the disease. The system also analyses which factors affect the risk probability the most and then suggests personalized solutions to control these factors. Also, the proposed system monitors the personal health records of the patients to control and compare over a long time.

3.1 Architecture

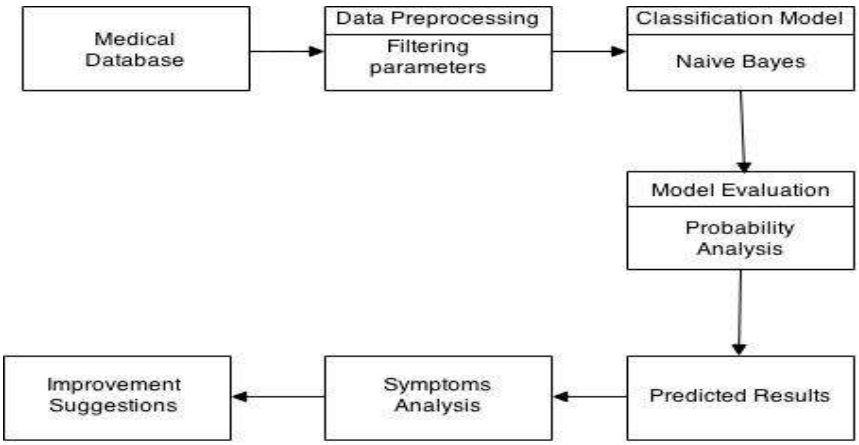

Fig. 1: Architecture diagram for proposed system

3.1.1 Medical Database: The Proposed System uses a large database acquired from a local pathology lab, which contains numerical values for the required factor.

3.1.2 Filtering Parameters: Stored data is pre-processed, empty spaces are filled and values are checked for upper and lower limits using data Warehousing Techniques.

3.1.3 Naïve Bayes Algorithm: Naïve bayes is used to calculate probabilities for each factor.

3.1.4 Probability analysis: Calculated probabilities from the database are analysed and compared with the user's values for predicting results.

3.1.5 Predicted Results: Probabilities for each factor are combined together and total probability of risk is calculated using the Naïve Bayes formulae.

3.1.6 Symptom Analysis: The factors that increase the risk of a disease are displayed and treatments to control each symptom are given.

3.1.7 Improvement suggestions: Data is provided to users to improve health and reduce risk of analysed diseases.

IV. FACTORS

Prediction of Diabetes is based on a number of numerical factor and statistics. The factors used in the proposed system are tabulated below.

Table 2: Factors for prediction

SR .no	Parameter	Description	Allowed values (Type)
1	Age	Age of the Subject	Discrete Integer values
2.	Gender	Gender of Subject	Male or female
3.	Body Mass Index(BMI)	Weight in kg/(Height in m)^2	Discrete Integer values
4.	Genetics	Family History of subject	Yes or No
5.	Blood Pressure	Diastolic blood pressure (mm hg)	Discrete integer values
6.	Pregnant	Number of times pregnant	Discrete integer values
7.	Plasma glucose Concentration	Blood sugar of the subject	Discrete integer values
8.	Smoking	Whether subject is smoking or not	Yes or No
9.	Drinking	Whether subject drinks occasionally	Yes or No

V. ALGORITHM

The proposed system uses Naïve Bayes Algorithm to analyze data and predict risks of diseases.

Input: Numerical or fixed value inputs for different factors.

VI. PROCESS

6.1 Run probabilities for each factor in the database within a fixed range. Probability of Diabetes risk with each Factor is calculated separately using:

P(Factor1|yes) = Number of people with Factor1 and Diabetes / Total Number of People

6.2 Find the range in which each input factor lies. Most numerical factors have fixed values of normal, high or low. Extract probabilities for particular range. eg. Cholesterol=250 means high Cholesterol

6.3 Mathematically the probability model for a classifier is a conditional model over a dependent class variable with a small number of outcomes or classes, conditional on several Factors F1 to Fn.

Using Bayes' theorem we rewrite the equation as:

P(Diabetes risk) = P(Factor1|Yes) * P(Factor2|yes) * ….

6.4 The probability is compared with a minimum value and analysed for risk assesment. High risk factors are analysed and used for personalised solutions.

Output: Display Risk probability and solutions for the risks.

VII. PROTOTYPE SCREENSHOTS

Given below is the prediction system for the prototype

Prediction Page

Fig. 2: Refers to the page where the researcher has to fill the details for predicting the disease

Result Page

Fig. 3: Refers to Results of the Predicted Disease.

VII. CONCLUSION

This system shows accurate results for prediction using Naïve Bayes. The Algorithm uses probabilities calculated from a large number of patients and hence has a low error rate which can be improvised by using a larger database. This system can be used extensively by patients to check their risk for certain diseases. In the future, this project can be expanded for many other diseases and their various symptoms.

REFERENCES

[1] G. Parthiban,A.Rajesh,S.K.Srivatsa,' Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method', International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[2] Ravneet Jyot Singh, Williamjeet Singh,' Data Mining in Healthcare for Diabetes Mellitus' International Journal of Science and Research (IJSR), Volume 3 Issue 7, July 2014

[3]. Abid Sarwar Vinod Sharma,'Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2', Special Issue of International Journal of Computer Applications (0975 – 8887) on Issues and Challenges in Networking, Intelligence and Computing Technologies – ICNICT 2012, November 2012

[4] Srideivanai Nagarajan, R.M.Chandrasekaran, and P.Ramasubramanian, 'Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes ', International journal of current research and academic review, ISSN: 2347-3215 Volume 2 Number 10 (October-2014) pp. 91-98

[5] Abdullah Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, 'Applications of data mining: Diabetes health care in young and old patients', Jounal of King Saud University- Computer and Information Science (2013).

[6]. Gaganjot Kaur Amit Chhabra, 'Improved J48 Classification Algorithm for the Prediction of Diabetes', International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014

[7]. Kavitha K , Sarojamma R M 'Monitoring of Diabetes with Data Mining via CART Method',International Journal of Emerging Technology and Advanced Engineering. Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 11, November 2012).

[8] http://www3.cs.stonybrook.edu

[9] Rupali Bhardwaj , Sonia Vatta,' Implementation of ID3 Algorithm'.' International Journal of Advanced Research in Computer Science and Software Engineering', Volume 3, Issue 6, June 2013

[10] Monali Dey, Siddharth Swarup Rautaray, 'Study and Analysis of Data mining Algorithms for Healthcare Decision Support System', (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 470-477.

[11] msdn.microsoft.com/en-IN/library/ms175595.aspx

[12] http://www.ncbi.nlm.nih.gov/