International Journal of Research in Engineering, Science and Management
Volume-3, Issue-8, August-2020
journals.resaim.com/ijresm | RESAIM Publishers

A Systematic Literature Review on Detection and Prevention of Diabetes Using Data Warehousing and Data Mining Techniques

S. Deepa[1*], B. Booba[2]
[1]*Research Scholar, Department of Information Technology, Vels Institute of Science, Technology & Advanced Studies, Chennai, India*
[2]*Professor, Department of Information Technology, Vels Institute of Science, Technology & Advanced Studies, Chennai, India*
*Corresponding author: s.deepaonline@gmail.com

***Abstract*: One of the most significant health issue faced by all the human being these days is diabetes. Diabetes is one of the dangerous diseases to cause health care crisis worldwide and also one of the leading causes of mortality and morbidity. The common sites of Diabetes have varied distribution in different geographical locations. The main objective of the present article is to conduct a systematic literature review on detection and prevention of Diabetes of various types i.e., Type – I, Type – II and Gestational Diabetes using several types of data mining and warehousing techniques. This would help the researchers to know various data mining algorithm and method for the prediction of diabetes mellitus. We have analyzed various publications and journals and selected 25 articles which represent various data mining and warehousing methods used for diabetes research for predicting diabetes. The various techniques used are OLAP operations, Decision Tree, Naive Bayes, k-NN, k-means algorithm, classification and clustering. The data mining and warehousing techniques applied in the selected articles were useful for retrieving useful information and framing new hypothesis for further experimentation and improving the health care for diabetic patients by predicting various diseases and find out the efficient ways to treat them in advance.**

***Keywords*: Diabetic patients, Data Mining and Warehousing Techniques, Decision Tree, Naive Bayes, K-NN, k- means Algorithm, OLAP Operations, Types of Diabetes- Type-I, Type-II and Gestational Diabetes.**

1. Introduction

Diabetes occurs when the glucose, or sugar, in the blood is poorly controlled and consistently high. There are three main types of diabetes mellitus namely Type 1 diabetes, Type 2 diabetes and Gestational diabetes.

Type 1 occurs when the body does not produce enough of the hormone that allows cells to absorb and use glucose. This hormone is called insulin. Type 1 diabetes can occur at any age, although it is more common in children and young adults.

Type 2 diabetes is a lifelong disease that keeps your body from using insulin the way it should. People with type 2 diabetes are said to have insulin resistance.

Gestational diabetes is diagnosed for the first time during pregnancy. Like other types of diabetes, it affects how your cells use sugar (glucose). Gestational diabetes causes high blood sugar that can affect your pregnancy and baby's health.

According to the World Health Organization (WHO), India had 69.2 million people living with diabetes in 2015. Nearly 98 million people in India may have type 2 diabetes by 2030, according to a study published in the 'Lancet Diabetes & Endocrinology' journal, found that the amount of insulin needed to effectively treat type 2 diabetes will rise by more than 20% worldwide over the next 12 years.

"The number of adults with Type-II diabetes is expected to rise over the next 12 years due to aging, urbanization, and associated changes in diet and physical activity," said Sanjay Basu from Stanford University, who led the research.

As an alternative to the tedious physical storage of resources it is important to develop a data warehouse specific to diabetes disease and a data mining model to predict diabetes earlier. If a machine learning technique is developed to store a person's medical and general record and predict his predisposition towards diabetes, its type and exact diagnostic method, physicians can directly start treatment immediately without wasting the precious time in different methods of diagnosis. There have been multiple data mining techniques in health care and allied industries and specifically with respect to Type-I & Type-II of diabetes.

2. Aims and Objectives

The main aim of the present literature review is to find out the articles which develop a multidimensional architectural diabetes data warehouse specifically to store and process diabetes-related database which include patient's general and medical records by using OLTP and OLAP technologies simultaneously.

The data will be coming from operational systems and external sources. To create the data warehouse, diabetes data

will be extracted using query engines from source systems like questionnaire, diabetes institute database, etc. which will be cleaned (e.g., to detect and correct errors), transformed (e.g., put into subject groups or summarized), and loaded into a data store (i.e., placed into a data warehouse) which can be further used to predict a person's predisposition towards diabetes and generate the risk level for a particular type of diabetes and the exact method of clinical diagnosis.

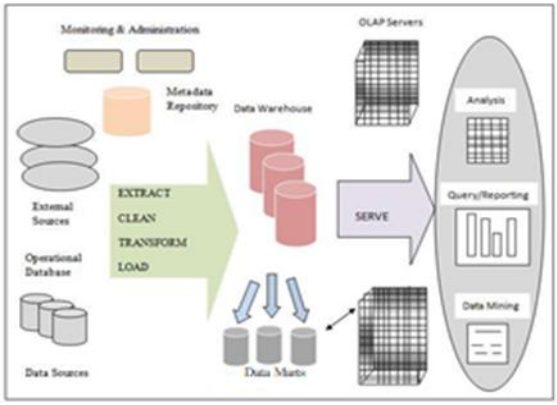

Fig. 1. DIABETES Data Warehousing Architecture using ECTL, OLTP, and OLAP Servers

3. Methodology

Various publications and journals has been analyzed which shows that the following data mining techniques have been applied for predicting diabetes.

A. Naive Bayes

The Naive Bayes classification is based on Bayes theorem of probability to predict the class of unknown data sets. It is one of the most efficient and scalable learning algorithms in data mining.

Naive Bayes model is easy to build and particularly useful for very large data sets. It is an eager learning classifier and it could be used for making predictions in real time. It is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.

Naïve Bayes Classifier and Collaborative Filtering together builds a recommendation system that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

B. Decision Tree

Decision tree as the name goes, it uses a tree-like model for decisions. It is a commonly used tool in data mining for deriving a strategy to reach a particular goal. Decision tree is a classification technique under predictive data mining which aims to find a method of early prediction of the disease of diabetes using a decision-tree algorithm. Diabetes is a condition in which the body is unable to regulate the level of sugar in blood which can lead to various serious conditions like neuropathy, retinopathy and nephropathy and diabetic foot etc. in human body. By using this technique, large amount of data and knowledge can be extracted from data sets by asking certain number of questions on the sample data and then classifying the information obtained into certain classes.

C. K- Nearest Neighbor's Algorithm (K-NN)

K-NN is one of the most essential classification algorithms in machine learning. It is also known as Lazy Learner method. It is a simple technique which stores all the cases and classified new ones based on the similarity measure. Since it is non-parametric, it does not make any underlying assumptions about the distribution of data. It is an instance based method which discovers the unidentified data points using previously known data points and classified data points according to voting system.

D. K-means Algorithm

K-means clustering is a method of vector quantization, originally from signal processing that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. It is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem.

The way k-means algorithm works is as follows:

1. Specify number of clusters k.
2. Initialize centroids by first shuffling the dataset and then randomly selecting k data points for the centroid without replacement.
3. Keep iterating until there is no change to the centroids.

E. Classification via clustering

In clustering, a group of different data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data. After the classification of data into various groups, a label is assigned to the group. It helps in adapting to the changes by doing the classification.

A clinical Decision Support System based on OLAP with data mining to diagnose whether a patient can be diagnosed with diabetes with probability high, low or medium. The system is powerful because it discovers hidden patterns in the data and can, it enhances real-time indicators and discovers bottlenecks and it improves information visualization.

4. Results

We have analyzed various publications and journals and selected 25 articles for our literature review out of which 10 articles are summarized in a tabular form which are shown below.

Table 1
Data Mining and Warehousing Techniques for Predicting Diabetes Mellitus

Author Name (Year)	Research Topic	Data Mining and Warehousing Methods used	Findings	Conclusion
Avinash Bharti, Sanjeev Sharma (2020)	Data Mining Techniques: A Survey	Clustering, Naïve Bayes, K-NN, Support Vector Machine, Partitioning method, Hierarchical method, Density-based method	Few Major issues- It becomes difficult to cover a vast range of data that can meet the client's requirement. We need a lot of background information, collected in data warehouses and analysed with various data mining algorithms to mine useful knowledge out of it.	In this paper, the authors have presented various data mining techniques and their applications and major issues.
Amira Hassan Abed, Mona Nasr (2019)	Diabetes Disease Detection through Data Mining Techniques	Clustering using k-means algorithm	The proposed cluster modeling used 2733 instances within 12 attributes for each one of the dataset. The used data is preprocessed in order to remove the inconsistent or unwanted data, and handle missing values to finally obtain faster processing time.	the research attempt to employ Data mining techniques based on the utilization of Patient dataset through the Business intelligence application to provide the significant results to make the right decision at the right time.
Md Imran Alam, Anu Bharti (2019)	Malady Prediction of Cardiovascular Diseases, Diabetes and Malignant Neoplasm in Lungs Based using Data Mining Classification Techniques	Bayes classification, C4.5, Naïve Bayes, Neural network.	Here c4.5 is superior to bayes technique	The principle objective of this paper was to recognize the most well-known information mining calculations, actualized in weka tools.
Steffi J.et.al., (2018)	Predicting Diabetes Mellitus using Data Mining Techniques	Decision Tree, Naïve Bayes, Logistic Regression,ANN, SVM	Decision Tree and Logistic Regression are equally good based on their Accuracy measures, the Naïve Bayes algorithm has the Second highest accuracy, followed by ANN and the most lowest accuracy is predicted in the SVM algorithms	Data mining classification algorithms are used to model actual Prediction of Diabetes Mellitus and a comparative analysis are made between them by making use of their Metric Measures say Accuracy, Precision, Sensitivity, Specificity and F1 Score.
R.S. Suryakirani, R. Porkodi (2018)	Comparative Study and Analysis of Classification Algorithms In Data Mining Using Diabetic Dataset	Classification Algorithm, Naïve Bayes, Random Tree, Decision Tree, J48.	J48 classifier gives better accuracy. The second best algorithm is the decision tree and then random tree and finally Naïve Bayes gives the least accuracy.	The performance of an algorithm is dependent on the domain and the type of the data set.
D. Jeevanandhini, E.Gokul Raj, V.Dinesh Kumar, N. Sasipriyaa (2018)	Prediction of Type2 Diabetes Mellitus Based on Data Mining	Decision Tree J48, KNN Classifier, Random Forest, SVM, K-Means, K- NN	SVM achieves higher accuracy as compared to other data mining techniques.	This study can be used to select best classifier for predicting diabetes.
Ramin Assar et.al. (2017)	Heart Disease Diagnosis Using Data Mining Techniques	Decision tree, K-nearest neighbor, Bayesian network, Support vector machine	SVM and Naïve Bayes achieved the highest accuracy, followed by KNN (k=7 resulted in the best accuracy as compared to other values) and decision tree, respectively.	Selecting different data mining techniques and implementing them on the selected dataset, SVM technique achieved the highest accuracy (84.33%).
Nimna Jeewandara, PPG Dinesh Asanka, (2017)	Data Mining Techniques in Prevention and Diagnosis Of Non Communicable Diseases	Clustering, Association rule, Decision tree, Regression, ANN, Naïve Bayes, K-means	Produced different accuracy, sensitivity and specificity according to their algorithms and the variables in large complex data sets.	The results can be more accurate and validated using these data mining techniques.
Haldurai Lingaraj et.al., (2015)	Prediction of Diabetes Mellitus using Data Mining Techniques: A Review	Decision Tree, Naïve Bayes, K-NN, clustering, ANN	Different approaches for the prediction of Diabetes Mellitus and its types were concentrated in this study	data mining techniques are applied in health care sector in order to predict various diseases and to find out efficient ways to treat them as well
Miroslav Marinov, M.S.,et.al., (2011)	Data-Mining Technologies for Diabetes: A Systematic Review	Selected 17 articles representing various data-mining methods used for diabetes research.	The applications of data-mining techniques in the selected articles were useful for extracting valuable knowledge and generating new hypothesis for further scientific research/experimentation and improving health care for diabetes patients.	Data mining can significantly help diabetes research and ultimately improve the quality of health care for diabetes patients.

5. Conclusion

Different approaches for the detection and prevention of Diabetic patients and its types are concentrated in this study. Various data mining techniques used to extract useful information from existing large volume of data which enable us to gain more knowledge and help to do further research in health care sector of diabetes mellitus.

In this way the data mining and warehousing techniques are applied in health care sector in order to predict various kinds of diabetes and to find out an efficient way to treat them.

References

[1] Avinash Bharti, Sanjeev Sharma (2020), Data Mining Techniques: A Survey, Tathapi (UGC Care Journal), Vol-19-Issue-13-May-2020, Pp. 429-434.
[2] Amira Hassan Abed and Mona Nasr (2019), "Diabetes Disease Detection through Data Mining Techniques", Int. J. Advanced Networking and Applications, Volume: 11 Issue: 01 Pages: 4142-4149(2019).
[3] Md Imran Alam, Anu Bharti (2019), Malady Prediction of Cardiovascular Diseases, Diabetes and Malignant Neoplasm in Lungs Based using Data Mining Classification Techniques, Journal of the Gujarat Research Society, Volume 21 Issue 16, pp. 1164-1179.
[4] Spandana Vadloori, Yo-Ping Huang and Wei-Chi Wu (2019), Comparison of Various Data Mining Classification Techniques in the Diagnosis of Diabetic Retinopathy, Acta Polytechnica Hungarica, Vol. 16, No. 9, pp. 27-46.
[5] J. Steffi, R. Balasubramanian and K. Aravind Kumar (2018), Predicting Diabetes Mellitus using Data Mining Techniques-Comparative analysis of Data Mining Classification Algorithms, International Journal of Engineering Development and Research, Volume 6, Issue 2, Pp.460-467.
[6] R. S. Suryakirani, R. Porkodi (2018), Comparative Study and Analysis of Classification Algorithms in Data Mining Using Diabetic Dataset, International Journal of Scientific Research in Science and Technology, Volume 4, Issue 2, pp. 299-304.
[7] Shuja Mirza, Sonu mittal, Majid Zaman (2018), Design and Implementation of Predictive Model for Prognosis of Diabetes using Data Mining Techniques, Volume 9, No. 2, pp. 393-398.
[8] Ramin Assari, Parham Azimi and Mohammad Reza Taghva (2017), Heart Disease Diagnosis Using Data Mining Techniques, International Journal of Economics & Management Sciences, Volume 6, Issue 3.
[9] M. Porkizhi (2017), A Study of Data Mining Techniques and its Applications, IJSART, Volume 3 Issue 4, pp. 1402-1406.
[10] K. Saravananathan and T. Velmurugan (2016), Analyzing Diabetic Data using Classification Algorithms in Data Mining, Indian Journal of Science and Technology, Vol 9(43).
[11] Haldurai Lingaraj et.al., Prediction of Diabetes Mellitus using Data Mining Techniques: A Review, Journal of Bioinformatics & Cheminformatics, Volume 1, Issue 1, 2015.
[12] Ravneet Jyot Singh, Williamjeet Singh (2014), Data Mining in Healthcare for Diabetes Mellitus, International Journal of Science and Research (IJSR), Volume 3 Issue 7, pp. 1993-1998, 2014.
[13] Miroslav Marinov et. al. (2011), "Data-Mining Technologies for Diabetes: A Systematic Review", Journal of Diabetes Science and Technology, Volume 5, Issue 6, 2011.
[14] https://www.geeksforgeeks.org/k-nearest-neighours/
[15] https:// upgrad.com/blog/cluster-analysis-data-mining/
[16] https://www.indiatoday.in/education-today/gk-current-affairs/story/98-million-indians-diabetes-2030-prevention-1394158-2018-11-22