International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 3| ISSN: 2456-3307

A Study of Performance Analysis and Shortcoming of Data Mining Models to Predict Diabetes

Ravinder Singh[1], Dr. Devender Kumar[2]

[1]Research Scholar, Department of Computer Science & Applications, Baba Mastnath University, Asthal Bohar, Rohtak, Haryana, India

[2]Associate Professor, Department of Computer Science & Applications, Baba Mastnath University, Asthal Bohar, Rohtak, Haryana, India

ABSTRACT

Since past few years data mining lot of attention related to knowledge like extracting methods in health care system like diabetes, cancer, CVS etc. There are lot of technique of data mining like decision tree, Naive base, KNN; J48 etc. are being used for prediction of diabetes. Diabetes is metabolic disorder related to poor absorption of insulin into body mussels or poor lowered secretion of insulin from pancreases. As this disease, this is main death causes disease in the world. So, prediction of these diseases with the help of data mining technique may help to protect many lives. In this study, we have to discuss various data mining technique, types of diabetes, application of these data mining technique.Prediction of diabetes or any other disease could play a significant role in health system. Data mining are very useful in the scenario. These techniques help in selection, understanding and designing of large size data to analysis the chances of diseases occurrence. Recently who has announced diseases a major cause of death worldwide. The prediction and identification early stage of diabetes can play major role to treat this disease significantly. Various data mining techniques like KNN, Decision tree, Naïve Bays etc. would be a significant asset for the researcher for gaining various data about diabetes, its causes, symptoms and possible treatment that have been using in the past and currently used by various physician. In this study we have briefly discussed various data mining techniques/models. Which have been currently used for diabetes prediction? Along with this discussion, we have also focused on performance and short coming of existing models/techniques time to time evaluated by researchers.

Keywords- Diabetes, predictive analytics, KNN, DT, insulin

I. INTRODUCTION

Data mining tools which helps in collecting the data from a large size of data set. It helps in analysing the large size sample by use of statically models and AI. This tool used for prediction of possible pattern's or trends in a particular set of data. The various techniques includes in data mining are:

To analysis various diseases in health care industry, data mining tools are being used for collecting, extracting, organization, analysing and utilization of all the data for the prediction of various life threating diseases. These tools required data in the form of structured or generalized set which could be easily analysed these analytical tools. The source of various data related to diseases with data mining techniques could play a significant role for diagnosis and prevention of different diseases.

At present diabetes is a prime disease for that all over the world. In this condition body is unable to produce and utilize the insulin. It is of two types. Type I, in which body is insulin depended. In this disease the immune system of body attack beta cell Langerhans present in pen crease. This is directly affecting the production of insulin with in human body. Approximately 90% of diabetes patient are suffered from type II diabetes. In which insulin production is normal but human mussel system become insensitive. In which Glucose Metabolism required for utilization of insulin. Type II diabetes is called non-insulin depended diabetes. It can be treated by uses different entity per glycaemia drugs. A rare but generally found in type III diabetes. This is seen in mainly during pregnancy time due to increased glycaemic level. This diabetes may lead to renal complication, cardiac diseases etc. By collecting all the data related to these, various types of diabetes and processing them with the help of data mining techniques may reduce chances of death due to diabetes. Data mining technique arrange, classified and analysis, these data according to the statically and AI methods. Various researchers in the health care field applies various statically tool to develop for analysis and refining these data set.

II. Predictive Analysis Techniques

A. Machine learning techniques

This technique deals with scientific field. It's improved the performance using data. In this technique computer program learn automatically on the base of AI.

1. **Supervised learning:**

This algorithm learns particularly for apurpose. In this all data are labelled and predict output from the training dataset.

2. **Unsupervised learning:**

This learning technique is used for clustering-based algorithm and especially data is un-labelled. It is determined unseen arrangement of records among variables.

B. Data mining techniques/models

Data mining referred to as "Knowledge mining from Data". It is helpful for finding the useful pattern or information and relationship among the huge amount of data. It has different technique as: -Classification, Association, Clustering etc.

C. Nature-inspired algorithm

Nature is a guide for human being from past era. It is significant for humans, animal and birds too. This algorithm is inspired from nature. Number of algorithms inspired from nature for optimization or search of any related problem.

It has mostly two types: -

1. Evolutionary Algorithm: - Genetic Algo.
2. Swarm Algorithm: - PSO Algo. and Ant Algo. Etc.

TABLE 1: Review the performance and shortcoming of data mining techniques for prediction of Diabetes

Sr. No.	Methodology	Dataset	Performance	Shortcoming	References
1	BL_WS mote Algo.	Clinical Data Set	The model accuracy is a high with 92%	Comparisons between the whole attributes and the important attributes	[1]

2	Neural Network & Decision Tree Model (mRMR)	Uzhow and PIMA Indian Dataset	Uzhow dataset is 0.8084 and PIMA Indians is 0.7721	Taking the suitable attributes	[2]
3	Firefly and Cuckoo Search Algorithms	Pima Indian Diabetes	Cuckoo-Fuzzy Accuracy- 81%	▶ Change no. of featuresused to train. ▶ Used nature-inspired algorithm.	[3]
4	NN, GA, Artificial Bee Colony algorithms	Pima dataset	Artificial Bee Colony-Accuracy-98.38%	For more optimal in classification problem arise in to change the indexes of weights	[4]
5	A Feed forward NN Algo.	Pima dataset	Network Model Accuracy value over 80%	For better prediction, required the Improve training method and changing activation function	[5]
6	Proposed and K-means Clustering Approach	Pima Indian Diabetes	Proposed System Accuracy-98.7%	More clusters are required for the best separation points and correct results	[6]
7	Elman NN Model	Pima Indian Diabetes	Classification Accuracy-95.7%	Required more function for better performance	[7]
8	Machine Learning Algorithm	UCI machine learning	Proposed Ensemble Method(PEM) Accuracy-90.36%	Required large amount of data	[8]
9	Decision tree and SMOTE Algorithm	Clinical dataset	Accuracy-94.7013%	To reduction in the class imbalance	[9]
10	NN Method Combined with Colonial Competition Optimization, Radial Base FunctionAlgorithm	Re-hospitalization of diabetic patients	Risk of re-hospitalization NN optimization Algorithm.	Use different algorithm like SVM,NN, Genetic, Evolutionary Algorithms	[10]
11	NN usingDragonfly Algorithm	Digestive and Kidney and National Institute of Diabetes	Enhance the accuracy and speed of classification	Comparisons for batter result	[11]
12	DM	Pima Dataset	Higher accuracy prediction of Diabetes.	Other refined procedure implemented for the study of DM	[12]
13	Genetic, Naïve Bayes and KNN Algorithm	Pima dataset	Accuracy-83.12%	Required more efficient techniques for better result.	[13]

14	Bat Optimization classification algorithm	PID dataset	Accuracy-73.91%	Used other algorithm for the improved accuracy	[14]
15	A and RBF NN algorithm	Pima Indian Dataset	Accuracy RBF NN-76.087% GA_RBF NN – 77.3913%	Can be use another kinds of diseases	[15]
16	Enhanced Class Outlier with Automatic Multilayer Perceptron	UCI (Pima Indians)	Accuracy-88.7%	For superior classification design and effective system	[16]
17	Deep Learning Approach	Online Machine Learning UCI Repository.	Performance is better than Shallow network	Result depend on size ofthe data set	[17]
18	NN with Genetic Algorithm	Pima Indians Diabetes Data Set	Proposed Approach with NN-86.5% With Genetic Approach-87.46%	Used other algorithm for better accuracy	[18]
19	K- means and Decision Tree	Pima Indian Diabetes Data	Accuracy-90.40%	Huge amount of data medical statics for the effective result	[19]
20	Evolutionary Algorithm	Pima Indian Diabetes dataset	Better Performance execution time & accuracy	For expand the work can use evolutionary search	[20]
21	SVM, KNN, Naïve Bayes, ANN algorithm	Diverse section of thesociety (More than 400)	Ensemble based Diabetes Diagnose-98.60%	Improve number of verity in the database	[21]
22	K-SVM algorithm	UCI Pima Indian dataset	Accuracy-99.74%	For potential Enhancement required to integrate one of the optimization techniques	[22]
23	Genetic algorithm MOE fuzzy classification	Pima dataset	Accuracy-83.0435%	Analysed and addressed missing data for the feature selection	[23]
24	ML with DM process	Electronic Health Records	SVM provide successful result	Use more medical statistics	[24]

25	KNN, Multilayer perceptron NN, Binary logistic regression	Multi-dimensional healthcare dataset	KNN higher accuracy	Use other data set and accuracy	[25]
26	SVM Algo.	Pima dataset	Accuracy-90.2%	Accuracy may be improved	[26]
27	MPSO-NN Algo.	Pima dataset	Accuracy-81.8%	Increased the number of instances	[27]
28	General Regression Neural Networks (GRNN)	Pima Indian Dataset	Better optimization Reduce complexity and computational cost	Use a better technique withsame dataset and features for better accuracy.	[28]
29	Data mining technique	Laboratories data	Improve accuracy	Work on accuracy	[29]
30	Firefly Algorithm, BAT Algorithm	Sree Diabetic Care Center	Improve various statically measure	Applying other techniques for increase the accuracy	[30]
31	With Chaotic Levy Flights	Pima data set	Flexibility of proposed system	This algorithm tested on other chronic diseases	[31]
32	GA and K-means	Pima Dataset	Accuracy-96.7%	Works on missing data	[32]

III. CONCLUSION

In this paper, works on different data mining techniques and models with performance and shortcoming of these models/techniques. These techniques work with different variable. Every model shows the different accuracy with particular variables. There is no one models/techniques available which are perfect in prediction the diabetes. Every model/technique has performance with shortcoming. These models/techniques provide the result with higher accuracy with the selection of attributes and dataset. In this situation feature selection method can be applied with the model/technique for higher accuracy and efficiency for prediction the diabetes.

IV. REFERENCES

[1]. Lei X. M. and Feng C., (2018) "The Establishment of Diabetes Diet Classification Model Based on BL_WSmote" DSIT2018, July 20–22, Singapore, Singapore, ACM.

[2]. Zou Q., Qu K., Luo Y., Dehui Yin, Ju Y.and Tang H., (2018) "Predicting Diabetes Mellitus With Machine Learning Techniques" Frontiers in Genetics, November,(9), 1-10.

[3]. Haritha, R., Babu, D. S., &Sammulal, P. (2018). "A Hybrid Approach for Prediction of Type-1 and Type-2 Diabetes using Firefly and Cuckoo Search Algorithms", International Journal of Applied Engineering Research, 13(2), 896-907.

[4]. Rashid, T. A., & Abdullah, S. M. (2018). "A Hybrid of Artificial Bee Colony, Genetic Algorithm, and Neural Network for Diabetic Mellitus Diagnosing" Aro-THE Scientific JOURNAL of Koya University, 6(1), 55-64.

[5]. Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. (2018) "A Feed- Forward Neural Network Model For The Accurate Prediction Of Diabetes Mellitus". INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, 7(8), 151-155.

[6]. Kadhm, M. S., Ghindawi, I. W., & Mhawi, D. E. (2018). "An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach", International Journal of Applied Engineering Research, 13(6), 4038-4041.

[7]. Sundaram, N. M. (2018). "An Improved Elman Neural Network Classifier for classification of Medical Data for Diagnosis of Diabetes", International Journal of Engineering Science, 8(3).16317-16321.

[8]. Alehegn, M., &Mulay, R. J. D. P. (2018) "Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm". International Journal of Pure and Applied Mathematics, 18(9), 871-878.

[9]. Mirza1S, Mittal S. & Zaman M. (2018). "Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree", International Journal of Applied Engineering Research, (13). 9277-9282.

[10]. Habibi, N., &Harouni, M. (2018), "Estimation of Re- hospitalization Risk of Diabetic Patients based on Radial Base Function (RBF) Neural Network Method Combined with Colonial Competition Optimization Algorithm". Majlesi Journal of Electrical Engineering, 12(1), 109-116.

[11]. Yasen, M., Al-Madi, N., Obeid, N., Sumaya, P., & Abdullah II, K. (2018), "Optimizing Neural Networks using Dragonfly Algorithm for Medical Prediction". Conference paper.

[12]. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). "Type 2 diabetes mellitus prediction model based on data mining". Informatics in Medicine Unlocked, 10, 100-107.

[13]. Patil, R. N., &Tamane, S. C. (2018). "Upgrading the Performance of KNN and Naïve Bayes in Diabetes Detection with Genetic Algorithm for Feature Selection". International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(1), 2456-3307.

[14]. Edlaa D.R. & Cherukua R. (2017) "Diabetes-Finder: A Bat Optimized Classification System for Type-2 Diabetes". 7th International Conference on Advances in Computing & Communications, ICACC 2017, Cochin, India 235-242.

[15]. Choubey, D. K., & Paul, S. (2017), "GA_RBF NN: a classification system for diabetes". International Journal of Biomedical Engineering and Technology, 23(1), 71-93.

[16]. Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., & Nawaz, R. (2017). "ECO-AMLP: A Decision Support System using an Enhanced Class Outlier with Automatic Multilayer Perceptron for Diabetes Prediction". arXiv preprint arXiv:1706.07679. 1-19.

[17]. Ramesh, S., Caytiles, R. D., & Iyengar, N. C. S. (2017). "A Deep Learning Approach to Identify Diabetes". Advanced Science and Technology Letters, 145, 44-49.

[18]. Dadgar, S. M. H., &Kaardaan, M. (2017), "A Hybrid Method of Feature Selection and Neural Network with Genetic Algorithm to Predict Diabetes", International Journal of Mechatronics, Electrial, Computer Technology, 7(24), 3397-3404.

[19]. Chen, W., Chen, S., Zhang, H., & Wu, T. (2017). "A hybrid prediction model for type 2 diabetes using K-means and decision tree". In Software Engineering and Service Science (ICSESS), 2017 8th IEEE International Conference on (pp. 386-390). IEEE.

[20]. Chandan Banerjee, Sayak Paul, Moinak Ghoshal (2017), "An Evolutionary Algorithm based Parameter Estimation using Pima Indians Diabetes Dataset", International Journal on Recent and Innovation Trends in Computing and Communication, 5(6), 374-377.

[21]. Sethi, H., Goraya, A., & Sharma, V. (2017). "Artificial Intelligence based Ensemble Model for Diagnosis of Diabetes". International Journal of Advanced Research in Computer Science, 8(5). 1540-1548.

[22]. Osman, A. H., & Aljahdali, H. M. (2017). "Diabetes disease diagnosis method based on feature extraction using K-SVM". International Journal of Advanced Computer Science and Applications, 8(1), 236-244.

[23]. Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., &Nalluri, S. (2017). "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset". In Computing Networking and Informatics (ICCNI), International Conference IEEE,1- 5.

[24]. 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 journal, 15, 104-116.

[25]. Selvakumar, S., Kannan, K. S., &GothaiNachiyar, S. (2017). "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques". International Journal of Statistics and Systems, 12(2), 183-188.

[26]. Tambade, S., Somvanshi, M., Chavan, P., & Shinde, S. (2017). "SVM based Diabetic Classification and Hospital Recommendation". International Journal of Computer Applications, 167(1), 40-43.

[27]. Ateeq, K., & Ganapathy, G. (2017). "The novel hybrid Modified Particle Swarm Optimization–Neural Network (MPSO-NN) Algorithm for classifying the Diabetes". International Journal of Computational Intelligence Research, 13(4), 595-614.

[28]. Alby, S., &Shivakumar, B. L. (2016). "A Novel Approach for Prediction of Type 2 Diabetes". International Journal of Advanced Research in Computer Science, 7(4).22-28.

[29]. Shetty, S. P., & Joshi, S. (2016). "A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique". I.J. Information Technology and Computer Science, 11, 26-32.

[30]. Thippa Reddy, G., &Khare, N. (2016). "FFBAT-optimized rule based fuzzy logic classifier fordiabetes". In International Journal of Engineering Research in Africa. 24, 37-152.

[31]. Soliman, O. S., &Elhamd, E. A. (2015). "A chaotic levy flights bat algorithm for diagnosing diabetes mellitus". International Journal of Computer Applications, 2(2), 56-63.

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

Cite this Article

Ravinder Singh, Dr. Devender Kumar, "A Study of Performance Analysis and Shortcoming of Data Mining Models to Predict Diabetes ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5, Issue 3, pp.680-686, May-June-2019.