International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-8, Issue-4, April 2022
ISSN: 2395-3470
www.ijseas.com

Prediction of Diabetes Using Data Mining Techniques

[1]Mrs.S.Ramya

Research Scholar (FT), [1]Department of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore-49.

[2]Dr.D.Kalaivani

Associate Professor & Head, [2]Department of Computer Technology, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore - 49.

Abstract:

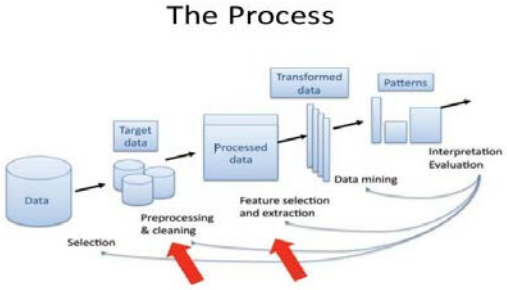
Diabetes is one of the major health issues among youngsters nowadays because of poor diet, family history, lifestyle habits. The medical diagnosis is important based on complicated risks that should be performed accurately and effectively. Based on the result and reports further investigation is needed through diagnosis and treatment are given to the patient. Many numbers of the data can be available in the healthcare systems. The availability and necessary medical data can be analyzed for data analysis tools to extract useful information through pattern recognition. Both data mining and knowledge discovery data are infinite applications including business and research scientific fields. Diabetes is one of the major applications nowadays where data mining tools are very useful to easily diagnose and find the solution. This paper, diagnoses diabetes through data mining tools such as SVM, association rule, clustering, and association. So, Data mining helps to predict and diagnose diseases with a low occurrence of risks. In this paper, the main focus is to make a present detailed survey of various data mining techniques and approaches that have been put to use for the prognosis of diabetes.

Keywords: Diabetes, data mining, machine learning, and knowledge discovery database.

1.INTRODUCTION

1.1 An overview of Datamining

Data mining is the core step in discovering the information from a huge set of data. Knowledge discovery is used in health care organizations to improve the quality of service. Data mining is to be quite successful in the healthcare sector in finding out the hidden patterns that are useful for disease prognosis. These data mining techniques have been successfully applied for the prognosis of diabetes. Diabetes mellitus commonly known as diabetes is a metabolic disorder condition that is characterized by a high level of sugar in the blood. Numerous data mining techniques have been used for designing the model that could aid physicians in predicting diabetes. Machine learning and Data Mining techniques are tools that can be used to improve the analysis and interpretation or extraction of knowledge from the data. The quality of service provides a treatment to the patient effectively to predict and diagnose the disease. The major challenge of health care organizations is mainly to provide a preferable treatment at an affordable cost at a point in time. The healthcare management should maintain the proper patient's database is stored and retrieved the data based on the patient diagnoses. A manual decision that leads to error can affect the quality of service to the patient. By maintaining the health records in computer-based easy to retrieve the data. "Data mining is an emerging field and has emerged as interdisciplinary in nature that has bought together techniques from machine learning, pattern recognition, statistics, databases and visualization for extraction of useful and specific information from large databases". For data mining the phrase "knowledge discovery in databases (KDD)" is often synonymously used. Data are any raw facts that can be in the form of numbers or texts which can be processed by computer systems into useful information. The goal of data mining is to fetch the pattern from a huge amount of data.



After patterns are fetched then they can be implemented in making certain decisions for the development of businesses. The following three steps involved are:-

- Exploration
- Pattern identification
- Deployment

Exploration: This is the first step of data exploration. This data is purified and converted into another form, and the determined nature of data according to the problem.

Pattern Identification: After exploring data, for the sake of specific variables it is refined and also defined for the formation of pattern identification. Then identify and choose the particular patterns that predict the best solution.

Deployment: Then patterns that are chosen are deployed for getting the desired outcome.

1.2 Causes and impact of diabetes disease:

Diabetes mellitus is a major health issue affected nowadays in youngsters because of an unhealthy diet, physical inactivity, and obesity. The madras diabetes. Research Foundation has identified a tool to identify a high risk of developing type2 diabetes mellitus (T2DM) in the future. The major risk is due to family history, personal history, age, physical inactivity. The majority of a case of diabetes can be divided into two categories, type 1 and type2. It is not only a disease but it can lead to heart attack, blindness, kidney diseases, etc. The analysis of diabetes is a major challenge in the medical field because of its abnormal, correlation structure, and complexity in nature. Machine learning methods applied in diabetes research is an important approach to extract information from a huge amount of data. It helps people to easily identify and rectify the disease. Diabetes mellitus commonly known as diabetes is often attributed to changing lifestyle of humankind; it is a metabolic disorder that is characterized by hyperglycemia- a condition of high concentration of glucose in the blood. It is the consequence of defective insulin secretion or defective insulin action or in some cases both and affects the metabolism of the body resulting in raised sugar levels in the blood. Chronic hyperglycemia has a hazardous effect on which leads to dysfunctioning and failures of several organs.

II. METHODOLOGY

We can separate our strategy into four primary segments as follows:

• Data Collection

• Data Preprocessing

• Data Training

• Applications of Machine Learning Algorithms

An overall work flow of our study has been shown in Fig. 1. Import dataset with 340 instances and 27 features Start Data Preprocessing Replace missing data with Mean, Median and Mode Feature Selection Apply Best First Search and Ranker Algorithm Find Most Significant Features Apply 10-Fold Cross-validation technique Apply Classification Algorithms Bagging Random forest Logistic Regression Determine Statistical Matrics Compare Performance End

A. Data Collection:

For this analysis, we have collected data from Khulna Diabetes Hospital, Khulna. The dataset includes total 340 instances with having 27 significant features for each instance. The dataset contains basic information of patients and two types of symptoms: Typical and Non-Typical. The Table I helps to understand the categories of the symptoms.

B. Data preprocessing: To handle missing information we've used two popular and useful functions in WEKA 3.8 (Waikato Environment for Knowledge Analysis) . First, ReplaceMissingValue function has been used to replace missing data. This function swaps every single missing information for nominal and numeric attributes with the modes and means [13]. We've used another function named Randomize which can fill-up the missing field without sacrificing too much performance [13].

C. Data Training: For training all the features of the dataset shown in Table II, we have used 10-Fold Cross-Validation technique. It is a re-sampling 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 [14]. The methodology has a solitary parameter considered K that alludes to the number of algorithms that a given information test is to be part into. It shuffles the dataset randomly, splits dataset into 10 groups and finally abridge the expertise of the model utilizing the example of model assessment scores [15].

D. Applications of ML Algorithms: After having the preprocessed and trained dataset we have applied three algorithms on the dataset. They are: Bagging, Logistic Regression and Random Forest. 1) Bagging (BAG): It is a procurement method that resamples the preparation information to make new models for each example that is drawn [16]. It makes a troupe of arrangement models for a learning plan where each model gives a similarly weighted forecast [14].

Input:
- R, a set of h training tuples
- t, the number of models in the ensemble
- A classification learning scheme (Decision Tree Algorithm, Naive Bayesian, etc.) Output: The ensemble, an associated model, L

III. DATA MINING ALGORITHMS AND TECHNIQUES

2.1 Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and creditrisk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms.

There are types of classification models:

- Support Vector Machines (SVM)
- Classification Based on Association

2.2. **Clustering:** Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Types of clustering methods:

- Model-based technique
- Partitioning technique
- Divisive technique

IV. SURVEY OF LITERATURE (DIFFERENT DATA MINING TECHNIQUES TO FIND OUT DIABETES DISEASES)

As data have become an integral part of any organization and to analyze it for discovering hidden knowledge has become inevitable for improvement in services, same is true for medical field where predictive data mining is used for prognosis of disease at an early stage to pre-empt its effects and to aid physicians in developing contingency plan. The available literature reveals majority of the work that has been carried out on diabetes has focused mainly on developing the methods for prognosis or diagnosis of type II diabetes to reduce its complications, in majority of the cases Pima India dataset has been used for experimentation though methods and tool used have varied.

To predict diabetes applied associative rule mining to discriminating continuous valued attributes an equal interval binning technique was used and for diabetic classification Apriori algorithm was applied and at the end association rules were generated for understanding relationship among measured fields used in prediction.

A Both implementation of two algorithms used are Suppport Vector Machine and Naïve Bayes was used by prognosis of diabetes had reached an accuracy of 97.6%

Hybrid prediction model for prognosis of type II diabetes was proposed by, the system was the combination of two data mining classifiers K-means clustering and decision tree algorithm, the system achieved an accuracy of 92.38%. To predict the risk of heart attack in a diabetic patient applied Naïve Bayes classifier on diabetic data with an accuracy of 74%.

In the research work used three data mining techniques SVM, KNN classify diabetes with an accuracy of 86% decision tree proved to be best among the three and was used further for designing a predictive model for diabetic prognosis.

For prediction of diabetes a model was designed by combining Genetic algorithm with fuzzy logic. The model achieved an accuracy of 80.5%.

Multi layer Perceptron, decision tree, and Naïve Bayes algorithms were applied by on Pima India diabetic dataset for prediction of diabetes, the model designed using Naïve Bayes had an accuracy of 76.30%.

A combination of OLAP and two data mining algorithms C4.5 and ID3 decision tree were used to develop decision support system for diabetic prediction with an accuracy of 74%.

For diabetic prognosis a hybrid predictive model was designed by using data mining classification algorithms SMO, J48, Bagging, J48, Naïve Bayes, Random Forest and AdaBoost were used. K-means clustering was used combined with these techniques for prediction of Positive and Negative cases of diabetes.

Discussions

For a machine learning systems perfection picking a decent assessment methodology is a vital step. Due to privacy concerns medical data is not easily available and is the main factor behind making research in medicinal field weighty. The analysis of work done on diabetes that is presented in this paper has been carried on Pima India dataset which is available online; the dataset consists of nine attributes with 768 instances which are considered to be main factors in diabetes. The attribute list is given in table 1.

Table 1. Pima India Attributes

S.no	Attribute name
1	Diastolic Blood Pressure
2	Plasma Glucose Concentration
3	Age
4	Diabetic Pedigree Function
5	Body Mass Index
6	2-Hr Serum Insulin
7	Class (yes or no)

All but seven attributes are directly related to diabetes while last attribute is used to differentiate between positive and negative cases of diabetes. The majority of the reviewed research presented in this survey used the same dataset for validating their models. The dataset has some of the limitations as it contains the data related to only female population while data related to male population is overlooked, some of the vital attributes like HbA1c value are not considered. There are many missing values in almost all attributes. Majority of the models have used classification techniques although the prediction accuracy has varied.

IV Conclusion:

In this paper was to present the various data mining techniques that have been applied for diabetic data mining for designing predictive models of diabetes. Application of data mining techniques for diabetes is task that puts up a great challenge but it has reduced the human effort drastically with increase in prognostic precision. Development of efficient data mining applications has led to reduction of both constraints cost and time in terms of human resources and expertise. A careful study of various data mining techniques was carried in this study and it could be concluded Decision tree, Support Vector Machines, Naïve Bayes and K-NN were used by researcher in majority of cases individual or some have used combined techniques in order to increase the predictive accuracy. In future, we will try to increase in finding even best techniques for diagnosing of diabetes

disease and also curing them and turn India into a healthy country. In future, we will conduct this study with more algorithms like ANN more specifically with Neuro Fuzzy Inference System, CNN (Convolution Neural Network) and advanced Ensemble Learning algorithms. An expert system can be developed with our analysis to predict diabetes more efficiently and effectively

References:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5592265/
2. **https://www.beneo.com/benefits/human-nutrition-benefit/blood-sugar-management**
3. American Diabetes Association. "Diagnosis and classification of diabetes mellitus." Diabetes care 37. Supplement 1 (2014): S81-S90.
4. Heikki, Mannila, Data mining: machine learning, statistics and databases, IEEE, 1996.
5. Huang F, Wang S, Chan CC. Predicting disease by using data mining based on healthcare information system. In: Granular computing (GrC), 2012 IEEE international conference on (pp. 191– 4). IEEE; August 2012.
6. P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequencing the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, August 2014, pp. 334-339.
7. Sudesh Rao, V. Arun Kumar, "Applying Data mining Technique to predict the diabetes of our future generations", ISRASE eXplore digital library, 2014.
8. Veena vijayan, Aswathy Ravikumar, " Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975-8887) vol. 95-No.17, June 2014.
9. P. Hemant and T. Pushpavathi, "A novel approach to predict diabetes by Cascading Clustering and Classification", In Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on IEEE, (2012) July, pp. 1-7.
10. Valdez, R., Yoon, P. W., Liu, T., & Khoury, M. J. (2007). Family History and Prevalence of Diabetes in the U.S. Population: The 6-year results from the National Health and Nutrition Examination Survey (1999 2004). Diabetes Care, 30(10), 2517-2522. doi:10.2337/dc07-0720