7th International Conference on Research Developments in Applied Science, Engineering & Management
Mahratta Chamber of Commerce, Industries and Agriculture, Tilak Road, Pune (India)
15th -16th June 2019 www.conferenceworld.in
ISBN : 978-93-87793-90-3

A STUDY OF COMPUTING METHODS PREDICTING DIABETES IN UCI BASED PIMA INDIANS DIABETES

[1]Ankit Srivastava, [2]Prof. (Dr.) Mohit Gangwar, [3]Asst. Prof. Manish Rai

[1]Mtech (CSE) ,Department of Computer Science and Engineering
RKDF College of Engineering Bhopal, India
[2]Dean Engineering, Bhabha University, Bhopal. India
[3]Department of Computer Science and Engineering
RKDF College of Engineering Bhopal, India

Abstract:

Diabetes is a Non-communicable disease (NCDs). NCDs also include some other diseases like as-stroke, heart disease, cancer and chronic lung cancer they together are responsible for almost 70% of the deaths worldwide. Diabetes mellitus Type II is most common in all NCDs diseases.

The diabetes dataset considered in this investigation may not consider some other significant components that are identified with gestational diabetes, as metabolic disorder, family history, propensity for smoking, sluggish schedules, some dietary designs and so on. Prediction of diabetes is very typical task in the last stages. There are many other environmental factors are responsible for it. There must be a diagnosis systems which are correctly predict for physicians to know whether a patient is diabetic or not.

Keywords: Data mining, PIDD, SVM, AI, ML

I.INTRODUCTION

Diabetes Mellitus has become a common health problem nowadays, which would affect people and lead to various disablements like cardio vascular disease, visual impairments, leg amputation and renal failure if diagnosis is not done in the right time [1]. Diabetes can influence individuals because of the absence of insulin in the blood. Insulin is a characteristic hormone emitted by the pancreas, which goes about as a key to open the body cells so sugar, starch and nourishment particles can be assimilated and henceforth be used by the cells to create vitality required for everyday life.

Diabetes is a major health concern affecting all age groups all over the world. Diabetes causes death, and may give rise to heart disease, blindness, kidney disease and other health related problems. It causes

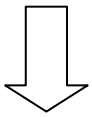
ISBN : 978-93-87793-90-3

mainly due to genetics or certain environmental conditions like obesity, lack of physical exercise, eating habits, unhealthy life style etc. However it can be controlled by proper control over diet and regular exercise, yoga etc [2]. Diabetes is a chronic disease in which body does not produce insulin or use it properly. This increase the risks of developing, kidney disease, blindness, nerve damage, blood vessel damage and contribute to heart disease [3]. There are two types of diabetes: one is type-1 diabetes-also called insulin dependent, which is usually diagnosed in children and juvenile; another is type-2 diabetes-which is often diagnosed in middle aged to elderly people. Patients with type-2 diabetes do not require insulin cure to remain alive, although up to 20% are treated with insulin to control blood glucose levels. It has been shown that 80% of type-2 diabetes complications can be prevented or delayed by early identification of people at risk [4].

In Type-1 diabetes:

The stages are:

- Normoglycemia
  (Normal glucose regulation)
- Hyperglycemia leads to :
  - Impaired Glucose Tolerance
    (Prediabetes)

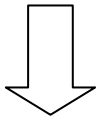


  - Diabetes Mellitus
    - Not insulin requiring.
    - Insulin requiring for control.
    - Insulin requiring for survival.
    -

Type-2 diabetes

It occurs due to combination of resistance to insulin action and deficient compensatory insulin secretory response. The stages are:

- Normoglycemia
  (Normal glucose regulation)
- Hyperglycemia leads to :
  - Impaired Glucose Tolerance
    (Prediabetes)



  - Diabetes Mellitus
    (Not insulin requiring)

The data set used in this paper is excerpted from the UCI Machine Learning Repository [5]. The original owner of this dataset is the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of this dataset from larger database, and all patients are females at least 21 years old of Pima Indian heritage.

Medical diagnostics is very troublesome and visual errand which is generally done by specialists' doctors. A specialist generally takes choices by assessing the present test aftereffects of a patient or the specialist doctor contrasts the patient and different patients with a similar condition by alluding to the past choices. Therefore, it is very difficult for

ISBN : 978-93-87793-90-3

a physician to diagnose hepatitis. For this reason, in recent times, many machine learning and data mining techniques have been considered to design automatic diagnosis system for diabetes.

General Symptoms of Diabetes:

1. Expanded thirst

2. Frequent urination

3. Loss of weight

4. Frequent hunger

5. Slow healing contamination

6. Obscured vision

7. Frequent vomiting

Diagnose test

1. Urine test

2. Fasting blood glucose level

3. Random blood glucose level

4. Oral glucose tolerance test

5. Glycosylated hemoglobin. [6]

A comparison analysis table is given which depict the past procedure creators, their techniques, advantage of their research work and limitation of their work. Further an improvement which can perform for better outcome is given in our proposed research.

II. LITERATURE REVIEW

There have been extensive studies of this dataset in the Machine Learning Literature. Various classification algorithms have been applied to the data set, and no algorithm performs exceptionally well.

In this paper [7] Diabetes mellitus is one of the most serious health challenges facing American Natives in the United States today. The publicly available Pima Indian diabetic database (PIDD) at the UCI Machine Learning Lab has become a standard for testing data mining algorithms to see their accuracy in predicting diabetic status from the 8 variables given. In this study we will try to predict the presence of diabetes based on ensemble of SVM and BP NN. The predictive accuracy was 88.04 which was the best accuracy and it was very promising with regard to the other classification systems in the literature for this problem.

In this paper [8] Diabetes is one of the leading causes of death, disability and economic loss throughout the world. Type 2 diabetes is more common (90-95% worldwide) type of diabetes. However, it can be prevented or delayed by taking the right care and interventions which indeed an early diagnosis. There has been much advancement in the field of various Machines learning algorithms specifically for medical diagnosis. But due to partially complete medical data sets, accuracy often decreases, results in more number of misclassification that can lead to harmful complications. An accurate prediction and

ISBN : 978-93-87793-90-3

diagnosis of a disease becomes a challenging research problem for many researchers. Therefore, aimed to improve the diagnosis accuracy we have proposed a new methodology, based on novel preprocessing techniques, and K-nearest neighbor classifier. The effectiveness of the proposed methodology is validated with the help of various quantitative metrics and a comparative analysis, with previously reported studies using the same UCI dataset focusing on pima-diabetes disease diagnosis. This is the first work of its kind, where 100% classification accuracy is achieved by feature reduction from eight to two that shows the out performance of the proposed methodology over existing methods.

In this paper [9] Parashar A. et al. (2014) have proposed Linear Discriminant Investigation and Support Vector Machine for the conclusion of Pima Indians Diabetes dataset, where LDA diminishes include subsets and SVM is capable to classify the data. They have likewise contrasted SVM and feed forward neural system (FFNN) yet our proposed SVM+LDA gives better order precision as 77.60% with 2 features. In this paper

[10] Healthcare industry contains very large and sensitive data and needs to be handled very carefully. Diabetes Mellitus is oneof the growing extremely fatal diseases all over the world. Medical professionals want a reliable prediction system to diagnose Diabetes. Different machine learning techniques are useful for examining the data from diverse perspectives and synopsizing it into valuable information. The accessibility and availability of huge amounts of data will be able to provide us useful knowledge if certain data mining techniques are applied on it. The main goal is to determine new patterns and then to interpret these patterns to deliver significant and useful information for the users. Diabetes contributes to heart disease, kidney disease, nerve damage and blindness. So mining the diabetes data in efficient way is a crucial concern. The data mining techniques and methods will be discovered to find the appropriate approaches and techniques for efficient classification of Diabetes dataset and in extracting valuable patterns. In this study a medical bioinformatics analyses has been accomplished to predict the diabetes. The Pima Indian diabetes database was acquired from UCI repository used for analysis. The dataset was studied and analyzed to build effective model that predict and diagnoses the diabetes disease. In this study we aim to apply the bootstrapping resampling technique to enhance the accuracy and then applying Naïve Bayes, Decision Tree and k Nearest Neighbors (kNN) and compare their performance.

ISBN : 978-93-87793-90-3

Table 1 -In diabetic disease predication there were different previously performed techniques. Which of some other common or major findings are given as follow .

Sn.	Author's	Methodologies	Finding
1	Tao et al.[11]	KNN,Naïve Bayes, Decision Tree, Random Forest, SVM and Logistic Regression	Concentrated on the accuracy of recall and got better result. Filtering criteria can be improved
2	Loannis et al.[12]	Naïve Bayes, Logistic regression ,and Svm	From the three algorithm Svm provided high accuracy of 84%
3	WeifengXu et al.[13]	NaïvBayes,Randomforest,Adaboost	Random forest classifier method better relative to other .in contrast ID3 provided the least accuracy than others.
4	Yunsheng et al. [14]	DISKR and KNN	An attribute which have less factor should be eliminated. Accuracy increase can be increase by removing outliers. Space complexity decreased.
5	Messan et al.[15]	GMM, ELM , ANN LR, and SVM	Fewer amounts of sample data used. Comparison of algorithm were done from those method artificial neural network provide better accuracy than other classifier.

ISBN : 978-93-87793-90-3

6	Ramiro et al.[16]	Fuzzy rule	Wrong treatment was reduced using fuzzy rule and recommendation system was developed for doctor
7	Swarupa et al.[17]	KNN,J48, ANN,zeroR, NB	CV parameter selection, Filtered classifier and simple cart Various dataset applied containing diabetes dataset. Cross validation not applied. NB shown high accuracy by providing accuracy of 77.01%.
8	Pradeep & Dr.Naveen [18]	Decision tree(J48)	J48 is noted as good accuracy provider algorithm. Feature selection has high role in the prediction area
9	Sajida et al.[19]	Adaboost, j48,and Bagging	Adaboost was shown improved accuracy than other method. International Journal of Pure and Applied Mathematics Special Issue 872
10	Santhanam and Padmavathi[20]	K-means with Genetic Algorithm ,and SVM	The integrated clustering and classification of algorithm done and provided better performance.

III.CONCLUSION

Data mining and classification is a significant idea while managing the substantial mismatch dataset. Data mining assumes a significant job in different fields such as artificial intelligence (AI) and machine learning (ML), statistics and database systems. The center goal of this examine is to improve the precision of prescient model. The exactness can be increment by improving the presentation of the

information, the algorithms or even by algorithm tuning.

In this paper survey approach over different mechanism is performed. Those techniques utilized by previous researcher. Thus the study shows that previous data can be more useful for diagnosis. Hence an efficient data mining approach is helpful to determine the disease over a patient. As the paper describe and work towards diabetes mellitus diseases which means body's ability to produce or respond to the hormone insulin is impaired, discussed here. Data mining techniques help in understanding and refining PIMA dataset.

III..REFERENCES

[1]World Health Organization, http://www.who.in/topics/diabetes_mellitius/en

[2] Hasan Temurtas, Nejat Yumusak and Feyzullah Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", "Expert Systems with Applications", 36,2009,8610-8615.

[3]R. Bellazzi, "Telemedicine and diabetes management: Current challenges and future research directions," J. Diabetes Sci. Technol., vol. 2, no. 1, pp. 98–104, 2008

[4]J.C.Pickup, G. Williams,(Eds), Textbook of diabetes, Blackwell Science, Oxford

[5] C L Blake, C J Merz. UCI repository of machine learning databases University of California, Irvine, Department of Information and Computer Sciences. 1998

[6]http://www.diabetes.ca/about-diabetes/types-of-diabetes.

[7] Rahmat Zolfaghari (2012). Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM. IJCEM International Journal of Computational Engineering & Management, Vol. 15 Issue 4, July 2012 ISSN (Online): 2230-7893 www.IJCEM.org

[8] Madhuri Panwar, Amit Acharyya, Rishad A. Shafik and Dwaipayan Biswas (2016). K-Nearest Neighbor Based Methodology for Accurate Diagnosis of Diabetes Mellitus. 2016 Sixth International Symposium on Embedded Computing and System Design (ISED) 978-1-5090-2541-1/16/$31.00 © 2016 IEEE

[9] Parashar A., Burse K., Rawat K. (2014). A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4, pp. 378-383, ISSN: 2277 128X.

[10] Uswa Ali Zia, Dr. Naeem Khan (2017). Predicting Diabetes in Medical Datasets Using Machine Learning Techniques. International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017 ISSN 2229-5518.

[11] Zheng, Tao, Wei Xie, LilingXu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International

ISBN : 978-93-87793-90-3

journal of medical informatics 97 (2017): 120-127.

[12] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).

[13] Xu, Weifeng, Jianxin Zhang, Qiang Zhang, and Xiaopeng Wei. "Risk prediction of type II diabetes based on random forest model." In Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on, pp. 382-386. IEEE, 2017.

[14] Song, Yunsheng, Jiye Liang, Jing Lu, and Xingwang Zhao. "An efficient instance selection algorithm for k nearest neighbor regression." Neurocomputing 251 (2017): 26-34.

[15] Komi, Messan, Jun Li, YongxinZhai, and Xianguo Zhang. "Application of data mining methods in diabetes prediction." In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on, pp. 1006-1010. IEEE, 2017.

[16] Meza-Palacios, Ramiro, Alberto A. Aguilar-Lasserre, Enrique L. Ureña-Bogarín, Carlos F. Vázquez-Rodríguez, Rubén Posada-Gómez, and Armín Trujillo-Mata. "Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus." Expert Systems with Applications 72 (2017): 335-343.

[17] Rani, A. Swarupa, and S. Jyothi. "Performance analysis of classification algorithms under different datasets." In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pp. 1584-1589. IEEE, 2016.

[18] Pradeep, K. R. and N. C. Naveen. "Predictive analysis of diabetes using J48 algorithm of classification techniques." In Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on, pp. 347-352. IEEE, 2016.

[19] Perveen, Sajida, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

[20] Santhanam, T., and M. S. Padmavathi. "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis." Procedia Computer Science 47 (2015): 76-83.