International Journal of Computer Applications Technology and Research
Volume 7–Issue 07, 255-258, 2018, ISSN:-2319–8656

Prediction of Heart Disease in Diabetic patients using Naive Bayes Classification Technique

Charu V.Verma
Research scholar (CSE)
Dr. C.V. Raman University
Bilaspur, India

Dr. S. M. Ghosh
Professor
Dr. C.V. Raman University
Bilaspur, India

Abstract: The objective of our paper is to predict the risk of heart disease in diabetic patients. In this research paper we are applying Naive Bayes data mining classification technique which is a probabilistic classifier based on Bayes theorem with strong (naive) independence assumptions between the features. Data mining techniques have been widely used in health care systems for prediction of various diseases with accuracy. Health care industry contains large amount of data and hidden information. Effective decisions are made with this hidden information by applying data mining techniques. These techniques are used to discover hidden patterns and relationships from the datasets. The major challenge facing the healthcare industry is the provision for quality services at affordable costs. A quality service implies diagnosing patients correctly and treating them effectively. In this proposed system certain attributes are consider in diabetic patients to predict the risk of heart disease

Keywords: Heart Disease, Diabetes, Data Mining, KDD, Naïve Bayes.

1. INTRODUCTION

The development of Information Technology has generated large amount of databases and huge data in various areas. The area includes health sector, financial sector, weather forecasting, education, manufacturing, fraud detection, bio information etc. The research in databases and information technology has given rise to an approach to store and manipulate this useful data for further decision making. Data mining is the process of discovering patterns in large data bases involving methods at the intersection of machine learning, statistics, and database systems [1]. Popularly data mining referred as knowledge discovery from the data. It is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams. The knowledge discovery is an interactive process, consisting by developing an understanding of the application domain, selecting and creating a data set, preprocessing, data transformation.

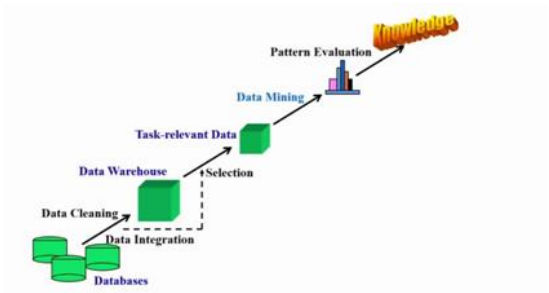


Figure 1: KDD knowledge discovery process

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods. Data mining plays a vital role in field of health sector since here huge amounts of data is generated which are too complex and voluminous to be processed and analyzed by traditional methods and difficult to handle manually. 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. In our busy schedule, most of the people work like a machine in order to live a deluxe and comfortable life in future and to earn more money, but during this type of situation people forget about their health and even don't take any proper rest. Because of this they affected from various type of diseases at a very early age, which cause our health as Diabetes, Heart Disease, Cancer, Eyes, Kidney failure and many more.

Diabetes Mellitus (DM) is commonly referred as Diabetes; it is the condition in which the body does not properly process food for use as energy. Most of the food we eat is turned into glucose or sugar for energy. The pancreas, an organ makes a hormone called insulin to help glucose get into the cells of our bodies. When a body is affected with diabetes, it couldn't make enough insulin or couldn't use its own insulin. This causes sugar to build up into blood. Several pathogenic processes are involved in the development of diabetes. These range from autoimmune destruction of the β-cells of the pancreas with consequent insulin deficiency to abnormalities that result in resistance to insulin action. Diabetes is a life threatening disease in rural and urban, then developed and under developed countries. The common symptoms for the diabetic patients are frequent urination, increased thirst, weight loss, slow-healing in wound, giddiness, increased hunger etc. Diabetes can cause serious health complications including heart disease, blindness, kidney failure and low-extremity amputations.

Types of Diabetes

Type 1Diabetes is called insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes. Autoimmune, genetic, and environmental factors are involved in the development of this type of diabetes. Type1 mostly occurs in young people who

are below 30 years. This type can affect children or adults, but majority of these diabetes cases were in children. In persons with type 1 diabetes, the beta cells of the pancreas, which are responsible for insulin production, are destroyed due to autoimmune system.

Type 2 Diabetes is called non-insulin-dependent diabetes mellitus (NIDDM) or adult-onset diabetes. In the type 2 diabetes, the pancreas usually produces some insulin the amount produced is not enough for the body's needs, or the body's cells are resistant to it. Risk factors for Type 2 diabetes includes older age, obesity, family history of diabetes, prior history of gestational diabetes, impaired glucose tolerance, physical inactivity, and race/ethnicity.

Gestational Diabetes is the third main form and occurs when pregnant women without a previous history of diabetes develop a high blood glucose level. The majority of gestational diabetes patients can control their diabetes with exercise and diet. In such cases between 10%-20% of them will need to take some kind of blood-glucose-controlling medications. In few cases this gestational diabetes may lead to type 2 diabetes in future. It affects on 4% of all pregnant women [2].

Heart disease is the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques in the diagnosis of heart disease. Heart diseases are the number 1 cause of death globally: more people die annually from heart disease than from any other cause. An estimated 17.7 million people died from heart disease in 2015, representing 31% of all global deaths. Of these deaths, an estimated 7.4 million were due to coronary heart disease and 6.7 million were due to stroke. Over three quarters of heart disease deaths take place in low- and middle-income countries. Out of the 17 million premature deaths (under the age of 70) due to non communicable diseases in 2015, 82% are in low- and middle-income countries, and 37% are caused by this disease. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management using counseling and medicines, as appropriate [3]. Now a day's heart disease is a major health problem and cause of death all over the world. Heart is a very valuable part of our body, and plays a very important role in our life. Our whole life depends on efficient working of heart. The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These "intermediate risks factors" can be measured in primary care facilities and indicate an increased risk of developing a heart attack, stroke, heart failure and other complications.

Heart disease is caused due to narrowing or blockage of coronary arteries. This is caused by deposition of fat on inner walls of arteries and also due to build up cholesterol. There are some of major heart disease factors which include Diabetes, high blood pressure, high cholesterol, obesity, family history, smoking, eating habits, alcohol that affects our whole body.

2. METHODOLOGY APPLIED

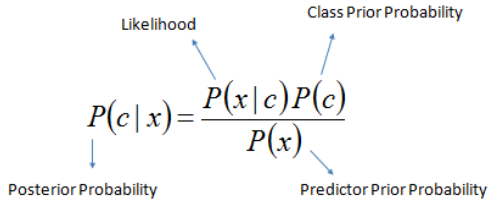
DATA MINING in health care has become increasingly popular because it can improved our patient care by early detecting of disease supports helping care providers for treatment programs and reduces the cost of health care. [4]

Data Mining is major anxious with the study of data and Data Mining tools and techniques are used for discovery patterns from the data set. The most important aim of Data Mining is to find patterns mechanically with least user input and efforts. Data Mining is an influential tool able of usage decision building and for forecasting expectations trends of market. Data Mining tools and techniques can be effectively functional in different fields in different forms. Many Organizations now begin using Data Mining as a tool, to contract with the aggressive surroundings for data analysis. By using Mining tools and techniques, different fields of business get advantage by simply assess various trends and pattern of market and to make rapid and efficient market trend analysis. Data mining is very helpful tool for the diagnosis of diseases.

Classification Technique

Classification derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible. Mainly classification is used to classify every item in a set of data into one of predefined set of classes or groups.

Naïve bayes is one of technique of classification used for prediction of data. Naive Bayes classifiers is a probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for large datasets such as in the field of medical science for diagnosing heart patients. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods [8]. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



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

2.1 Equations:

• P(c|x) is the posterior probability of class (target) given predictor (attribute).

• P(c) is the prior probability of class.

• P(x|c) is the likelihood which is the probability of predictor given class.

• P(x) is the prior probability of predictor Where C and X are two events. Where C and X are two events. Such Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods [4].

2.2 Processing Data set

The patient data set is compiled from UCI data repositories as combined data from Statlog data set and Cleveland Clinic Foundation for heart patients. Here we are taken 14 attributes with nominal values from the database that are considered for the required prediction they are age, sex, chest pain(cp), Blood Pressure(trestbsp), diabetes(fbs)(its value is always 1), ECG(restecg), Heart Rate(thalach), exang, oldpeak, slope, thal, blood Cholesterol(chol) and num (heart disease diagnosis). Where if the num value is present then there is presence of heart disease and if the num value is absent then no heart disease.

2.3 Tool used

Waikato Environment for Knowledge Analysis (WEKA) has been used for prediction due to its proficiency in discovering, analysis and predicting patterns. It was developed at the University of Waikato in New Zealand and easiest way to use is through a graphical user interface called Explorer. Weka is a collection of machine learning algorithms for data mining tasks, written in Java and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization

2.4 10 fold cross validation

Cross-validation, a standard evaluation technique, is a systematic way of running repeated percentage splits. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once[7]. Divide a dataset into 10 pieces (“folds”), then hold out each piece in turn for testing and train on the remaining 9 together. This gives 10 evaluation results, which are averaged. In “stratified” cross-validation, when doing the initial division we ensure that each fold contains approximately the correct proportion of the class values. Having done 10-fold cross-validation and computed the evaluation results, Weka invokes the learning algorithm a final (11th) time on the entire dataset to obtain the model that it prints out [6].

We will simply define and calculate the accuracy, sensitivity, and specificity from the confusion matrix.

True positive (TP) = the number of cases correctly identified as patient

False positive (FP) = the number of cases incorrectly identified as patient

True negative (TN) = the number of cases correctly identified as healthy

False negative (FN) = the number of cases incorrectly identified as healthy

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy= (TP+TN)/TP+TN+FP+FN

Sensitivity: The sensitivity of a test is its ability to determine the patient cases correctly. To estimate it, we should calculate the proportion of true positive in patient cases. Mathematically, this can be stated as:

Sensitivity=TP/TP+FN

Specificity: The specificity of a test is its ability to determine the healthy cases correctly. To estimate it, we should calculate the proportion of true negative in healthy cases. Mathematically, this can be stated as:

Specificity=TN/TN+FP

3. EXPERIMENTAL RESULT

The result in our experiment is given of heart disease in diabetic patients. In our experiment we had proposed Naive bayes classification technique as experimental result shown in the figure 2. Accuracy, Sensitivity, Specificity has been calculated from the both confusion matrix of both the performed experiments. For heart disease model accuracy is 89.41%, Sensitivity is 46.05%, Specificity is 55.55%

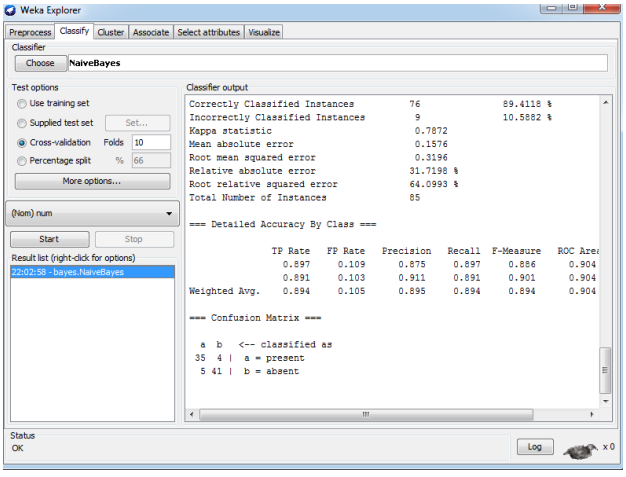


Figure 2 Result window of heart disease data set having Diabetes

Table 1. Result of detailed accuracy class of diabetes patient having heart disease

	Tp	Fp	Pre	Recall	Fm	ROC

Heart disease with diabetes	0.894	0.105	0.895	0.894	0.894	0.904

4. CONCLUSION

In this research paper application of data mining is used to analyze the clinical dataset to detect diseases and diagnosis based on the data and the attributes provided. In the proposed works Naive Bayes classification technique helps to predict heart disease in diabetic patient. From the system we get confusion matrix from which we can predict accuracy of the applied Naïve Bayes algorithm. The result shows that accuracy of our applied algorithm is 89.41% in the prediction of risk of heart disease in diabetic patient. As a future work, further data analysis has been planned to perform other data mining algorithms to improve the classification accuracy.

5. REFERENCES

[1] https://en.wikipedia.org/wiki/Data_mining

[2] http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] https://www.researchgate.net/publication/312188365 ”Heart Attack prediction using Data mining technique”

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

[5] https://www.futurelearn.com/courses/data-mining-with-weka/0/steps/25384.

[6] https://www.openml.org/a/estimation-procedures/1

[7] http://www.saedsayad.com/naive_bayesian.htm

[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4614595