Prediction of Diabetes Using Support Vector Machine

Harwinder Kaur,ME Student,University Institute of Engineering(CSE),Chandigarh University Gharuan,Mohali,Punjab,India,harwinder.cu11@gmail.com

Gurleen Kaur,Assistant Professor, University Institute of Engineering(CSE),Chandigarh University Gharuan,Mohali,Punjab,India,gurleen.cse@cumail.in

ABSTRACT - Data mining is described as the process in which important information is taken out from the raw data. To acquire essential knowledge it is important to extract more data. Diabetes is a chronic disease that is also known as Non-Insulin Dependent Diabetes Mellitus, or Adult Onset Diabetes Mellitus. The adequate insulin is produced by the patient, which cannot be utilized by body due to lack of sensitivity to insulin by the cells of the body. In this existing work, the technique of Support Vector Machine(SVM) is applied for the prediction of diabetes. The SVM classifier has less accuracy and high execution time for the prediction. To improve the accuracy of prediction the voting based classification approach will be applied for the diabetes prediction. The proposed method will be implemented in python and results will be analyzed in terms of accuracy, precision, recall and execution time.

Keywords: Support Vector Machine, Data mining, Diabetes prediction.

I. INTRODUCTION

Introduction to Data Mining

Data mining is described as the process in which important information is extracted from the raw data. In order to acquire essential knowledge it is essential to extract large amount of data. This process of extraction is also known as misnomer. Currently in every field, there is large amount of data is present and analyzing whole data is very difficult as well as it consumes a lot of time. This present data is in raw form that is of no use

hence a proper data mining process is necessary to extract knowledge [1]. The process of extracting raw material is characterized as mining. This is a world where having a lot of information leads to power and success and this is possible only because of sophisticated technologies such as satellites, computers. With the advent in the technology in the mass digital storage and computers it becomes easy to handle large amount of information by which different types of data is stored.

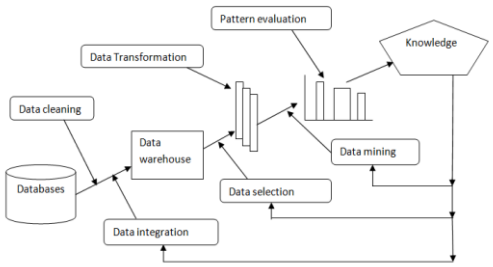


Fig1:data mining process

- **Association:** Association is classified as the best technique among others in the data mining technique. For the transaction of the similar data from one particular image to other, can be done with the help of association in which a pattern is discovered on the basis of relationship [2]. This association technique has been utilized to predict the presence of diabetes in the body and also provide the analysis about the relationship of different attributes. For the prediction of the disease, all the risk factor in the patients is sorted out.
- **Clustering:** In the data mining technique, clustering is a technique in which clustering of objects are identified. An automatic technique has been utilized for this purpose as it has the similar characteristics. This clustering technique defined the classes and objects in order to define the process that how objects are assigned into a predefined classes. The prediction of heart disease becomes feasible with the help of clustering technique in which list of patients which have same risk factor are clustered. A separate list of patients is made using this technique.
- **Classification:** In the data mining technique, classification is a method that is used in machine learning. In classification of the data, each item in the data set is classified into predefined set of classes or groups[5]. Decision trees, linear programming, neural network and statistics are the mathematical techniques that have been utilized by the classification method.
- **Prediction:** The connection between ward variable and free factor in the data mining -technique is

 DOI : 10.35291/2454-9150.2019.0076

discovered by the prediction technique. In the various fields this techniques can be utilized in order to anticipate benefit for future. Therefore, dependent variable is referred as profit and independent variable is referred as sale. Historical sale and profit data has been utilized for the prediction of profit using a fitted regression curve [3].

- **Diagnosing Diabetes using Data Mining Techniques**

The gathering of metabolic diseases in which a person has high blood sugar is commonly referred as diabetes and in the scientific term as Diabetes mellitus. There are twso reasons for the presence of high blood sugar in the body: (1) enough insulin is not produced by the pancreas, (2) no response by cells to the produced insulin. Hence, it is the condition that occurs in the human body due to absence of appropriate insulin[8]. There are various types of diabetes exists such as diabetes insipidus.

Diabetes is due to either the pancreas not creating enough insulin or the cells of the body not reacting appropriately to the insulin delivered. There are three main kinds of diabetes mellitus:

- **Type 1:** One of the types of Diabetes mellitus (DM) which is also referred as previously "insulin-dependent diabetes mellitus" (IDDM). Due to failure of pancreas, it results in DM as it is unable to produce required insulin. The causes of type-1 are unknown. The main target of the type-1 diabetes is youngsters and beneath 20 years old. The main effects of this type 1 diabetes are to damage the functioning of pancreatic cells in the body. The secretion of insulin in the body is nil in the type-1 diabetes due to which patient suffers throughout their life and depend on insulin injection. Regularly follow exercises and healthy diet can maintain the health of type1 diabetic patients [7].
- **Type 2:** It is the disease in which DM resist the occurrence of insulin in the body or described as the situation where cells fails to acknowledge insulin appropriate. With the increase of disease, the insulin produces less in the body. This type previously was known as "non-insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The significant reason of this disease is exorbitant body weight hence, appropriate exercise is required[4].
- **Gestational diabetes's:** It is the third form which occurs when women who is pregnant suffer from high blood sugar levels without a previous history of diabetes. As per analysis, it is analysed that around 18% of pregnant women have this type of sickness. The gestational diabetes occur when pregnancy conceive during older age [6].

- **TYPE 2 diabetics in Data Mining:**

Type II - Diabetes is a chronic disease that is also known as Non-Insulin Dependent Diabetes Mellitus (NIDDM), or Adult Onset Diabetes Mellitus. The adequate insulin is produced by the patient, which cannot be utilized by body due to lack of sensitivity to insulin by the cells of the body. At the age of 40, type II disorder occurs mostly in human beings. Devastating results are provided by the diabetic foot among the chronic diabetic as it produce various complications. Loss of sensation is also experienced by the Diabetes patients in their feet even a small injury can cause infection that is very difficult to cure. Foot ulcers problem also occur in the 15% of patients suffering with diabetes due to nerve damage and reduced blood flow. Diabetes in the person minimizes the vision to see and also cause common blindness and cataracts the diabetic person. Every year more than 50,000 leg amputations take place in India due to diabetes [7].

II. LITERATURE REVIEW

Han Wu, et.al (2018) implies to predicting type 2 diabetes mellitus (T2DM). Nowadays, influence of diabetes mellitus is increased and it affects more families. Due to this disease millions of people in the worldwide us suffering. The main objective in this document is to improve the accuracy of the prediction model and to more than one dataset model is made adaptive in nature. Proposed model comprised of two parts based on a series of preprocessing strategy [13]. The two phases improved K-means algo or the logistic regression algorithm. So as to contrast the outcomes and different techniques the Pima Indians Diabetes Dataset and the Waikato Environment was utilized for Knowledge Analysis toolkit. As per performed experiments, it is concluded that proposed model show better accuracy as compared to other methods and also provide the sufficient dataset quality. In order to evaluate the quality of the model it is applied to other diabetes dataset, in which good performance is shown by both the methods.

P. Suresh Kumar, et.al (2017) proposed a model that overcome all the problems such as clustering and classifications from the existing system by applying data mining method. This method is to diagnose this type of diabetes and from the collected data a security level for every patient. There are various affects of this disease due to which most of the research is done in this area [12]. All the collected data of the 650 patient's was used in this paper for the investigation purpose and its affects are identified. In order to cluster the entire dataset Simple k-means algorithm was used. It is divided into three datasets such as cluster-0 - gestational diabetes, cluster-1 for type of 1 diabetes, cluster-2 for type of 2 diabetes. In the classification model, this clustered dataset was used as input that is used for the classification process such as patient's risk levels of diabetes as mild, moderate and

 DOI : 10.35291/2454-9150.2019.0076

severe. In order to diagnose diabetes, performance analysis of different algorithms was done. On the basis of obtained result the performance of each classification algorithm is measured.

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 [10]. 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. Type 2 diabetes (TTD) is the most common type of diabetes. 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. In the data mining process, support vector machine (SVM) was utilized that acquire all the information extract all the data of patients from previous records. The early detection of TTD provides the support to take effective decision.

Aiswarya Iyer, et.al (2015) It defines for diagnosis of diabetes for actual world entity.The major point of this it detects the diabetes at early stage.In this we have read decision tree and naïve bayes are used for diagnosis of diabetes. And at the end can say that proposed model give the best and effective result.

III. PROBLEM FORMULATION

The expectation examination is the technique which can predict the future outcomes from the current data. When you have diabetes, you figure out to make sure you maintain glucose levels within your target range goals – not too high, not too low. That implies figuring out when and what you will eat for meals and snacks, when you will analyze blood sugar and how to fit in exercise. The prediction analysis techniques are based on the clustering and classification. In this, medical data is analyzed to predict the regional diseases. The data is collected from the UCI repository for predicted analysis. The SVM classifier is used to classify the data into certain number of classes. To analyze the diabetes, it is very hard to apply machine learning and data mining in every single research study. We will analyze different techniques and apply on the dataset. We will try to generate the efficient result. The existing improvement directly increase accuracy of classification and less execution time.

3.2 OBJECTIVES

1. The study of different prediction based algorithms and then analyzes it for data mining.
2. The proposed improvement will be based on SVM algorithm which calculate relation between attributes of the dataset
3. The existing and proposed algorithm will be compared in terms of time and accuracy.

4. RESEARCH METHODOLOGY

The prediction analysis is the technology which can predict the future possibilities from the existing dataThe steps are following:

Step 1: Input the diabetes data for classification.

Step 2: Pre-process data to remove missing data.

Step 3: It will check the condition if the data is classified then it will display the predicted result.

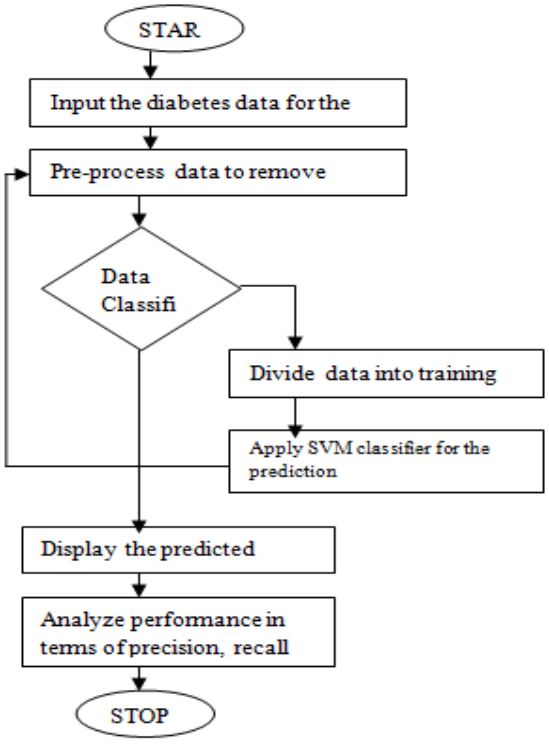
Step 4: If data is not classified then it will classified into training data and testing data.

Step 5: After this it will apply SVM classifier for prediction.

Step 6: Then it will pre-process the data for removing the missing data.

Step 7: Then it will display the predicted result.

Step 8: At the end, it will analyze the performance in terms of accuracy.



IV. RESULT AND DISCUSSION

We have used python for find the better accuracy.We have taken the dataset from the UCI repository.

 DOI : 10.35291/2454-9150.2019.0076

```
accuracy using SVM:  75.32467532467533 %
Predicted result for last 20 rows
     Pregnancies  Glucose  Age  Outcome  predict_svm
748            3      187   36        1            1
749            6      162   50        1            1
750            4      136   22        1            1
751            1      121   28        0            0
752            3      108   25        0            0
753            0      181   26        1            0
754            8      154   45        1            1
755            1      128   37        1            1
756            7      137   39        0            0
757            0      123   52        1            0
758            1      106   26        0            0
759            6      190   66        1            1
760            2       88   22        0            0
761            9      170   43        1            1
762            9       89   33        0            0
763           10      101   63        0            0
764            2      122   27        0            0
765            5      121   30        0            0
766            1      126   47        1            0
767            1       93   23        0            0
```

Fig.3. It shows the prediction of last twenty columns.

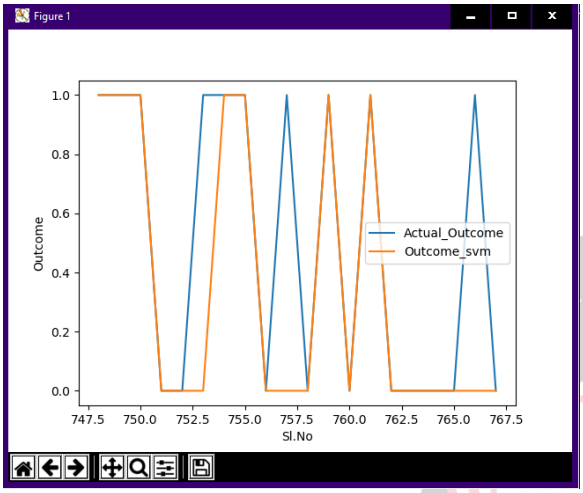


Fig.4.Graph of actual outcome and SVM outcome

It gives the accuracy of 75.3

V. CONCLUSION

A chronic disease that causes major causalities in the worldwide that is Diabetes. All 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. Therefore, data mining was used as it provides the optimal results and help in knowledge discovery from data. It is analyzed that diabetes can be predicted with classifications To implement prediction analysis, whole data is split into training and test sets. For the diabetes prediction , SVM classifier used. Certain parameters are used for the analysis of classifier.

VI. FUTURE SCOPE

The research study has only targeted patients with diabetes. Readmission prediction model has to be generated for other key health conditions also such as Heart disease, kidney disease etc. in Indian Healthcare system. Various other key features in the medical records, like family history (to find hereditary information), emotional status, socioeconomic status and lifestyle habits to be collected and analyzed. Living with diabetes is challenging and distressful. Diabetic patient's condition cannot be understood from medical charts. The discussion among specialist and patient can also be gathered and analyzed which could help to takeout important features corresponding to patient's willingness and gratitude by mining techniques. This data may improve the keen models to recognize patients at high danger of readmission.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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

[6] Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal (2014), "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.

[7] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "Weather Forecasting using Incremental K-means Clustering", vol. 8, 2014, pp. 142-147.

[8] Chew Li Sa, Bt Abang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, M. (2014), "Student performance analysis system (SPAS)", in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, vol.15, 2014, pp.1-6.

[9] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal 15 (2017) 104–116

[10] Bayu Adhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, "Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine", Vol. 11, issue 3, pp. 12-23, 2008.

[11] Zhiqiang Ge, Zhihuan Song, Steven X. Ding, Biao Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", 2017 IEEE. Translations and content mining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.

[12] 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.

[13] 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.

 DOI : 10.35291/2454-9150.2019.0076