International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 10 | Oct 2020 www.irjet.net p-ISSN: 2395-0072

Diagnosis and Prediction of Diabetes Patient data by using Data Mining Techniques

Er.Lovepreet Singh[1], Er.Parminder Singh[2], Dr. Naveen Dhillon[3]

[1]M.Tech scholar in R.I.E.T, Phagwara
[2]Head of Computer Science Department in R.I.E.T, Phagwara
[3]Principal in R.I.E.T, Phagwara

Abstract - *Diabetes is a disease which is affecting many people now-a-days. Diabetes is a chronic disease caused due to the expanded level of sugar addiction in the blood. Various automated information systems were outlined utilizing various classifiers for anticipate and diagnose the diabetes. Due to its continuously increasing rate, more and more families are unfair by diabetes mellitus. Most diabetics know little about their risk factor they face prior to diagnosis. Data mining approach helps to diagnose patient's diseases. It has played an important role in diabetes research. It would be a valuable asset for diabetes researchers because it can unearth hidden knowledge from a huge amount of diabetes-related data. The primary target of this examination is to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. Clustering is the process of partitioning the data or objects into the same class and data in one class is more similar to each other than to those in other cluster. In this research we present the comparison of different clustering techniques using Waikato Environment for Knowledge Analysis or in short (WEKA) by using diabetes data set. The algorithm or methods tested are DBSCAN, filtered Cluster and K-MEANS clustering algorithms. This research present a comparative analysis for various clustering techniques on diabetes dataset.*

***Key Words*: DIABETES, K-MEANS, DBSCAN AND WEKA.**

1. INTRODUCTION

Data Mining is used to invent knowledge out of data and exhibiting it in a condition that is easily understandable to humans. It is a process to inspect large amounts of data collected. Information technology plays a vital role for implementing the Data mining techniques in various sectors like banking, education, etc. [1].In the field of medical domain data mining can be effectively used for the prediction of diseases by using various data mining techniques. There are two predominant goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest [2]. Description focuses on finding the patterns detailing the data that can be interpreted by humans. Basic conception of growth and characteristics affecting diabetes from external sources is very much essential before constructing predictive models. The idea is to predict the diabetes and to find the factors responsible for diabetes using data mining methods [3].Data mining techniques can be used for early prediction of the disease with greater quality in order to save the human life and it will also reduce the treatment cost.

Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced. There are three main types of diabetes mellitus:

Type 1 DM results from the pancreas's failure to produce enough insulin. This type was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". The cause is unknown. The type-1 diabetes is affected by the young people and below 20 years of age [4]. In type 1 the pancreatic cells will get affected and fail to function. Because of nil secretion of insulin, the type-1 diabetic people suffer throughout their life and depend on insulin injection. The type1 diabetic patients should regularly follow exercises and healthy diet as suggested by dietitians [5].

Type 2 DM begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses a lack of insulin may also develop. This type was previously referred to as "non-insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The most common cause is excessive body weight and not enough exercise [6].

- Gestational diabetes is the third main form and occurs when pregnant women without a previous history of diabetes develop high blood sugar levels [7]. According to recent study of diabetes, it is found that around 18% of pregnant women have diabetes. Pregnancy during older age may have a risk of developing the gestational diabetes [8].

2. LITERATURE REVIEW

Fikirte Girma Woldemichael, et.al (2018) proposed a study in which the data mining techniques were used to predict the patients suffering from diabetes. J48, naïve bayes, back propagation and support vector machine were some commonly known classifiers used to predict diabetes from patients in the recent research studies. Large value learning rate was included in the neural networks used in these

systems such that the performance of systems could be improved.

Ioannis Kavakiotis, et.al (2017) reviewed the various machine learning based Diabetes mellitus (DM) detection techniques proposed by different authors over the time. From several clinics and biological fields, the data was collected to create the datasets on which the experiments could be performed. The supervised learning techniques were used in around 85% of the experiments and the remaining ones preferred unsupervised learning techniques.

Yu-Xuan Wang, et.al, (2017) proposed a method to design operating system that used the support of data mining and machine learning. With the help of this it becomes possible to discover a new, automatized way by which optimization of the complex algorithms become simple and easy to use. For the validation of the proposed method cache design was utilized that automatically control the replacement of cached contents to make decisions. All the collected data from the system was analyzed when reply is obtained from a data miner. As per performed experiments, it is concluded that proposed method provides effective results.

Zhiqiang Ge, et.al, (2017) presented a review on existing data mining and analytics applications by the author which is used in industry for various applications. To the semi-supervised learning algorithms an application status was given in this paper. In the process of industry both the methods unsupervised and supervised machine learning is widely used for approximately 90%-95% of all applications. In the recent years, the semi-supervised machine learning has been introduced. Therefore, it is demonstrated that an essential role is played by the data mining and analytics in the process of industry as it leads to develop new machine learning technique.

Bayu Adhi Tama, et.al (2016) presented in this paper a chronic disease that causes major causalities in the worldwide that is Diabetes. As per International Diabetes Federation (IDF) around the world estimated 285 million people are suffering from diabetes. This range and data will increase in nearby future as there is no appropriate method till date that minimize the effects and prevent it completely. The major issue was the detection of TTD as it was not easy to predict all the effects. Therefore, data mining was used as it provides the optimal results and help in knowledge discovery from data

Jahin Majumdar, et.al, (2016) proposed a model in which SVM techniques were used as it provided the accuracy and heavy in the computational functions. The accuracy level of SVM is measured with the help of dataset. In order to improve the data classification and pattern recognition in Data Mining mainly feature selection various existing approaches were focused and experimented. As per performed experiments, it is concluded that comparison between the existing techniques was done in order to find out the best method. The theoretical limitations of existing algorithms were overcome by proposed method.

Aiswarya Iyer, et.al (2015) presented a real world medical problem for the automatic diagnosis of diabetes. The key treatment for the detection of diabetes is its early detection. In this paper, for the actual diagnosis of diabetes to model local and systematic treatment, Decision Trees and Naïve Bayes were utilized. As per performed experiments, it is concluded that proposed model provide the effective and efficient results. Performance of the proposed model was investigated for the diabetes diagnosis problem and on the basis of result adequacy of the proposed model was demonstrated.

3. RESEARCH METHODOLOGY

The diabetes prediction problem is complex in nature due to available of large number of attributes in the dataset. The voting is the classification method which is the combination of several classification methods. Several classifiers are collected to generate one ensemble classifier called the voting based classifier. For exploiting the various peculiarities of each algorithm, they are trained and evaluated in parallel individually. Either “hard” or “soft” voting is implemented in this classifier. The final class label is predicted as the class label that has been predicted most frequently by the classification models in “hard” voting. However, the class-probabilities are averaged for predicting the class labels in “soft” voting.

Data mining is the extraction of intriguing patterns or information from huge stack of data. In other words, it is the exploration of links, associations and overall patterns that prevail in large databases but are hidden or unknown. Data mining is used in classification, clustering, regression, association rule discovery, sequential pattern discovery, outlier detection, etc. Clustering techniques have numerous applications in various fields including, artificial intelligence, pattern recognition, bioinformatics, segmentation and machine learning.

The methodology describes all the steps according to which comparative analysis of clustering algorithms is performed.

Step1. Choose the clustering algorithms: To perform the comparative analysis, these clustering algorithms are chosen namely K-means, filtered cluster and Make Density.

Step2. Choose the dataset: The various data set has been chosen from specific location where it is stored. The file format is CSV.
Step3. Load data on WEKA: Load data file for further analysis.
Step4. Normalize data: After loading of the dataset the next step is to normalize the dataset using the WEKA tool through filter tab. Select normalize filter and apply on the same data set. Save the result using save button.
Step5. Apply clustering algorithms: Apply the all clustering algorithms on un-normalize as well as normalize dataset.
Step6. Store the result: After running all algorithms, results are stored into the tabular forms and based on number of iteration, sum of squared error, time taken to build clusters, correctly clustered data, and comparative analysis is performed.
Step7. Plot the graph: Represent results in graphical format.

4. RESULT AND DISCUSSION

The Proposed system was executed on WEKA tool. First the entire dataset was pre-processed by applying Simple K-means, DBSCAN and Filtered Cluster algorithm. We will analysis of different parameters by applying dataset of diabetes like execution time, accuracy of data and number of clusters

4.1 DATA SET

I am using "information sharing in Diabetes" dataset it has real values there are approximately 800 instances. Some of them will be used as an input attributes and other are used as an output attributes.

Table 1: Attributes of Data Set

PREGNANT
PLASMA
BLOOD PRESSURE
SKIN
SERUM INSULIN
MASS INDEX
PEDIGREE FUNCTION
AGE
CLASS

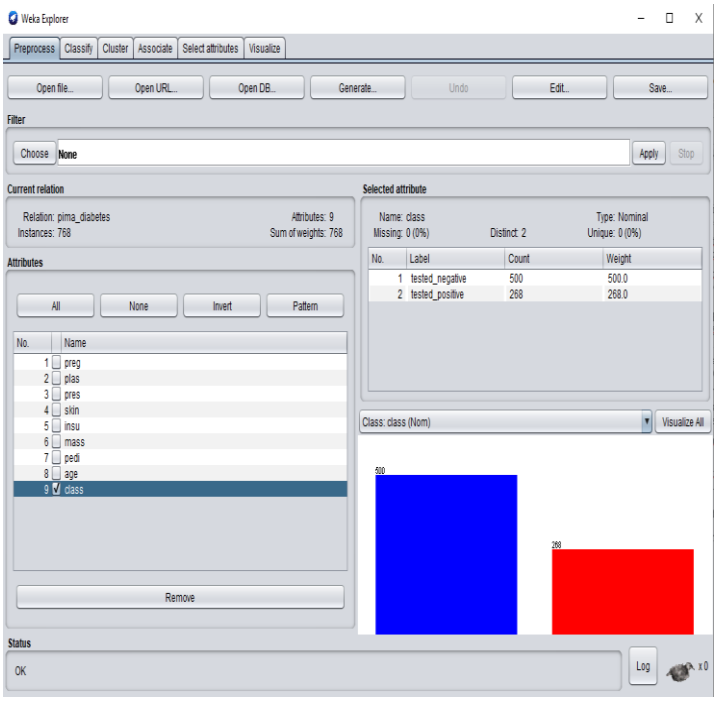

Figure 1: WEKA Explorer interface

In **Figure** 1 the last attribute of dataset is taken as a class attribute by the WEKA. This attribute contains two types .The count of number of instances under each type in the dataset is shown numerically as well as graphically.

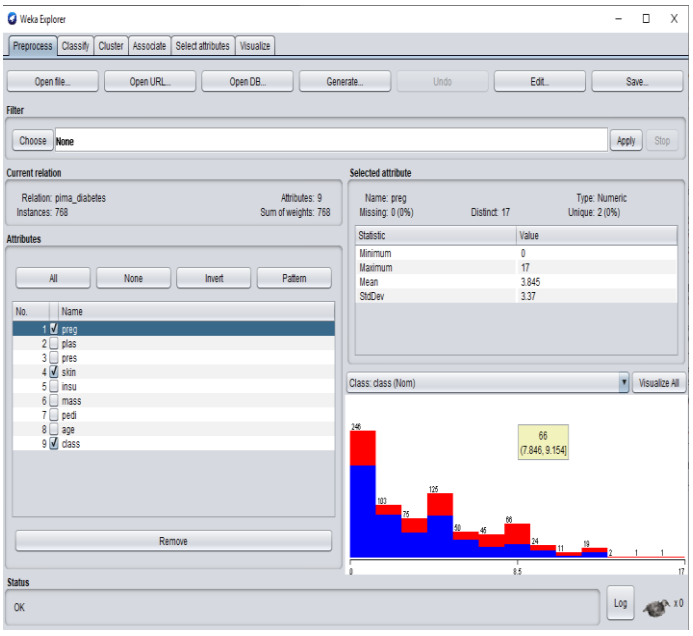

Figure 2: Diabetes skin instances

In **Figure** 2 shows the count of number of skin instances under each type in the dataset is shown numerically as well as graphically.

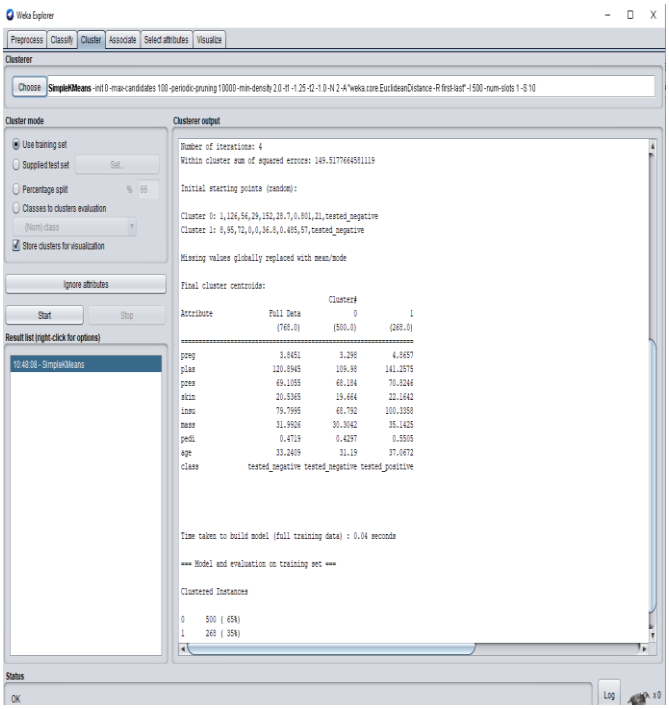

Figure 3: Shows the implementation of K-Means

4.2 RESULT

We are using the WEKA tool for implementation thesis. To import the dataset click on open file and select the data set, which is must be a CSV file. After implementation of these algorithms on data set, the following results obtained.

Table 2: Results of Diabetes Data Set

Parameters	DBSCAN	K-MEANS	Filtered cluster
NO. OF CLUSTERS	2	2	2
ERROR RATE	101.97	149.51	261.83
COMPUTAION TIME	0.01	0.04	0.01
ACEESING TIME	FAST	SLOW	FAST

5. CONCLUSIONS

Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, but not limited to, numerical analysis, pattern matching and areas of artificial intelligence such as machine learning, neural networks and genetic algorithms. In this research, WEKA an open source data mining tool is used for the analysis of diabetes database.

In this research data mining technique applied to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes or not. Different classification algorithms are applied to Pima Indians Diabetes Database and the result obtained is tabulated in table. This research can be extended by applying association mining. For this subset of dataset is converted into required form. This work extends to utilize the implementation of different data.

REFERENCES

[1] Alexis Marcano-Cede˜no, Diego Andina, "Data mining for the diagnosis of type 2 diabetes", IEEE, Vol. 11, issue 3, pp. 9-19, 2016.

[2] B. M. Patil, R. C. Joshi, Durga Toshniwal, "Association rule for classification of type -2 diabetic patients", 2010 Second International Conference on Machine Learning and Computing, Vol. 8, issue 3, pp. 7-23, 2010.

[3] Prova Biswas1,2, Ashoke Sutradhar3, Pallab Datta, "Estimation of parameters for plasma glucose regulation in type-2 diabetics in presence of meal", IET Syst. Biol., 2018, Vol. 12 Iss. 1, pp. 18-25, 2018.

[4] MS.Tejashri n. Giri, prof. S.r.todamal, "data mining approach for diagnosing type 2 diabetes", international journal of science, engineering and technology, vol. 2 issue 8, 2014.

[5] P. Suresh Kumar and V. Umatejaswi, " Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017.

[6] M. Sharma, G. Singh, R. Singh, "Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques", Elsevier, vol. 5, pp. 202-222, 2017.

[7] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", ScienceDirect, Vol. 11, issue 3, pp. 12-23, 2018.

[8] Yan Luo, Charles Ling, Ph.D., Jody Schuurman, Robert Petrella, MD, "GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing", 2014 IEEE International Conference on Data Mining, Vol. 9, issue 8, pp. 12-23, 2014.

[9] Abdelghani Bellaachia and Erhan Guven (2010), "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, 2010, pp. 234-239.

[10] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, vol. 7, 2010, pp. 123-128.

[11] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, vol. 12, 2012, pp. 959-963.

[12] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S (2012), "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, vol. 7, 2012, pp-56-62.

[13] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," Middle-East Journal of Scientific Research, vol. 5, 2012, pp. 959-963

[14] Kajal C. Agrawal and Meghana Nagori (2013), "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm", International Conf. on Advances in Computer Science and Electronics Engineering, vol. 23, 2013, pp. 546-552.

[15] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, 2017, pp- 215-227

[16] Fikirte Girma Woldemichael, Sumitra Menaria, "Prediction of Diabetes Using Data Mining Techniques", 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Pages: 414 – 418