

# A Proposed Methodology For Diabetes Prediction Using Data Mining

Sagar D. Soparkar<sup>1</sup>, Abhinay A Patil<sup>2</sup>, Shubham V. Narkhede<sup>3</sup>, Dhiraj S Mahajan<sup>4</sup>, Prof. B.B.Thakre<sup>5</sup>

<sup>1,2,3,4,5</sup> Dept of computer engineering

<sup>1,2,3,4,5</sup> Padm. Dr. V.B. Kolte College of Engineering, Malkapur

**Abstract-** Here we propose a system that allows users to get instant guidance on their health issues through an intelligent health care system online. The system is fed with various symptoms and the disease associated with those systems. Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is 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.

**Keywords-** DM, LDA, GP etc.

## I. INTRODUCTION

To analyze large amount of data, data mining technique is used. For each subfield of Clinical Predictions, and also presented how clinical data warehousing in combination with data mining can help administrative, clinical, research and educational aspects of Clinical Predictions. There are various challenges of data mining which can be discussed in health prediction.

There are various applications of data mining such as telecommunication industry, commercial industry, biological data analysis and many more. With the advance changes happening in the technology, especially in the field of health organization a lot of data is produced day by day. Since there is need of analysis of data and the amount of data analyzed is in large amount, so there is need of excessive knowledge regarding the technology of data mining. For health diagnosis E-healthcare applies data mining and telecommunication techniques. E-health was primarily used for patient data

analysis and disease diagnosis at various levels. There are some patients who require continuous checkup and might need doctor help immediately.

The purpose is to diagnose whether the person is affected by diabetes or not using K Nearest Neighbor classification technique. The diabetes dataset is taken as the training data and the details of the patient are taken as testing data.

Data mining is a subfield in the subject of software engineering. It is the methodical procedure of finding examples in huge data sets including techniques at the crossing point of manufactured intelligence, machine learning, insights, and database systems. The goal of the data mining methodology is to think data from a data set and change it into a reasonable structure for further use. Our examination concentrates 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.

Data mining is a significant tool in medical databases, which enhances the sensitivity and/or specificity of disease detection and diagnosis by opening a window of relatively better resources [4]. Applying machine learning and data mining methods in diabetes research is a pivotal way to utilizing plentiful available diabetes-related data for extracting knowledge. The severe social impact of the specific disease makes DM one of the main priorities in medical science research, which inevitably produces large amounts of data. Therefore, there is no doubt that machine learning and data mining approaches in DM are of great concern on diagnosis, management, and other related clinical administration aspects [5].

In order to achieve the best classification accuracy, abundant algorithms and diverse approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches, and association rule learning. Most noted among the aforementioned ones are the following: Calisir and Dogantekin proposed LDA-MWSVM, a system for diabetes diagnosis [6]. The system performs feature

extraction and reduction using the Linear Discriminant Analysis (LDA) method, followed by classification using the Morlet Wavelet Support Vector Machine (MWSVM) classifier. Gangji and Abadeh [7] presented an Ant Colony-based classification system to extract a set of fuzzy rules, named FCSANTMINER, for diabetes diagnosis. In [8], authors regard glucose prediction as a multivariate regression problem utilizing Support Vector Regression (SVR). Agarwal [9] utilized semi-automatically marked training sets to create phenotype models via machine learning methods. Ensemble approaches, which utilize multiple learning algorithms, have been confirmed to be an effective way of enhancing classification accuracy.

The current system is a manual and file based one, we realize that system. We are going to build must give the solutions for wastage of time and space which affect the efficiency of the daily activities performed at the hospital. In previous system there is no location tracker for patient and doctors. There is no any feedback system in existing system for taking a feedback from patient. If the patient requires an instant diagnosis on their disease then they have to go doctor but it is not possible to everyone to identify disease at home instantly. Today's health prediction system is so much time consuming.

## II. LITERATURE SURVEY

The health industry has been growing a lot from past few years. This technique has gained a lot of importance in medical areas. It has been calculated that a care hospital may generate five terabytes of data in the year. In our day to day life we have lot of other problems to deal with and we neglect our health problems. So in order to overcome such problem we have designed user friendly website which helps users to get diagnosed from their residence at any time. We also provide an option for booking an appointment with the doctor to discuss health related problems and get diagnosed properly. Sajida et al. in [1] discusses the role of Adaboost and Bagging ensemble machine learning methods using J48 decision tree as the basis for classifying the Diabetes Mellitus and patients as diabetic or non diabetic, based on diabetes risk factors. Results achieved after the experiment proves that, Adaboost machine learning ensemble technique outperforms well comparatively bagging as well as a J48 decision tree. Orabi et al. in [2] designed a system for diabetes prediction, whose main aim is the prediction of diabetes a candidate is suffering at a particular age. The proposed system is designed based on the concept of machine learning, by applying decision tree. Obtained results were satisfactory as the designed system works well in predicting the diabetes incidents at a particular age, with higher accuracy using Decision tree.

Pradhan et al in [3] used Genetic programming (GP) for the training and testing of the database for prediction of diabetes by employing Diabetes data set which is sourced from UCI repository. Results achieved using Genetic Programming gives optimal accuracy as compared to other implemented techniques. There can be significant improve in accuracy by taking less time for classifier generation. It proves to be useful for diabetes prediction at low cost. Rashid et al. in [4] designed a prediction model with two sub-modules to predict diabetes-chronic disease. ANN(Artificial Neural Network) is used in the first module and FBS (Fasting Blood Sugar) is used in the second module. Decision Tree (DT) is used to detect the symptoms of diabetes on patients health. Nongyao et al. in [5] applied an algorithm which classifies the risk of diabetes mellitus. To fulfill the objective author has employed four following renowned machine learning classification methods namely Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes. For improving the robustness of designed model Bagging and Boosting techniques are used. Experimentation results shows the Random Forest algorithm gives optimum results among all the algorithms employed.

M. Durairaj, V. Ranjani presented a paper that aims to make a detailed study report of different types of data mining applications in the healthcare sector and to reduce the complexity of the study of the healthcare data transactions. Also presents a comparative study of different data mining applications, techniques and different methodologies applied for extracting knowledge from database generated in the healthcare industry. Finally, the existing data mining techniques with data mining algorithms and its application tools which are more valuable for healthcare services are discussed in detail[1].

Divya Tomar and Sonali Agarwal have presented a brief introduction of data mining techniques such as classification, clustering, association, regression in health domain and their advantages and disadvantages. This survey also highlights applications, challenges and future issues of Data Mining in healthcare[2].

R.Karthiyayini, J.Jayaprakash have presented a paper which analyses the various results generated by implementing the Apriori algorithm of Association technique. The focus of this paper is to provide precise information about chronic diseases for public[3].

Priyanka Vijay Pawar, Megha Sakharam Walunj, and Pallavi Chitte presents a methodology to predict diseases based on user input symptoms. They have built a prototype to demonstrate the efficiency of these methods which will inform users about the disease they are suffering from. It predicts

probable diseases by mining data sets and provides suggested doctors and remedial solutions[4].

Gitanjali J, C. Ranichandra ,M.Pounambal has presented a method for identifying frequency of diseases in particular geographical location for a given period of time using Apriori data mining technique based on association rules is proposed[5].

Abdullah Saad Almalaise Alghamdi discussed about the importance of data mining using medical data then discussion of general data mining techniques has been presented. Furthermore, methodology describes the conceptual model for the extraction of rules on medical databases finally result can guide the relationship between the different attributes presented in the data. In this regard, they applied FP growth algorithm for extracting rules from the medical data[6].

### III. VARIOUS ISSUES OF DIABETES PREDICTION IN DATA MINING

Diabetes Mellitus is a chronic disease for which there is no known cure except in very specific situations management concentrates on keeping blood sugar levels as close to normal as possible without causing hypoglycemia.

This can be controlled with diet, exercise and use of appropriate medications. Diabetes Mellitus occurs throughout the world and it is more in developed countries. The increase in rates in developing countries follows the trend of urbanization and life style changes, including a “western-style” diet. This is because of less awareness.

The purpose of data mining is to extract useful information from large databases or data warehouses. Data mining applications are used for commercial and scientific sides [1].

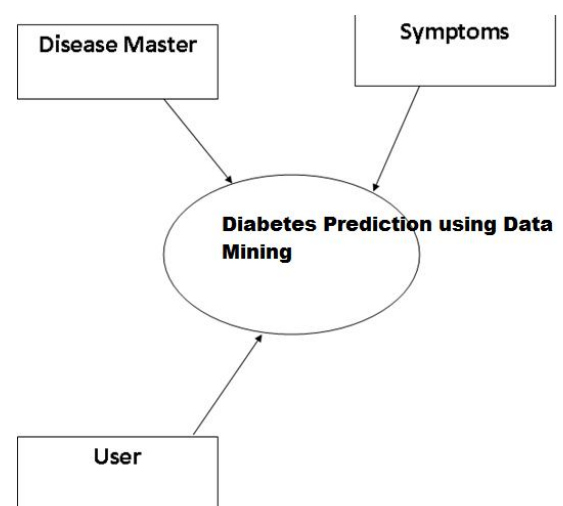
Data mining is process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [2].

KDD process may consists several steps: like data selection, data cleaning, data transformation, pattern searching i.e. data mining, finding presentation, finding interpretation and finding evaluation [3].

Diabetes is a chronic disease characterized by a long treatment cycle, numerous complications (e.g., kidney and eye

diseases), and recurrent illness. With advances in the informatization of medicine, medical industries with large amounts of complicated patient data are keen to extract information from this data to assist the development of these industries. Simultaneously, they also seek to be capable of alleviating the challenges faced by medical personnel, through the forthcoming development of smart medicine. The use of machine learning and other artificial intelligence methods for the analysis of medical data in order to assist diagnosis and treatment is one of the manifestations of smart medicine with the most practical significance.

With the improvement of the living standards of our people and the westernization of our diet, the incidence, mortality, and morbidity of diabetes have significantly increased and have a serious impact on our health. In 2006, Shang [1] made use of the survey data of Xinjiang chronic disease integrated prevention and control demonstration site in the New Urban District of Urumqi in 2004 and surveyed 2031 people over the age of 18 in three communities in the district. The results showed the relationship between diabetes and age and gender: the prevalence of male and female rose with age, because the decrease of glucose tolerance with age and the improvement of living standard are the reasons for the increased incidence. Overweight and obesity are one of the risk factors of diabetes mellitus. The survey found that the prevalence of diabetes in people with BMI>24 was 10.58%, the prevalence of diabetes in people with BMI≤24 was 4.31%, two groups prevalence by chi-square test was P <0.01, and there was a significant difference between the two groups, indicating that overweight and obese individuals are more susceptible to diabetes.



In 2009, Su [2] analyzed the related factors of diabetes in the New Urban District of Urumqi in Xinjiang. The results showed that age, gender, height, weight, and BMI

associated with diabetes were not statistically significant. However, the waist circumference, systolic blood pressure, and triglyceride are factors that are positively correlated with diabetes.



#### IV. CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases

#### REFERENCES

[1] Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. *International Journal of Engineering and Technology (IJET)* 5, 2903–2908.

- [2] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University - Computer and Information Sciences* 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.
- [3] Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. *International Journal of Computer Applications* 54, 21–25. doi:10.5120/8626-2492.
- [4] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. *Advances in Intelligent Systems and Computing* 1, 763–770. doi:10.1007/978-3-319-11933-5.
- [5] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: *Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016)*, pp. 451–455.
- [6] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE*. pp. 5–10.
- [7] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 476–491. doi:10.1109/34.589207.
- [8] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications* 09, 1–16. doi:10.4236/jilsa.2017.91001.
- [9] Garner, S.R., 1995. Weka: The Waikato Environment for Knowledge Analysis, in: *Proceedings of the New Zealand computer science research students conference*, Citeseer. pp. 57–64.
- [10] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree based diabetes prediction model, in: *International Conference on Advanced Software Engineering and Its Applications*, Springer. pp. 99–109.
- [11] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management*
- [12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. *Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology*