A Review on Prediction and Analysis of Multiple Diseases in Healthcare using Data Mining

Miss Vaishali G. Mohature [1], Prof.J.M.Patil [2]

[1,2]*Department of Computer Science and Engineering SSGMCE, Shegaon, India*

Abstract: Now a days, there are many applications are used for searching results on web. In this system we are predicting multiple diseases by applying data mining technique. Data mining is the process of discovering interesting pattern and large amount of data. The main aim of this project is to build a basic decision support system which can determine and exact previously unseen patterns, relation and concepts related with multiple diseases. Data mining used of large data set. Data set used is Pima Indian diabetes dataset. Classification represents a data mining technique that requires collecting various of information and data for their attributes in order to be analyzed. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like SVM, Naive Bayes, Decision Tree, Decision Table etc. In this research we review on diabetes prediction. And various dataset used for diabetes prediction.

Keywords- Prediction, Data mining, Various Techniques, Classification, and Dataset.

I.INTRODUCTION

Data mining is a relatively new concept used for retrieving information from a large set of data. Mining means using available data and processing it in such a way that it is useful for decision-making. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining thus has evolved based on human needs which can help humans in identifying relationship patterns and forecasts based on pre-set rules and stipulations built into the program (Eapen, 2004).

Data mining is subfield in the subject software engineering.. There are currently a lot of health institutions that has been developed such as hospitals and medical centers which are crucial to maintain and improve the health of the community around us. The classification and prediction techniques can be improved by concatenating association rules with it to find out the frequently used elements. Data Mining is used for many attributes. In computer science for the past few years, the health industry has been growing significantly that leads to insurmountable piles of data to be calculated. The present study focus on developing a diabetic prediction system based on data mining methods. Nowadays, data mining is the most important technique in health system. The large amount of data is generated in healthcare. The main aim of at analyzing the various data mining techniques in the recent years. Using data mining techniques the peoples to predict diabetes has gain major popularity. Data mining is the process of discovering correlations, patterns through large amount of data stored in repositories, database and warehouse. Diabetes is the fast growing disease among the youngsters. In diabetes a person generally suffering from high blood sugar. There are different type of disease predicted in the data mining i.e. sugar breast cancer lung cancer ,thyroid etc. the main aim of prediction of diabetes a candidate is suffering at a particular age. The proposed system is designed based on the concept of pattern matching algo. The algorithm and FP growth are used to frequent item sets from data base in the previous research works the implementation and the accuracy testing of algorithm such as decision tree Naive Bayes carried out in which Neural network[1].

Data mining is the process of discovering correlations, patterns or relationships through large amount of data stored in repositories, databases and data warehouse. Many techniques or solutions for data mining and knowledge discovery in databases are very widely provided for classification, association, clustering and regression, search, optimization. Diabetes is the fast growing disease among the youngsters. In diabetes a person generally suffering from high blood sugar. Intensify thirst, Intensify hunger and frequent urination are some of the symptoms caused due to high blood sugar. There are different type of disease predicted in the data mining i.e. sugar, breast cancer, lung cancer, thyroid etc. Data mining is a relatively new concept used for retrieving information from a large set of data.

Mining means using available data and processing it in such a way that it is useful for decision-making. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining thus has evolved based on human needs which can help humans in identifying relationship patterns and forecasts based on pre-set rules and stipulations built into the program (Eapen, 2004).

Data mining is a relatively new concept used for retrieving information from a large set of data. Mining means using available data and processing it in such a way that it is useful for decision-making. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining thus has evolved based on human needs which can help humans in identifying relationship patterns and forecasts based on pre-set rules and stipulations built into the program (Eapen, 2004).

DIABETES

Diabetes is a chronic disease that occurs when the human pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces, which leads to an increase in blood glucose levels. There are generally three types of diabetes,
TYPE 1- In this type of diabetes, the pancreatic cells that produce insulin have been destroyed by the defense system of the body. This type can be caused regardless of obesity. Type 1 diabetes can occur in childhood age. TYPE 2-In this case the various organs of the body become insulin resistant, and this increases the demand for insulin or fails to produce insulin. Type 2 generally occurs in the middle age groups. GESTATIONAL DIABETES-It is a type of diabetes that tends to occur in pregnant women due to the high sugar levels as the pancreas don't produce sufficient amount of insulin.

Side effect of diabetes,

In addition to the symptoms, diabetes can cause long term damage to our body diabetes affects our blood vessels and nervous and heart attack and stroke therefore can affect any part of the body. The risk is greater for people with diabetes, who have progressed cholesterol and blood pressure levels. If the family has diabetic history also increases heart problems. To reduce the risk and pick up any problems early: Have the blood pressure checked at least every six months, but more often if person have high blood pressure or are taking medication to lower this. Have the test of HbA1c checked at least every year it may need to be checked three to six monthly. Have the cholesterol checked at least yearly. Further pathology tests such as an electrocardiogram (ECG) or exercise stress test may also be recommended by doctor. Heart disease and blood vessel are common problems for many people who don't have their diabetes under control. Blood vessel damage and nerve damage may also cause foot problems that, in rare cases, can lead to amputations. People having diabetes are ten times occurred cause to damage the whole body.

II.LITRATURE REVIEW

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.

II. LITRATURE SURVEY

Various data mining techniques have been used for study and analysis of various diseases like hepatitis, cancer and diabetes is also one of them data mining plays an important role in diabetes prediction in healthcare. The health prediction system will rely on its implementation of data mining, which is referred to as mining knowledge and information from a large amount of data sets. The medical industry is just one of many fields in society that collects vast amount of information that can utilized helpfully by data mining. Data mining can improve the medical industry with eliminating current health disparities by easily providing answers to complex medical cases to solve and eliminate any time consumptions created from making a clinical decision. The main aim of this paper is to find out best classifier from different classification algorithm that can be used to predict disease on applying data set of the patients [3]. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithm works better in diagnosing different diseases. Data Mining and Machine learning algorithms gain its strength due to the

capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study. This research work focuses on pregnant women suffering from diabetes. In this work, Naive Bayes, SVM, and Decision Tree machine learning classification algorithms are used and evaluated on the PIDD dataset to find the prediction of diabetes in a patient. Experimental performances of all the three algorithms are compared on various measures and achieved good accuracy [4].

Orabi et al. in [5] designed a system for diabetes prediction, whose main aim is the prediction of diabetes a candidate is suffering at a particular age. The proposed system is designed based on the concept of machine learning, by applying decision tree. Obtained results were satisfactory as the designed system works well in predicting the diabetes incidents at a particular age, with higher accuracy using Decision tree [7].

Mohamed EL Kourdi et al. [3] have proposed this system in which Naive Bayes (NB) which is a factual machine learning algorithm is utilized to order Arabic web documents. This system utilizes K-Nearest Neighbor for estimating values that will maintain a strategic distance from financial distress and bankruptcy. In the current review k-Nearest Neighbor characterization technique, have been examined for economic estimating. Lately, after the situation of worldwide financial emergency, the quantity of bankrupt organizations has risen. Since organizations' financial distress is the principal phase of bankruptcy, utilizing financial proportions for anticipating financial distress have pulled in a lot of consideration of the scholastics and economic and financial institutions.

Kevin Beyer et al. [6] have proposed this system that tries to explain what happens when dimensionality increases. While dimensionality builds the separation between the nearest and the most distant point gets to be distinctly irrelevant and in this manner the execution is influenced. This may prompt to wrong forecast. Additionally, increment in measurement ought to be dismissed however much as could be expected. In example acknowledgment, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric strategy utilized for order and relapse. The K-NN algorithm is among the easiest of all machine learning algorithms. Both for order and relapse, it can be valuable to relegate weight to the commitments of the neighbors, so that the closer neighbors contribute more to the normal than the more far off ones. Data mining can improve the medical industry with eliminating current health disparities by easily providing answers to complex medical cases to solve and eliminate any time consumptions created from making a clinical decision [2]. Based on the finding from the literature review, most health prediction system contains around more than one data mining algorithms to predict the diseases with all of it indicating Naïve Bayes algorithm to be the data mining algorithm that produces the best and most accurate result out of all the other data mining algorithm. Generally, the data mining algorithm will be considered and selected depending on the size of the dataset to be tested on its prediction accuracy.

Abhishek Taneja in his paper mentions that the algorithms with search limitations and constraints are also introduced to reduce the count of association and also for validation purpose [8].

III.DATA MING TECHNIQUES

1. Decision tree

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes.

2. Naïve base classifier

Naïve Bayes is a classification technique with a notion which defines all feature are independent and unrelated to each other. Naïve Bayes can be a powerful predictor. This technique very useful for large datasets. Naive Bayes is a machine learning classifier which employs the Bayes theorem. Naive Bayes is known to beat even profoundly advanced grouping techniques. Bayes theorem provides a simple method of calculating posterior probability P(c in x) from P(c), P(x) and P (x in c).

Look at the equations below:

P (C in X) = P (X in C) P(C) /P(X)
P (C in x) is the posterior probability of class (C, target) given predictor (x, attributes).
P (C) is the prior probability of class.

P (X in C) is the likelihood which is the probability of predictor given class.
P (X) is the prior probability of predictor.

3. Support Vector Machine

SVM is one of the standard set of supervised machine learning. model employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes [7]. For better generalization hyperplane should not lies closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors. The Accuracy of the experiment is evaluated using WEKA interface. The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyperplane which is defined by $wT\ x + b = -1$ and the hyperplane defined by $wT\ x + b = 1$ this distance is equal to 2 _w_. This means we want to solve max 2 _w_.

IV. MOTIVATION AND SCPOPE

In healthcare organization and medical diagnosis data mining and machine learning technique can be implemented to treat some hazardous disease. Besides all the other disease heart, hepatitis, miasmas are responsible to take more lives than any other disease such as cancer and HIV. Thus there is a need to predict theses disease in real time to prevent future hazards and to avoid future issues and problems. A tabular dataset is constructed by encapsulating all of the medical digital data and is used for future experiments of prediction analysis data mining is one of the most important techniques in healthcare industry. Then the proposed system of prediction disease can be developed many directions which have vast scope of improvement the system. Also Increase the accuracy of algorithm. It is also used multiple disease prediction. Working on some more attributes so to tackle diabetes.
In future the system helps to everyone. In medical diagnosis, data mining has been widely used for predicting diseases through diagnosis. Various data mining models have to be considered and compared in order to create an intelligent health prediction system.

V.CONCLUSION

In this paper the various data mining techniques are used to predict the person is diabetes or not. Using data mining techniques the healthcare management predicts the disease and diagnosis of diabetes. In this we used Naïve Bayes, SVM classifier and decision tree classifier for better prediction of diabetes disease. In future we used more attributes for prediction. In the data mining methods to aid people to predict diabetes has gain major popularity. The Naïve Bayes has highest accuracy to predict the person is diabetic or not as compare to SVM and Decision tree classifier. All above methods used to predict diabetes. But if the Patient is detected as diabetes firstly there is a need of finding Control and Un-control condition of diabetes. Because if Patient has diabetes in Un-control condition, may be the patient has severe effect on Patient's Organ like Heart, Eye, Kidney etc. So there is need of finding early Stage which may be help patient for reducing the Severity on Organ or Halting the Severe Effect on Organ.

VI.REFERENCES

[1] Ajinkya kunjir , Harshal Sawant ,Nuzhat F. Sheikh 2017 "Data Mining and visualization for prediction of Multiple Disease in healthcare".

[2] M. Durairaj and V. Ranjani, "Data mining applications in healthcare sector: a study.,"*International Journal of Scientific & Technology Research,* vol. 2, no. 10, pp. 29-35, 2013.

[3] Isha Vashi, Prof. Shailendra Mishra, "A Comparative Study ofClassification Algorithms for Disease Prediction in Health Care", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.

[4] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14.

[5] Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016.Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer.Springer.pp.420–427.

[6] K Beyer, J Goldstein,R Ramakrishnan and U Shaft, "When is 'Nearest neighbor' Meaningful?" 2014.

[7] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491.

[8] Abhishek Taneja. "prediction of heart disease using Data minig techniques" Oriental Journal of computer Science technology .December 2013. Vol .6,No (4): Pgs.457-466.