Prediction of Diabetes using Data Mining Technique: A Review

MWNL de Silva[1] and DU Vidannagama[2]

[1]*Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka*
[2]*Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka*
[1]mwndesilva001@gmail.com
[2]udeshika@kdu.ac.lk

***Abstract -** Data mining plays an efficient role in prediction of diseases in health care industry. Diabetes has become one of the major global health problems at present. According to the WHO 2014 report, around 422 million people worldwide are suffering from diabetes. Diabetes is a metabolic disease where the improper management of blood glucose levels led to risk of many diseases like heart attack, kidney disease, eye etc. Many algorithms have been developed for the prediction of diabetes and accuracy estimation. This paper gives detailed evaluation of existing data mining methods used for prediction of diabetes.*

More than twenty research papers had been referred during this research work. And according to those research papers, decision tree related algorithms had been used in most of the research works and this algorithm has given the best performance also. So that, to have the best accuracy and performance in data mining related projects, decision tree algorithm and related algorithms can be used.

Further, it gives some idea about the tools which can be used in data mining. WEKA, Orange, MATLAB, Tanagra and Rapid Miner are some of the data mining tools which are commonly used. Some of the researchers have used more than one tool for their research works. But almost all of the referred papers have used the data mining tool WEKA.

Finally, this paper gives the idea that, using Decision tree related algorithms may give high performance and high accuracy in data mining related works.

Keywords— Data mining, Diabetes, Algorithms

I. INTRODUCTION

Data mining is described as the process of discovering associations, patterns and trends to search through a large amount of data stored in repositories, databases, and data warehouses. Data mining adopts a sequence of pattern recognition techniques and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining can also be known as a process that requires goals and objectives to be specified. Data mining is the process that has played an important role in health care domain. Data mining techniques would be an asset for the diabetes predicting researchers because it can expose hidden knowledge from a huge amount of diabetes related data. Data mining is the method of retrieving useful information and patterns from massive dataset. Especially in health care, the dataset size is huge and dynamic in nature and hard to predict based on statistics. In the context of clinical field, many data mining techniques are proposed.

Diabetes is a chronic disease and a major public health challenge which is occurred worldwide. It occurs when a body is not able to react or outgrowth properly to insulin, which is needed to maintain the rate of glucose. Diabetes can be controlled with the help of insulin injections, a healthy diet and regular exercise but it cannot be cured completely. Diabetes leads to much other disease such as blindness, blood pressure, heart disease, and kidney disease and nerve damages. There are three prime types of diabetes mellitus: Type 1 Diabetes Mellitus results from the body's inability to produce insulin. This was previously referred to as insulin-dependent diabetes mellitus. Type 2 Diabetes Mellitus occurs because of insulin resistance which is a condition in which cells fail to use insulin properly, and for sometimes, also with an absolute insulin deficiency. This type was previously referred to as non-insulin-dependent diabetes mellitus. Gestational diabetes is the third main form and occurs when a pregnant woman previously seems diagnosis of diabetes develop a high blood glucose level.

In order to automate the overall process of diabetes prediction and severity estimation, diabetic databases are needed. This repository of diabetic database helps in the identification of impact of diabetes on various human organs. More the accuracy of prediction is kept, more the chances of accurate severity estimation. Therefore, this paper has presented different prediction methods of diabetes and evaluation of them. Remaining of the research paper is organized as follows: Section-II briefs some of the commonly used data mining techniques for prediction of diabetes with the help of related research works, Section-III presents how the research papers were selected and the findings, Section-IV discusses open source tools for data mining, section v provides an over-all summary of the selected research papers and finally Section-Vi provides the Conclusion of the research work.

II. LITERATURE REVIEW

Generally, data mining consists of several algorithms and techniques for picking out various patterns from large data sets. Data mining techniques can be classified into two categories: supervised learning and unsupervised learning. In supervised learning, a model can be built prior to the analysis. Then the algorithm can be applied to data in order to estimate the parameters of the model. Classification, Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining are some of the common examples of supervised learning.

In unsupervised learning, a model can be created or hypothesis prior to the analysis. (Garg, July, 2013) Applying the algorithm directly to the dataset and observing the results is done in unsupervised learning. Then a model can be created based on the obtained results. Clustering is one of the examples of unsupervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbor have been used for knowledge discovery from large data sets (Sharma, June, 2017). A set of research papers has been selected which includes diabetes predictions using different datamining techniques. A comparative study on different data mining techniques has been done using the selected research papers below.

(Sisodia, 2018) had designed a model which can predict the likelihood of diabetes in patients with maximum accuracy. Therefore, three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes has been used in this experiment to detect diabetes at an early stage.

Experiments had been performed on Pima Indians Diabetes Database (PIDD) which was sourced from UCI machine learning repository. The performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained shows that Naive Bayes outperforms with the highest accuracy of 76.30% comparatively to other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

The main goal of (Joshi, October, 2017) was to predict the diabetes disease and compare the algorithms to find which algorithm provides high accuracy. Finally, the author had tried to select the best algorithm to predict the diabetes disease at early stage. The author has examined how patients' characteristics as well as measurements disturb diabetes cases.

A total of 768 instances were in the selected data set from PIDD (Pima Indian Diabetes Data Set). In this paper, the most known predictive algorithms have been applied, such as KNN, Naïve Bayes, Random forest, and J48. By using these algorithms, the researchers made an ensemble hybrid model by combining individual techniques/methods into one in order to increase the performance and accuracy.

According to the writer, in most studies decision tree algorithm has provided high accuracy. And in this study, Weka and java were used as the tools to predict diabetes dataset.

The paper (Zaveri, April, 2017) had explored various Data mining techniques such as classification, clustering and association which are used by healthcare organization to increase their capability for making decision regarding patient health.

The objective of this paper was to summarize the different algorithms of data mining for the major life-threatening diseases in the field of medical predication. The main focus was to use different algorithm and combination of several targeted attributes for different types of disease identification or predication using data mining techniques. And also, the author has presented a comparative study of different data mining applications and techniques applied for extracting knowledge in healthcare industry.

Finally, the analysis has shown that it is very difficult to name a single data mining algorithm as the most suitable algorithm for the diagnosis and/or prediction of diseases. Because sometimes some algorithms perform better than others, but there are cases when a combination of the best properties of some of the algorithms together give more effective results.

The aim of (Mirza, April, 2018) was to design an efficient data mining procedure for prediction of diabetes by extracting knowledge form the historical medical records. The data was obtained from leading diabetic diagnostic centers of Srinagar (J&K) India.

The data set obtained, contains the record of almost all age groups of population. The main focus was on type 2 diabetes, as it is the most common type affecting nearly 90% of the diagnosed population. The data set contained record of 734 patients. After proper scrutiny of the data, decision tree classifier was applied on it using Waikato environment (WEKA tool) for knowledge analysis's J48 decision tree classifier to develop model. The model achieved an accuracy of 92.5068%.

The aim of (Shetty, 2016) was to discover new and useful patterns to provide meaningful and useful information for the users about the diabetes. Here, a diabetes prediction and monitoring system had been designed and implemented using ID3 classification algorithm. The symptoms causing diabetes were identified and were applied to the prediction model and then, the prediction was done. The monitoring module has analyzed the laboratory test reports of the blood sugar levels of the patient and provides proper awareness messages to the patient through mail and bar chart.

Further, it helped the user to know whether they are diabetic or non-diabetic. According to the research paper the system also raises awareness among the user and helps to keep track of their health status.

The research by (Sathya, 2016) focused on prediction of diabetes using ID3 classification algorithm. ID3 algorithm was well suited for robust missing value. ID3 has given better performance when compared with many other classification algorithms. The UCI Machine learning diabetic dataset has been used for implementation in ID3 algorithm. The dataset contained around 50 attributes and 10Z, 1767 instances. Among this, around 100 instances with 50 attributes have been used for implementation. This implementation has been done with three different modes of runs with 10-fold cross validation of training and test data.

The comparative analysis has been made between the different modes of run and the accuracies and other measures were tabulated and charted for analysis. It was found that the data cleaning with supervised learning method gives better accuracy of 63% while unsupervised learning provided 56% and actual data set without preprocessing gives the poor result of 48% accuracy.

The paper by (Sengamuthu, May, 2018) explored the early prediction of diabetes using various data mining techniques. The dataset has taken 768 instances from PIMA Indian Dataset to determine the accuracy of the data mining techniques in prediction. The analysis has proved that Modified J48 Classifier provide the highest accuracy than other techniques.

Further the author stated that, in the medical field accuracy in prediction of the diseases is the most important factor rather than the execution time. In the analysis of data mining techniques and tools modified J48 Classifier gives 99.87% of highest accuracy using WEKA & MATLAB tool.

The goal of the data mining methodology is to take data from a data set and change it into a reasonable structure for further use. In this research paper, (Shetty, 2017) the examination concentrated on this part of Medical conclusion learning design through the gathered data of diabetes and to create smart therapeutic choice emotionally supportive network to help the physicians. The primary target of this examination was to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes utilizing diabetes patient's database. In this system, the author has proposed the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient’s database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

The best thing about the system is that, it would give accurate results whether the patient is diabetic or not with the help of the knowledge base of the larger dataset that they are going to use.

The paper (Iyer, January, 2015) aimed at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. The research hoped to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients. A summary of the above details can be given as follows.

Author	**Algorithm used**	**Selected data set**
(Sisodia, 2018)	Naïve Bayes SVM Decision Tree	Pima Indians Diabetes Database (PIDD)
(Joshi, October, 2017)	KNN Naïve Bayes Random forest J48	Pima Indians Diabetes Database (PIDD)
(Mirza, April, 2018)	J48 decision tree classifier	Pima Indians Diabetes Database (PIDD)
(Shetty, 2016)	ID3 algorithm	Data sets are obtained from laboratories

(Sathya, 2016)	ID3 algorithm	UCI Machine learning diabetic dataset
(Sengamuthu, May, 2018)	J48 classifier	Pima Indians Diabetes Database (PIDD)
(Shetty, 2017)	Bayesian algorithm and K-NN algorithm.	The diabetes dataset is obtained from following laboratories
(Iyer, January, 2015)	J48 algorithm Naïve Bayes	Pima Indians Diabetes Database (PIDD)

Table 1. Summary table

III. METHODOLOGY

Nearly twenty-five research papers which were related to diabetes prediction and data mining were referred to collect relevant information during this research work. This collection includes review papers and research papers which are about automated prediction systems.

According to those research papers, decision tree related algorithms have been used in most of the mentioned research works, as it is well known of being one of the most important classifiers which is easy and simple to implement. And also, the researchers have used some modified versions of decision tree algorithm.

C4.5 is an algorithm which is used to generate decision trees. And C4.5 is an extension of ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and C4.5 is often referred to as a statistical classifier. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. The following chart presents the algorithms usage in referred research works during the study.

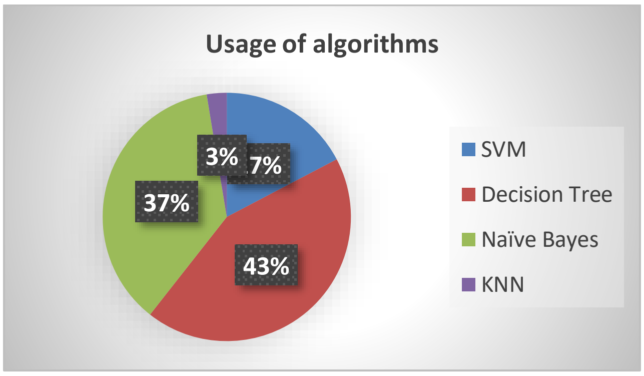

Figure 1. Usage of algorithms in research works

IV. OPEN SOURCE TOOLS FOR DATA MINING

A. Rapid Miner

Rapid Miner is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. It offers an integrated environment useful in machine learning, text mining, data mining, business analytics and predictive analytics. The tool supports various steps useful in data mining including result optimization, visualization and validation (Umadevi, April, 2017).

B. WEKA

WEKA is a world-leading open-source system for data mining. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be appropriate straight to a dataset or called from your own Java code. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. It has the ability to show how are the various relationships between the data sets, clusters, predictive modelling and visualization (Garg, July, 2013).

C. Orange

Orange is an Open source data visualization and analysis for noise and experts in Data mining through visual programming or Python scripting. Orange incorporates various components useful in data pre-processing, feature filtering and scoring, model evaluation, exploration and modelling techniques.

D. MATLAB

A proprietary programming language developed by the MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python. When doing data mining, a large part of the work is to manipulate data. Indeed, the part of coding the algorithm can be quite short since MATLAB has a lot of toolboxes for data mining (Kaur, July, 2014).

E. Tanagra

Tanagra is a free data mining software for academic and research purposes. It proposes several data mining methods for data analysis, statistical learning, machine learning and databases area. The main purpose of Tanagra project is to give researchers and students easy to use data mining software to analyze either real or synthetic data (Radha, August, 2014).

Author	Tool used
(Umadevi, April, 2017)	Rapid Miner
(Garg, July, 2013)	WEKA
(Kaur, July, 2014)	MATLAB
(Radha, August, 2014)	Tanagra

Table 2. Summary table

When referring the selected research papers, the researcher found that almost all the researches were done using the WEKA data mining tool. In some occasions, researchers have used more than one tool for their predictions.

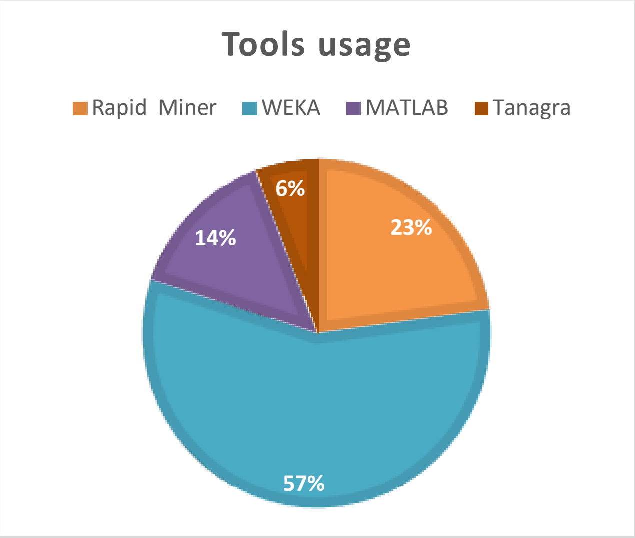

Figure 2. Usage of data mining tools in research works

V. SUMMARY

Author	Algorithm used	Selected data set	Accuracy of the algorithm used	Tool used
(Sisodia, 2018)	Naïve Bayes SVM Decision Tree	Pima Indians Diabetes Database (PIDD)	76.3% 65.10% 73.82%	WEKA
(Joshi, October, 2017)	KNN Naïve Bayes Random forest J48	Pima Indians Diabetes Database (PIDD)	Decision tree algorithm provides highest accuracy	WEKA
(Mirza, April, 2018)	J48 decision tree classifier	Pima Indians Diabetes Database (PIDD)	92.5068 %	WEKA
(Shetty, 2016)	ID3 algorithm	Data sets are obtained from laboratories	94%	Not mentioned
(Sathya, 2016)	ID3 algorithm	UCI Machine learning diabetic dataset	63%	WEKA
(Sengamuthu, May, 2018)	J48 classifier	Pima Indians Diabetes Database (PIDD)	99.87%	WEKA & MATLAB
(Shetty, 2017)	Bayesian algorithm and K-NN algorithm.	The diabetes dataset is obtained from following laboratories	Not mentioned	Not mentioned
(Iyer, January, 2015)	J48 algorithm Naïve Bayes	Pima Indians Diabetes Database (PIDD)	74.86% 79.56%	WEKA

Table 3. Summary table

VI. CONCLUSION

Detection of diabetes in its early stages is very much helpful in providing treatments, since diabetes is a chronic disease.

This review paper concentrates about various data mining techniques and methods which are used for the early prediction of various types of diabetes from the medical data set given by the patient. The literature review gives some idea about the systems which have been implemented using the mentioned data mining techniques.

Further, the review paper gives some idea about the open source data mining tools which help in mining data. Finally, this analysis shows that it is very difficult to name a single data mining algorithm as the most suitable algorithm for the diagnosis and/or prediction of diseases. Because sometimes some algorithms perform better than others, but there are cases when a combination of the best properties of some of the algorithms together give more effective results.

Sometimes the final output of a prediction may get different due to the size of the data set also. The suitable algorithm should be selected according to the data set which is going to be used.

When referring the selected research articles, an idea can be taken that most researchers have preferred to use decision tree related algorithms in their research works. And also, this algorithm has given best performance in most cases. So that, an idea can be taken that the decision tree algorithm and related algorithms may give better performance than the other algorithms.

REFERENCES

Azrar, A., 2018. Data Mining Models Comparison for Diabetes Prediction. *(IJACSA) International Journal of Advanced Computer Science and Applications,* 9(8), p. 4.

Balpande, V., January, 2017. Review on Prediction of Diabetes using Data Mining Technique. *International Journal of Research and Scientific Innovation (IJRSI),* Volume 4, p. 4.

C, T., 2016. A Survey on Diabetes Mellitus Prediction Using Machine Learning Technique. *International Journal of Applied Engineering Research,* 11(3), p. 5.

Dokania, N. K., May, 2018. COMPARATIVE STUDY OF VARIOUS TECHNIQUES IN DATA MINING. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY,* p. 8.

Garg, S., July, 2013. Comparative Analysis of Data Mining Techniques on Educational Data sets. *International Journal of Computer Applications (0975 – 8887),* p. 5.

Iyer, A., January, 2015. DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUE. *International Journal of Data Mining & Knowledge Management Process (IJDKP),* 5(1), p. 15.

Joshi, R., October, 2017. Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology (IRJET),* 4(10), p. 10.

K.Priyadarshini, September, 2017. A Survey on Prediction of Diabetes Using Data Mining Technique. *International Journal of Innovative Research in Science, Engineering and Technology,* 6(11), p. 5.

Kaur, A., March-April, 2018. HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY. *International Journal of Advanced Research in Computer Science,* 9(2), p. 4.

Kaur, G., July, 2014. Improved J48 Classification Algorithm for the prediction of Diabetes. *International Journal of Computer Applications,* 98(22).

Kumar, B. S., December, 2016. A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis. *International Journal of Advanced Research in Computer and Communication Engineering,* 5(12), p. 5.

Meng, X.-H., 16 October 2012. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung Journal of Medical Sciences,* p. 7.

Mirza, S., April, 2018. Applying Decision Tree for Prognosis of Diabetes Mellitus. *International Journal of Applied Research on Information Technology and Computing,* 9(1), p. 7.

pradhan, M., April, 2011. predict the onset of diabetes disease using Artificial Neural Network. *International Journal of Computer Science & Emerging Technologies,* 2(2).

Radha, P., August, 2014. Predicting Diabetes by consequencing the various Data mining Classification Techniques. *International Journal of Innovative Science, Engineering & Technology,* 1(6), p. 5.

Sa'di, S., October, 2015. COMPARISON OF DATA MINING ALGORITHMS IN THE DIAGNOSIS OF TYPE 2 DIABETES. *International Journal on Computational Science & Applications (IJCSA),* 5(5), p. 12.

Sathya, S., 2016. An Effective Prediction of Diabetics Using ID3 Classification Algorithm. *Middle-East Journal of Scientific Research,* p. 5.

Sengamuthu, R., May, 2018. VARIOUS DATA MINING TECHNIQUES ANALYSIS TO PREDICT DIABETES MELLITUS. *International Research Journal of Engineering and Technology (IRJET),* 5(5), p. 4.

Sharma, D. A., August, 2017. A RESEARCH REVIEW ON COMPARATIVE ANALYSIS OF DATA MINING TOOLS, TECHNIQUES AND PARAMETERS. *International Journal of Advanced Research in Computer Science,* 8(7), p. 7.

Sharma, D., June, 2017. A Literature Survey on Data Mining Techniques to Predict Life Style Diseases. *International Journal for Research in Applied Science & Engineering Technology,* 5(6), p. 8.

Shetty, D., 2017. Diabetes Disease Prediction Using Data Mining. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS),* p. 5.

Shetty, S. R. P., 2016. A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique. *I.J. Information Technology and Computer Science, 2016,,* p. 7.

Sisodia, D., 2018. Prediction of Diabetes using Classification Algorithms. *Inernational Conference on Computational Intelligence and Data Science (ICCIDS 2018),* p. 8.

Umadevi, D. B., April, 2017. A Survey on Prediction of Heart Disease Using Data Mining Techniques. *International Journal of Science and Research (IJSR),* 6(4), p. 5.

Zaveri, P. S. H., April, 2017. A Comparative Study of Data Analysis Techniques in the domain of Medicative care for Disease Predication. *International Journal of Advanced Research in Computer Science,* 8(3), p. 3.