

DISEASE PREDICTION USING DATA MINING

Mr. Tanaji* | Dr. Balaji Koturu**

*Research. Scholar, Department of Computer Science and Information, Himalayan University, Itanagar, Arunachal Pradesh, India.

**Research Supervisor, Department of Computer Science and Information, Himalayan University, Itanagar, Arunachal Pradesh, India.

DOI: <http://doi.org/10.47211/tba.2023.v08i01.004>**ABSTRACT:**

Data mining is described as sifting via very huge amount of records for beneficial information. Some of the vital and famous records mining techniques are classification, clustering, prediction. Data mining methods are used for range of applications. In health care industry, data mining plays an necessary position for predicting diseases. For detecting a sickness number of exams should be required from the patient. But the use of data mining method the wide variety of check be reduced. This reduced check plays an important position in time and performance. This method has an benefits and disadvantages. This paper analyzes how data mining techniques are used for predicting one of kind sorts of diseases. This paper reviewed the research papers which mainly targeted on predicting heart disease, Diabetes, breast cancer.

Key Words: Data mining, disease, heart disease, breast cancer, diabetes.

ABOUT AUTHORS:

Author Mr. Tanaji is Ph.D. Scholar at Himalayan University, Itanagar, Arunachal Pradesh, India.



Author Dr. Balaji Koturu Research Supervisor, Himalayan University, Itanagar, Arunachal Pradesh, India. He has published articles in national and International Journal.

INTRODUCTION:

Data Mining is the manner of extracting hidden know-how from massive volumes of raw data. The understanding should be new, no longer obvious, and one have to be in a position to use it. Data mining has been defined as “the nontrivial extraction of earlier unknown, implicit and doubtlessly useful records from data. It is “the science of extracting useful data from massive databases”. It is one of the duties in the method of understanding discovery from the database. Data Mining is used to find out knowledge out of data and presenting it in a shape that is without problems recognize to humans. It is a manner to look at massive amounts of statistics robotically collected. Data mining is most useful in an exploratory analysis because of nontrivial statistics in massive volumes of data. It is a cooperative effort of human beings and computers. Data mining is the process of selecting, discovering and modeling huge amounts of data. This process has become an increasingly insidious activity in all areas of medical science research. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization; and includes classification and regression techniques. Best effects are finished with the aid of balancing the knowledge of human experts in describing issues and desires with the search skills of computers. There are two essential desires of statistics mining have a tendency to be prediction and description. Prediction includes some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand description focuses on discovering patterns describing the facts that can be interpreted by way of humans. There are unique sorts of illnesses estimated in facts mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes etc

HEART DISEASE

There are various factors that contribute to the development of heart disease, such as age, gender, family history, smoking, high blood pressure, high cholesterol, diabetes, obesity, and sedentary lifestyle. Predicting heart disease involves analyzing these risk factors and identifying patterns and relationships that can help in early detection.

Data mining techniques, such as classification algorithms, can be used to predict the likelihood of an individual developing heart disease. These algorithms are trained on historical data that includes information about patients' risk factors and their heart disease status. By learning from this data, the algorithms can identify patterns and create prediction models. Feature selection plays a crucial role in heart disease prediction. Relevant features, such as age, gender, blood pressure, cholesterol levels, and smoking status, are selected to build accurate prediction models. These models can then be used to assess an individual's risk of developing heart disease based on their specific combination of risk factors. Machine learning algorithms, like decision trees, logistic regression, and support vector machines, are commonly used in heart disease prediction. These algorithms analyze the selected features and create decision rules to classify individuals as either having heart disease or being at low risk. Additionally, data mining techniques can be combined with other technologies, such as genetic testing and imaging data, to enhance the accuracy of heart disease prediction. Genetic information can provide insights into an individual's predisposition to heart disease, while imaging data can help identify structural abnormalities or blockages in the arteries. It's important to note that while data mining techniques can assist in predicting heart disease, they should always be used in conjunction with medical expertise. The predictions made by these models should be validated and interpreted by healthcare professionals to ensure accurate diagnosis and appropriate treatment. In conclusion, predicting heart disease involves analyzing risk factors, selecting relevant features, and using data mining techniques to create prediction models. These models can help identify individuals at high risk of developing heart disease, enabling early intervention and preventive measures. Medical information mining has excess potential for exploring the hidden patterns in the information sets of the medical domain. These patterns can be utilized for clinical prognosis for widely dispensed in raw clinical statistics which is heterogeneous in nature and voluminous. This information ought to be accessed in an equipped form. This gathered information can be integrated to form a medical institution facts system. Data mining technological know-how affords a user-oriented approach to novel and hidden patterns in the data. From the evaluation of World Health Organization, they estimated 12 million deaths manifest worldwide, each and every year due to the heart diseases. Half the deaths happen in United States and different developed countries due to cardio vascular diseases. On the above discussion, it is considered as the primary motive at the back of deaths in adults. Heart disorder kills one man or woman each and every 34 seconds in the United States. The following paper reviewed about predicting of coronary heart disorder the use of records mining technique. Heart disease is the leading cause of death in the U.S. At any point in your life, either you or one of your loved ones will be forced to make decisions about some aspect of heart disease. Knowing about the structure and functioning of the heart, in particular how angina and heart attacks work, will enable to make informed decisions about your health. Heart disease can strike suddenly and need to make decisions quickly.

Decision tree: is one of the popular and important classifiers which is easy and simple to implement. It doesn't have domain knowledge or parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results attained from Decision Tree are easier to interpret and read.

Naive Bayes: is a statistical classifier which assigns no dependency between attributes. To determine the class the posterior probability should be maximized. The advantages are one can work with the naïve bayes model without using any Bayesian methods. Here Naïve Bayes Classifiers performs well.

Several factors contribute to this damage. They include:

1. Smoking, including secondhand smoke
2. High amounts of certain fats and cholesterol in the blood
3. High blood pressure
4. High amounts of sugar in the blood due to insulin resistance or diabetes
5. Blood vessel inflammation

BREAST CANCER:

Breast cancers has turn out to be a frequent cancer in women. Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow up into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Some common signs of breast cancer include a lump or thickening in the breast or underarm area, changes in breast size or shape, nipple discharge or inversion, and skin changes such as redness or dimpling. However, it's important to note that these symptoms can also be caused by non-cancerous conditions. The disease occurs almost entirely in women, but men can get it, too. For instance, it affects one in every seven ladies in the United State. The mammography is the typical method for breast cancers diagnosis. However, the radiologists exhibit substantial variability in how they interpret a mammogram. Moreover, Elmore indicated that 90% of radiologists diagnosed fewer than 3% of cancers and 10% identified about 25% of the cases. The satisfactory needle aspiration cytology is any other strategy adopted for the prognosis of breast most cancers with extra precise prediction accuracy. However, the common right identification rate is around 90%. Generally, the cause of all the associated research is identical to distinguish between patients with breast cancer in the malignant group and patients without breast most cancers in the benign group. There are three predictive center of attention of most cancers prognosis: 1) prediction of most cancers susceptibility (risk assessment), 2) prediction of most cancers recurrence and 3) prediction of most cancers survivability.

DIABETES:

Diabetes is a chronic condition that affects how your body regulates blood sugar (glucose) levels. There are two main types of diabetes: type 1 and type 2. Type 1 diabetes is an autoimmune disease where the immune system mistakenly attacks and destroys the insulin-producing cells in the pancreas. Insulin is a hormone that helps regulate blood sugar levels. People with type 1 diabetes require daily insulin injections or the use of an insulin pump to manage their blood sugar levels. On the other hand, type 2 diabetes is more common and often develops later in life. It occurs when the body becomes resistant to the effects of insulin or doesn't produce enough insulin to maintain normal blood sugar levels. Type 2 diabetes can often be managed through lifestyle changes like a healthy diet, regular exercise, and sometimes medication.

Both types of diabetes can have serious health complications if not properly managed. High blood sugar levels over time can lead to problems with the heart, kidneys, eyes, nerves, and blood vessels. That's why it's important for individuals with diabetes to work closely with healthcare professionals to monitor their blood sugar levels, take medications as prescribed, and make necessary lifestyle adjustments. Diabetes mellitus, or simply diabetes, is a chronic disease that occurs when the pancreas is no longer able to make insulin, or when the body cannot make good use of the insulin it produces. Insulin is a hormone made by the pancreas, which acts like a key to let glucose from the food we eat pass from the blood stream into the cells in the body to produce energy. All carbohydrate foods are broken down into glucose in the blood. Insulin helps glucose get into the cells. Not being able to produce insulin or use it effectively leads to raised glucose levels in the blood (known as hyperglycemia). Over the long-term high glucose levels are associated with damage to the body and failure of various organs and tissues. Insulin is one of the most necessary hormones in the body. It aids the physique in converting sugar, starches and other meals items into the power needed for each day life. However, if the body does no longer produce or top use insulin, the redundant amount of sugar will be driven out by using urination. This sickness is referred to diabetes. The purpose of diabetes is a mystery, even though weight problems and lack of workout appear to possibly play full-size roles. Based on the American Diabetes Association in November 2007, 20.8 million young people and adults in the United States (i.e., about 7% of the population) have been diagnosed with diabetes.

DATA MINING TECHNIQUES**Naive Bayes**

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is simple to build, with no difficult iterative parameter estimation which makes it particularly useful for very large datasets.

Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of manipulating the posterior probability.

It's a classification algorithm that's based on Bayes' theorem. The "naive" part comes from the assumption that the predictors are independent of each other. This simplifies the model-building process and makes it especially useful for handling large datasets. When we talk about the classifier, we have a class variable (let's say "stolen?") that represents whether a car is stolen or not, given certain conditions. These conditions are represented by the features or parameters, denoted as x_1, x_2, x_3 , and so on. These features could be things like the color, type, and origin of the car. To calculate the probability of the class variable given the features ($P(y|x_1, x_2, x_3, \dots, x_n)$), we can use the equation. By substituting the values for each feature from the dataset and expanding the equation using the chain rule, we can calculate the probability. In the equation, we have $P(x_i|y)$, which represents the probability of each feature given the class variable. We multiply these probabilities together for all the features, and also multiply it with the prior probability of the class variable ($P(y)$). The denominator, which represents the probability of the features occurring together, remains the same for all entries in the dataset.

To simplify the equation, we can remove the denominator and use proportionality instead. So, the probability of the class variable given the features is proportional to the prior probability of the class variable multiplied by the probabilities of each feature given the class variable. In some cases, the classification can indeed have more than two outcomes. It depends on the problem and the dataset. The Naive Bayesian classifier can handle multivariate classification as well.

CLASSIFICATION USING CLUSTERING

Clustering is the process of grouping identical elements. This method may additionally be used as a preprocessing step before feeding the data to the classifying model. The attribute values want to be normalized earlier than clustering to keep away from high cost attributes dominating the low fee attributes. Further, classification is performed based on clustering. Experiments had been carried out with Weka 3.6.0 device. Data set of 909 records with thirteen attributes. All attributes are made specific and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering is not performing well when compared to other two methods.

In the survey of Naive bayes have been used to predict attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. The clinical dataset is having been collected from one of the leading diabetic research institutes in Chennai. The records of 500 patients are taken. The data is analyzed and implemented in WEKA ("Waikato Environment for Knowledge Analysis") tool. Data mining finds out the valuable information hidden in huge volumes of data. Weka tool is a collection of machine learning algorithms for data mining techniques, written in Java. It consists of data pre-processing, classification, regression, association rules, clustering and visualization tools. We have used Naïve bayes method to perform the mining and classification process. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process.

DECISION TREE:

Decision tree models are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions. The prediction could be to predict categorical values (classification trees) when instances are to be placed in categories or classes. Decision tree is a classifier in the structure of a tree structure where each node is either a leaf node, indicating the value of the target attribute or class of the examples, or a decision node, specifying some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance.

CONCLUSION

In this survey paper the trouble of summarizing the special algorithm of data mining are used in the subject of clinical prediction are discussed. The main focal point is on using exclusive algorithm and aggregate of a number of goals attributes for unique types of sickness prediction using data mining. In this study two different data

mining classification techniques was used for the prediction of various diseases and their performance was compared in order to evaluate the best classifier. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for medical applications. First we talk about the coronary heart disease prediction, in that machine studying algorithms namely naive bayes, Decision List. Of these the classification accuracy of the naive bayes algorithm is higher when in contrast to different algorithm. Here the accuracy is in contrast to the three classifiers particularly Decision Tree, Naïve bayes and classification by using clustering. To locate predictive policies in medical statistics set the three necessary steps are generated in this algorithm. The coronary heart disorder is recognized for diabetic sufferers the use of naïve bayes method. Of these the creator concluded that naïve bayes classify 74% of input situations correctly.

REFERENCES:

1. Nazari Nezhad, S., Zahedi, M. H., & Farahani, E. (2022). Detecting diseases in medical prescriptions using data mining methods. *BioData mining*, 15(1), 29. <https://doi.org/10.1186/s13040-022-00314-w>.
2. Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International journal of nanomedicine*, 13(T-NANO 2014 Abstracts), 121–124. <https://doi.org/10.2147/IJN.S124998>.
3. Maheswari, S., & Pitchai, R. (2019). Heart Disease Prediction System Using Decision Tree and Naive Bayes Algorithm. *Current medical imaging reviews*, 15(8), 712–717. <https://doi.org/10.2174/1573405614666180322141259>