IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Literature Survey On Different Techniques Used For Predicting Diabetes Mellitus

[1]Reshma R,[2]Dr.Anjana S Chandran

[1]Student,[2]Assistant Professor

[1]Department of Master of Computer Application

[1]SCMS School of Technology and Management ,Muttom,Aluva,India

Abstract: In today's world diabetes is the major health challenges in India. It is a group of a syndrome that results in too much sugar in the blood. It is a protracted condition that affects the way the body mechanizes the blood sugar. Prevention and prediction of diabetes mellitus is increasingly gaining interest in medical sciences. The aim of this paper is to conduct a survey on different techniques that are used for predicting diabetes.

Index Terms **- Diabetes, Machine learning, data mining, Multiperceptron, K-Nearest Neighbors ,Logistic Regression, Random forest,Random tree.**

I. INTRODUCTION

Diabetes Mellitus (DM) is commonly referred as Diabetes; it is a common chronic disease and poses a great threat to human health. Diabetes is a long lasting disease and its affects people worldwide. It happens when the body is not capable of producing enough insulin. Insulin which is secreted by pancreas, one of the most important hormones in the body, which is needed to maintain the level of glucose. It may produce the symptoms of frequent urination, increased thirst and hunger. Diabetes can be controlled with the help of insulin injections, a healthy diet and regular exercise. Diabetes can also leads to other disease such as blindness, blood pressure, heart disease, and kidney disease etc[1]. There are four types of diabetes.

Type 1 Diabetes: - Insulin isn't delivered sufficiently by pancrers, then type-1 diabetes may take place in body. It may arise any stage of life. e.g.: kids, youngsters [2].
Type 2 Diabetes: - Insulin doesn't in adequate amount and it is not sufficient for body need, then type-2 diabetes arises. Because of parent's inheritance, seniority, corpulence expands the danger of getting type 2 diabetes. For the most part happens at 40 years old.
Gestational Diabetes: -It is the third principle shape, significantly happens with the pregnant ladies because of abundance glucose equal level in the body.
Pregestational Diabetes: It is another form of diabetes and it occurs when insulin-subordinate diabetes earlier getting to be pregnant [2].

II. LITERATURE REVIEW

This section of the paper is dedicated to some of the research work done in the field of medical diagnosis using machine learning and data mining techniques.

Sonu Kumari and Archana Singh proposed [3] an intelligent and effective methodology for the automated detection of Diabetes Mellitus using Neural Network. The paper [4] approached the aim of diagnoses by using ANNs and demonstrated the need for preprocessing and replacing missing values in the dataset being considered. Through the Modified training set, a better accuracy was achieved with lesser time required for training the set. Sajida[8] by using CPCSSN(Canadian primary care sentinel surveillance Network) dataset and three machine learning methods to predict the diabetes Disses (DD) in early stage to safe human life at from early death .On this study Bagging ,Adaboost,and decision tree(J48) were used to predict the diabetes and the researcher was compare the result of those methods and concluded that Adaboost method was provide effective and better accuracy than the other methods in weka data mining tools. Sadri [20] used Naive Bayes, RBF Network and J48 datamining algorithms for diagnosing type II diabetes. They used WEKA tool. Finally they found Naive Bayes, having the accuracy rate of 76.96% than other algorithms. In this paper[27],Prediction of diabetes is done using ensemble voting classifiers for pima Indian diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 80% and 81% is achieved for data set by using 10-fold cross validation and by spitting data into 30% testing and 70% training. J. Pradeep Kandhasamy, S. Balamurali [58] this research study compare the performance of algorithms those are used to predict diabetes using data mining techniques. Also authors classifiers J48 Decision Tree, KNearest Neighbors, and Random Forest, Support Vector Machines to classify patients with diabetes mellitus. Authors compared four prediction models for predicting diabetes mellitus using 8 important attributes under two different situations. One is before pre-processing the dataset. Here the studies conclude that the decision tree J48 classifier achieves higher accuracy of 73.82 % than other three classifiers. After pre-processing, the dataset given more accurate

result when compared to the previous studies. In this case, both KNN (k=1) and Random Forest performance much better than the other three classifiers and they provide 100% accuracy. From this we can come to know that after removing the noisy data from our dataset it will provide good result for our problem.

This paper[5] shows how the Data mining classification algorithms say Naïve Bayes, Logistic Regression, C5.0, SVM and ANN are used to model actual Prediction of Diabetes Mellitus and a comparative analysis are made between them by making use of their Metric Measures say Accuracy, Precision, Sensitivity, Specificity and F1 Score. As a results of the research work, the C5.0 and Logistic Regression are equally good based on their Accuracy measures.Rahul and Minyechil Alehegn [7]studied various data mining techniques and its application . Application of machine learning algorithm were applied in different medical data sets. Single algorithm provided less accuracy than ensemble one.In most study decision tree provided high accuracy. In this study hybrid system Weka and java are the tools to predict diabetes dataset. This paper[18] is about various supervised classifier machine learning algorithms that were applied onto the training set which was obtained by eliminating attributes that did not have much context towards predicting diabetes. This was done using the chi-squared test and only that attributes which were ranked highest and was given more weightage and more likely to predict the onset of diabetes was considered. It was seen that on this training set the Neural Networks algorithm provided the most accurate results. The paper[44] analyses about the three types of diabetes and their causes. It also uses the prediction, classification technique. This provides the higher accuracy for the disease prediction. The research paper[45] explores about various Data mining algorithm approaches of data mining that have been utilized for diabetic disease prediction. In this paper Classification and Naive Bayes is one of the most used algorithms for the prediction of disease .

Pradeep & Dr.Naveen [10] in this paper, the performance of machine learning techniques were compared and measured based on their accuracy. The accuracy of the technique is vary from before pre-processing and after pre-processing as they identified on this study. This indicates the in the prediction of diseases the pre-processing of data set has its own impact on on the performance and accuracy of the prediction. Song [6]describe and explain different classification Algorithms using different parameters such as Glucose, Blood Pressure, Skin Thickness, insulin, BMI, Diabetes Pedigree, and age. The researches were not included pregnancy parameter to predict diabetes disease (DD). In this research, the researchers were using only small sample data for prediction of Diabetes. The algorithms were used by this paper were five different algorithms GMM, ANN, SVM, EM, and Logistic regression. Finally. The researchers conclude that ANN (Artificial Neural Network) was providing High accuracy for prediction of Diabetes. P. Chen[19] in their work have performed statistical testing on medical measurement index results of both patients with diabetes and without diabetes. They have further used boosting algorithms to give excellent classification of diabetes model based on the given medical data. In this study [39]a medical bioinformatics analyses have been accomplished to predict the diabetes. The WEKA software was employed as mining tool for diagnosing 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 Trees and (KNN) and compare their performance.

Yunsheng[9] in his study was the new approach that used KNN algorithm by removing the outlier/OOB(out of bag) using DISKR(decrease the size of the training set for K-nearest neighbour .and also in this study the storage space was minimized. There for ,the space complexity is become less and efficient after removing a parameters or instances which have less effect or factor the researchers got better accuracy. V. Kumar and L. Velide,[21] used Data mining Approach for Prediction and Treatment Of diabetes Disease. The techniques they used as Naïve Bayes, JRip, J48 (4.5), DT, NN .They used WEKA tool for implementation. They got 68.5% of accuracy level for J48 algorithm. The research paper [46] elaborates about detailed review of existing data mining methods used for prediction of diabetes. It also gives about the types of diabetes disease Type1, type2, and type3.The aim of the diabetes is to predict the diabetes with the help of Data mining methods such as the K-Nearest Neighbor Algorithm, Bayesian Classifier, Naive Bayesian Classifier, Bayesian Network, all the methods are used for prediction of diabetes. This paper also mentions about the effects of diabetes on patients . In this paper[54,]the proposed methodology aims at providing an efficient hybrid classification framework for predicting and monitoring the Diabetes disease. The main aim of this research is to identify and construct models that would assist medical practitioners in an efficient way by the way benefiting the people to attain longer life in this world.

Monika and Pooja [11] have discussed about previous data processing techniques to retrieve information and current development in the research of medical sciences. Further we have elaborated terminologies and techniques of learning in data mining and machine learning. In this paper, D. Jeevanandhini , E. Gokul Raj ,V. Dinesh Kumar, N. Sasipriyaa [12] conducted performance analysis for type2 diabetes mellitus dataset to improve the accuracy by using clustering and classification algorithm .Here they compared the four prediction model using 8 important attributes .From this studies concludes that Support Vector Machine (SVM) classifier achieves higher accuracy of 77.82 % than other three classifiers. Dr. K. Thangadurai and N.Nandhin[13] used various data mining algorithm and are applied on Pima datasets. It is found that the genetic algorithm gives a better performance over five data mining algorithm. In this paper[49] we have used Matlab tool for analysis and performed comparison of selected classification algorithms. After the comparative analysis we examined that neural network algorithms is more accurate and has less error rate. Our interface also provides the user the choice of selecting suitable prediction algorithm. We conclude that ANN has more precision than other models.

This assessment paper[2] focuses on several predictive analysis procedures and approaches and it is utilizing premature estimate of a several cases of diabetes from patient record. The different approach analytics procedures are applied in health records field for foreseeing case of diabetes and to find out effective ways to treat them in better manner. In this paper [37],the AROC for the proposed GBM model is 84.7% with a sensitivity of 71.6% and the AROC for the proposed Logistic Regression model is 84.0% with a sensitivity of 73.4%. The GBM and Logistic Regression models perform better than the Random Forest and Decision Tree models. In this work[38], we have analyzed the early prediction of diabetes by taking into account various risk factors related to this disease using machine learning techniques. Extracting knowledge from real health care dataset can be useful to predict diabetic patients. To predict diabetes mellitus effectively, we have done our experiments using four popular machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN) and C4.5 decision tree, on adult population data to predict diabetes mellitus.

Detection and analysis of clinical activity is the most important issue in real time scenario, because the lack of training samples and sufficient data's make these processes much complicated. There are several different methods to diagnosis and prognosis diabetes mellitus. This survey[16] presents a various techniques of the data mining approach to solve the diabetes disease diagnosis problem. From the analysis we discover several problems and finds in clinical datasets handling process. Sowjanya [17] had developed an android application-based solution to overcome the deficiency of awareness about DM in his paper. The application used the DT classifier to predict diabetes levels for users. The system also provided information and suggestions about diabetes.This paper[14] used an ensembling method, specifically with the "Vote" technique that combined three decision tree classification methods (Random Forest, NB Tree, and LMT). The ensembling method improved the prediction accuracy to AUC = 0.922. The study[15] shows the potential of ensembling and SMOTE approaches for predicting incident diabetes using cardiorespiratory fitness data. Devi epresents the development of an amalgam model for classifying Pima Indian diabetic database (PIDD). This amalgam model combines K-means with K Nearest Neighbour (KNN). They compare the results of simple KNN with cascaded K-means and KNN for the same k-values. The results are then compared by measuring the statistical measures such as accuracy, sensitivity and specificity and calculated using WEKA tool. For k=5, K-means and KNN has accuracy of 97% while the simple KNN has the accuracy of 73.17% and Amalgam KNN has accuracy of 97.4%. For k=3, Amalgam KNN has accuracy of 96.87% and simple KNN has accuracy of 72.65%.The author concluded that performance of the algorithm increases if the value of K increases .V.AnujaKumari, R.Chitra[22], used SVM with Radial Basis Function Kernal for classification of diabetes disease. They used MATLAB, R2010a for implementation. They found the accuracy rate as 78%.

In Preeti Verma, Inderpreet Kaur , Jaspreet Kaur paper[36] work,the performance of this method is evaluated using 10-fold cross validation accuracy, confusion matrix. The obtained classification accuracy using 10-fold cross validation is 96.58% in comparison with other spline SSVM technique. The results of this study showed that the modified spline SSVM was effective to detect diabetes disease diagnosis and this is very promising result compared to the previously reported results.In Dr. B .L. Shivkumar and S Thiyagarajan work[23], an effective machine learning algorithm is proposed for the classification of type dm patients. This machine learning algorithm used for classification will find the optimal hyper-plane which divides the various classes. Sneha and Tarun[24] proposed a method that aims to focus on selecting the attributes that ail in early detection of Diabetes Miletus using Predictive analysis. The result shows the decision tree algorithm and the Random forest has the highest specifcity of 98.20% and 98.00%, respectively holds best for the analysis of diabetic data. Naïve Bayesian outcome states the best accuracy of 82.30%. The research also generalizes the selection of optimal features from dataset to improve the classifcation accuracy. This paper[47] focuses that the use of data mining algorithms can be very helpful in early prediction and in consequence early precautions before the diagnosis of disease. The main goal of this paper is to provide a comparison and suggest best algorithm which can be used for the pattern recognition or prediction in healthcare fields. After the implementations of these algorithms it can be said that for PID dataset Decision Tree gives best accuracy 75.65%. The tool used for testing and validation is Rapid Miner while all algorithms worked with 70:30 ratio for training and testing.

In this paper[33],prediction of diabetes is done using ensemble voting classifiers for pima Indian diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 80% and 81% is achieved for data set by using 10-fold cross validation and by spitting data into 30% testing and 70% training.The paper [32] approached the aim of diagnoses by using ANNs and demonstrated the need for preprocessing and replacing missing values in the dataset being considered. Through the Modified training set, a better accuracy was achieved with lesser time required for training the set. In this paper [28],classification techniques such as Binary Logistic Regression, Multilayer Perceptron and K-Nearest Neighbor are classified for diabetes data and classification accuracy were compared for classifying data. A Class wise KNN(CKNN) methodology[34] for classification of diabetes dataset was proposed where the preprocessing of the dataset is done using normalization and an improvised model of KNN algorithm, i.e., class wise KNN algorithm is applied on the dataset for classification. This method achieves an accuracy of 78.16%. Rohan bansal[35] used KNN classifier for the diagnosis of diabetes; the attributes are selected using Particle swarm optimization (PSO) techniques. This method is proved to provide a prediction accuracy of 77% .

In Abdullah [40] work, Oracle Data miner and Oracle Database 10g used for Analysis and storage respectively .The parameters or factors were identified in this study .The target variables were identified based on their percentage .This study concentrated on the treatment of the patient .The patient divided into two categories old and young based on their age and predict their treatment .For both young and old diet controle indicates high percentage on this study. The treatment predictive percentage done by support vector machine. Xue-Hui Meng [41] in this study, the researchers used different data mining techniques to predict the diabetic diseases using real world data sets by collecting information by distributed questioner .In this study SPSS and weka tools were used for data analysis and prediction respectively and compare three techniques ANN, Logistic regression, and j48 .Finally it was concluded as j48 machine learning technique provide efficient and better accuracy.

Byoung Geol Choi, Seung-Woon Rha , Suhng Wook Kim , Jun Hyuk Kang , Ji Young Park , and Yung-Kyun Noh[26] successfully developed and verified a T2DM prediction system using machine learning and an EMR database, and it predicted the 5-year occurrence of T2DM similarly to with a traditional prediction model. In Kemal Polat work[29],Discussing Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) are used for diagnosis of diabetes disease. Also, proposed a new cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. The proposed system includes two stages. The first stage, they used Generalized Discriminant Analysis to discriminant feature variables between healthy and patient (diabetes) data as pre-processing step. The second stage, they used LS-SVM for classification of diabetes dataset. While LS-SVM obtained 78.21% classification accuracy using 10-fold cross validation, the proposed system called GDA-LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation and it is very promising compared to the previously reported classification techniques.

Hasan Temurtas [30] a multilayer neural network structure, trained by Levenberg–Marquardt (LM) algorithm and a probabilistic neural network structure were used. The results of the study were compared with the results of the previous studies that also focused on diabetes disease diagnosis and by using the same UCI machine learning database obtains 79.62% accuracy. The classification accuracy of MLNN with LM obtained by this study using correct training was comparatively better than those obtained by other studies except the classification accuracies by Polat and Gunes. Santi Wulan [31] Implemented MKS-SSVM technique to improve accuracy of the result

has been developed by many researchers. It is called Multiple Knot Spline SSVM (MKS-SSVM).Implement an experiment on Pima Indian diabetes dataset to evaluate the effectiveness of our method. The accuracy of previous results of this data is still below 80% using SSVM that is smooth support vector machine. Then, the proposed MKS-SSVM showed better performance in classifying diabetes disease diagnosis with accuracy of 93.2% which is better than previous reported results. It can be concluded from the study[63] that hybrid deep learning provides the most satisfactory results for prediction of diabetes. Least error rate and highest area under the ROC curve, accuracy and precision values provide evidence of better performance as compared to pure SVM and pure Deep Learning models.

In our proposed method[57], we have used attribute selection filter to select the subset of attributes from original data and then we have applied J48and decision stump data mining classification techniques which are used to predict diabetes. The performances of classifiers are evaluated through the confusion matrix in terms of accuracy and execution time. The Random Tree Algorithm gives 86.59% which is providing better Accuracy than other classifier's accuracy and also Random Tree algorithmtakes very minimum time to classify data sets than other classifiers.In the research paper[48],proposed system used well known and most commonly used machine learning algorithms. Algorithms used in this study are J48, KNN, NB, and Random Forest. The proposed method provides better accuracy of 93.62% in case of PIDD using stacking meta classifier. In case of large dataset 130-us hospital an ensemble method provides better accuracy than single prediction algorithm. This paper [50]summarizes the most frequently used Data Mining tool and techniques and the type of datasets and attributes that has been used to predict the diabetes disease in patients. It also includes the performance and accuracy of each algorithm that has been used in this study. The studies revealed that the classifiers would yield better results when the data is being pre-processed.

In this paper[53] ,we have used K- nearest neighbor algorithm for the diagnosis of diabetes mellitus. We have calculated accuracy and error rates for K=3, 5. The result is showing that as the value of k increases, accuracy rate and error rate will also increase. KNN is one of the most effective Artificial intelligence algorithms that is widely used for diagnostic purposes. More accurate and efficient results can be obtained through KNN. In this paper[55],various classification approaches had been implemented in data mining process. These approaches have been used to divide the data into different sets so that easily relation between different attributes can be identified. Different data mining techniques have been used to help health care professionals in the diagnosis of diabetes disease. Other data mining techniques are also used including kernel density, automatically defined groups, bagging algorithm, and support vector machine. In this paper[56],we use three different classification algorithms used in this experiment namely, ZeroR, OneR and Naïve Bayes. This experiment shows that Naïve Bayes is the fastest and ZeroR is the slowest. The performance comparison is found using weka data mining tool. Each model is converted into rules and those rules are incorporated into this application. It is found Naïve Bayes outperforms OneR and ZeroR. This research paper[60] is about Machine learning, that has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.

In this study, Quan Zou, Kaiyang Qu , Yamei Luo , Dehui Yin , Ying Ju and Hua Tang[25] used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used.Thangara[42] predicts the diabetic disease from clinical database by using Neural Network algorithm .Chitkara[43] done research to recognize the vocal qualities of patients having type 2 diabetes mellitus using MDVP.This paper [52] discuss about Machine learning on administrative data ,provides a powerful new tool for population health and clinical hypothesis generation for risk factor discovery, enabling population-level risk assessment that may help guide interventions to the most at-risk population. Using the approach described herein, it is possible to identify patients likely to develop type 2 diabetes with Prediction Of Type 2 Diabetes From Claims Data 285 at least 67% better PPV compared with traditional risk assessment methods for 0–2 years into the future. The extensive set of risk factors recovered by our method, for different stages of disease onset, can be a basis for additional hypothesis testing in medical research laboratories. Finally, our approach is general enough to be applied to different outcomes of interest, to build predictive models for different years into the future, and to analyze the risk factors as they emerge at different stages before the onset.

In this paper[59], we are currently testing the same data set divided identically into the same training and forecasting sets using both logistic regression and a linear perceptron model. We plan to have data to compare the ROC curves for this prediction using all three methods - ADAP, logistic regression and a linear perceptron. This paper [62]is using amalgamated technique the resultant values and derived with more accuracy where as the linear regression produced the confusion matrix thus resulting the compact the precise formation of weights based on characteristics and attributes where as LS-SVM classifies the estimation of probabilistic model where scheme can define the range in which the prediction can be made more perfectly based on type of categorization in diabetes and precautions respectively. This paper [64]concentrates about various data mining techniques and methods which are used for the early prediction of various diabetes. Data mining is a techniques used to extract useful information from existing large volume of data which enables you to gain more knowledge. Therefore applying Data mining methods and techniques will helps to predict the diabetes and also reduces the treatment cost. In this way data mining techniques are applied in medical data domain in order to predict diabetes and to find out efficient ways to treat them as well.

III. EXISTING PREDICTIVE ANALYSIS TECHNIQUES

Machine learning techniques: Machine learning technique is a technique which is deal in scientific field and it is learn from experience. The tenure ML is closely related to AI.

1. **Supervised learning :**Machine learning technique which is learning particularly a purpose i.e. Objective purpose that is an appearance of a form unfolding the statistics. e.g.: -GA algorithm, KNN algorithm, SVM algorithm etc.
2. **Unsupervised learning :** The machine learning technique which is trying on the way to determine the unseen arrangement of records among variables[2]

Data mining techniques : Data Mining is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. Due to the wide availability of huge amounts of data in electronic forms, and the imminent need for turning such data into useful information and knowledge for broad applications including market analysis, business management, and decision support, data mining has attracted a great deal of attention in information industry in recent years[5].

Nature-inspired algorithm: Nature is a great teacher and guide for human from ancient era and it is important for humans, birds and animal too. Nature inspired algorithm is algorithm which is inspired from nature, animal and bird specification and many algorithms inspired from nature for either search or optimization of any problem. Nature-inspired algorithm are mainly two types: -Evolutionary algorithm and swarm algorithm. Evolutionary algorithm example is Genetic algorithm and swarm algorithm example are PSO algorithm, Ant algorithm and it is played an important role in every field of human life[2].

IV. METHODS

4.1 Neural Networks/Multilayer Perception - Multi Layer Perception can be characterized as Neural Network and Artificial insight without capability. Neural systems, have a wonderful capacity to get significance from entangled or uncertain information, and can be utilized to concentrate examples and distinguish patterns that are too mind boggling to be in any way seen by either people or other PC procedures[18].

4.2 Logistic Regression-The logistic function is also called as the sigmoid function and it is developed by statisticians to describe the properties of population growth in ecology, that rising quickly and maxing out to the carrying capacity of the environment. It is an S-shaped curve that can obtain any real-valued number and mapped into the values between 0 and 1.

$$1 / (1 + e^{-value})$$

where e denotes the base of the natural logarithms and the value is the actual numericalvalues that we want to transform[5].

4.3 Adaboost.M1: AdaBoost (Adaptive Boosting) is a Boosting machine learning meta-algorithm which theoretically can be used to significantly reduce the error of any learning algorithm that consistently generates classifiers whose performance is a little better than random guessing. It is a nominal class classifier using the Adaboost M1 ensemble method which means only nominal class problems can be solved. It is Often dramatically improves performance, but sometimes over fits. On the other hand, Bagging is an ensemble method that creates separate samples of the training dataset and creates a classifier for each sample. It reduces the variance

.

Each instance in the training dataset is weighted. The initial weight is set to:

$$weight(xi) = 1/n$$

Where xi is the i'th training instance and n is the number of training instances

4.4 Random Forest

It is supervised learning, used for both classifcation and Regression. Te logic behind the random forest [44, 45] is bagging technique to create random sample features. The diference between the decision tree and the random forest is the process of fnding the root node and splitting the feature node will run randomly.

The Steps are given below

a. Load the data where it consists of "m" features representing the behaviour of the datase.
b. Te training algorithm of random forest is called bootstrap algorithm or bagging technique to select n feature randomly from m features, i.e. to create random samples, this model trains the new sample to out of bag sample(1/3rd of the data) used to determine the unbiased OOB error.
c. Calculate the node d using the best split. Split the node into sub-nodes.
d. Repeat the steps, to fnd n number of trees.
e. Calculate the total number of votes of each tree for the predicting target. The highest voted class is the fnal prediction of the random forest[24].

4.5 IBK

In Weka, nearest neighbor classification algorithm is known as IBK (the IB stands for Instance-Based, and the K allows us to specify the number of neighbors to examine). IBK is a useful data mining technique that allows us to use past data instances with known output values to predict an unknown output value of a new data instance. It predicts very accurate but often perform slow, generally perform well for large value ok K[65].

4.6 J48

This is an open source Java execution of basic C4.5 choice tree calculation. J48 is a straightforward technique to manufacture a choice tree from the preparation information by beginning at the top, with the entire preparing dataset[7].

4.7 SVM

This is an arrangement and relapse expectation device that utilizations AI hypothesis to amplify prescient exactness while consequently maintaining a strategic distance from over-fit to the information[65].

4.8 Naive Bayes

This arrangement method depends on the Bayes' hypothesis. This classifier accept that the nearness of a specific component in a class is inconsequential to the nearness of some other element[7].

4.9Random Tree

It is a tree which is formed by a stochastic process. Random trees normally refer to randomly built trees which have nothing to do with machine learning. However the popular machine learning framework Weka uses the term to refer to a decision tree built on a random subset of columns[68].

V. PERFORMANCE EVALUATION METRICS

- Accuracy = (TP + TN)/(TP + FP + TN + FN)
- Sensitivity = TP/(TP+FN)
- Specificity = TN/(FP+TN)
- Precision = TP/(TP+FP)
- F1 Score = (2*Precision*Recall)/(Precision+Recall)

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively.
The model with highest the sensitivity, specificity, and accuracy is the best predictive model.

VI. CONCLUSION

This paper summarizes the most frequently used techniques and the type of datasets and attributes that has been used to predict the diabetes disease in patients. It also includes the performance and accuracy of each algorithm that has been used in this study.. Certain papers also focus on reducing the misclassification or correlated datasetsStill there is wide gap towards the accuracy and computation speed that needs to be addressed. There is a wide opportunity for the researchers to focus on the Diabetes dataset.

REFERENCES

[1] K.Priyadarshini1, Dr.I.Lakshmi.2017. A Survey on Prediction of Diabetes Using Data Mining Technique from International Journal of Innovative Research in Science, Engineering and Technology, ISSN(Online): 2319-8753

[2] Sonali Vyas , Rajeev Ranjan , Navdeep Singh , Arohan Mathur.2019. Review of Predictive Analysis Techniques for Analysis Diabetes Risk.

[3] Kumari, Sonu, and Archana Singh .2013. A data mining approach for the diagnosis of diabetes mellitus. Intelligent Systems and Control (ISCO), 7th International Conference on. IEEE.

[4] T.Jayalakshmi and Dr.A.Santhakumaran.2010.A Novel Approach for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks", International Conference on Data Storage and Data Engineering,159-163.

[5] J. Steffi, Dr.R.Balasubramanian.2018.Predicting Diabetes Mellitus using Data Mining Techniques from International Journal Of Engineering Development And Research. Volume 6, Issue 2 ,ISSN: 2321-9939

[6] Komi, M., Li, J., Zhai, Y., & Zhang, X. 2017, June. Application of data mining methods in diabetes prediction. In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on (pp. 1006-1010). IEEE.

[7] Rahul Joshi, Minyechil Alehegn.2017.Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach from International Research Journal of Engineering and Technology, p-ISSN: 2395-0072

[8] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. 2016. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115-121.

[9] Song, Y., Liang, J., Lu, J., & Zhao, X.2017. An efficient instance selection algorithm for k nearest neighbour regression. Neurocomputing, 251, 26-34.

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

[11] Monika , Pooja Sharma.2018. Survey on Prediction and Analysis of Diabetic Data using Machine Learning Techniques from International Journal of Computer Sciences and Engineering, E-ISSN: 2347-2693

[12] D. Jeevanandhini , E. Gokul Raj ,V. Dinesh Kumar, N. Sasipriyaa.2018. Prediction of Type2 Diabetes Mellitus Based on Data Mining from International Journal of Engineering Research & Technology (IJERT), chnology (IJERT) ISSN: 2278-0181

[13] Dr. K. Thangadurai, N.Nandhini.2016. Comparison of data mining algorithms for prediction and diagnosis of diabetes mellitus from International Journal of Scientific & Engineering Research, Volume 7, Issue 5, ISSN 2229-5518

[14] Manal Alghamdi, Mouaz Al-Mallah, Steven Keteyian , Clinton Brawner , Jonathan Ehrman and Sherif Sakr.2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project

[15] M Nirmala Devi, Balamurugan.S Appavu alias ,U.V Swathi.2013.An amalgam KNN to predict Diabetes Mellitus, IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, Madurai, Tamil Nadu, India

[16] B. Senthil Kumar , Dr. R. Gunavathi.2016. A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis from International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified, ISSN (Online) 2278-1021

[17] Ms. K Sowjanya.2015. MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. IEEE International Advance Computing Conference (IACC).

[18] Ambika Rani Subhash, Ashwin kumar UM.2019. Accuracy of Classification Algorithms for Diabetes Prediction from International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958

[19] K. Saravananathan, T. Velmurugan, "Analyzing Diabetic Datausing Classification Algorithms in Data Mining", International Journal of Science and Technology, Vol. 9 (43), November 2016

[20] S. sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi.2015. Comparison of Datamining Algorithms in the Diagnosis of Type II Diabetes, IJCSA, vol. 5, no. 5.

[21] V. Kumar and L. Velide.2014. A Data mining Approach for Prediction and Treatment Ofdiabetes Disease from IJSIT

[22] V. A. Kumari and R. Chitra.2013.Classification of Diabetes Disease using Support Vector Machine from IJERA

[23] Kessler, R. C.2016.Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports,Molecular psychiatry.

[24] N. Sneha and Tarun Gangil.2019. Analysis of diabetes mellitus for early prediction using optimal features selection,Journal of Big Data

[25] Quan Zou, Kaiyang Qu , Yamei Luo , Dehui Yin , Ying Ju and Hua Tang.2018.Frontiers in Genetics

[26] Byoung Geol Choi, Seung-Woon Rha , Suhng Wook Kim , Jun Hyuk Kang , Ji Young Park , and Yung-Kyun Noh.2019,Yensai Medical Journal

[27] Prema N S, Varshith V, Yogeswar J.2019. Prediction of Diabetes using Ensemble Techniques from International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878,

[28] S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar.2017. Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques from International Journal of Statistics and Systems, ISSN 0973-2675 ,pp. 183-188

[29] K. Polat, S. Gunes and A. Aslan.2008.A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine, Expert Systems with Applications, vol. 34(1), pp. 214–221

[30] T. Hasan, Y. Nejat, T. Feyzullah.2009., A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications, vol. 36, pp. 8610- 8615

[31] Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain and S.P. Rahayu.2009. A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis", Journal of Computer Science vol.5 (12), pp.1006-1011, ISSN 1549-3636

[32] T.Jayalakshmi and Dr.A.Santhakumaran, 2010.A Novel Approach for Diagnosis of Diabetes Mellitus Using Artificial Neural Network from International Conference on Data Storage and Data Engineering,, pp. 159-163.

[33] Prema N S, Varshith V, Yogeswar J.2019. Prediction of Diabetes using Ensemble Techniques from International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878

[34] Y. A. Christobel and C. Sivaprakasam.2013.New Classwise K Nearest Neighbor (Cknn) Method For The Classification Of Diabetes Dataset, Int. J. Eng. Adv. Technol., vol. 2, pp. 396–400.

[35] R. Bansal, S. Kumar, and A. Mahajan.2017. Diagnosis of diabetes mellitus using PSO and KNN classifier from International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017, pp. 32–38.

[36] Preeti Verma , Inderpreet Kaur , Jaspreet Kaur.2016. Review of Diabetes Detection by Machine Learning and Data Mining from International Journal of Advance Research , Ideas and Innovations in Technology, ISSN: 2454-132X

[37] Hang Lai1, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi1 and Xin Gao.2019. Predictive models for diabetes mellitus using machine learning techniques,BMC Endocrine Disorders Article.

[38] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus,IEEE.

[39] S.Saru and S.Subashree.2019.Analysis And Prediction Of Diabetes Using Machine Learning from International Journal of Emerging Technology and Innovative Engineering Volume 5, Issue 4 , (ISSN: 2394 – 6598).

[40] 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(2), 127-136.

[41] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

[42] Thangarasu Gunasekar and Assoc. Prof. Dr. Dominic P.D.D.2104. Prediction of Hidden Knowledge from Clinical Database using Data mining Technique, IEEE 978-1-4799-0059-6.

[43] Ayush Anand, Divya Shakti, Prediction of Diabetes Based on Personal Lifestyle Indicators, IEEE, International Conference on Next Generation Computing Technologies, pp. 673-676, 2015.

[44] K. Sharmila and S. Manicka.2015. "Efficient Prediction and Classification of Diabetic Patients from big data using R," International Journal of Advanced Engineering Research and Science, vol. 2.

[45] NIyati Gupta, A. Rawal, and V. Narasimhan.2013, Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data, IOSR Journal of Computer Engineering, vol. 11, no. 5, pp. 70-73.

[46] Sassanian and G. Hari Sekaran.2015. Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients fromInternational Journal of Science and Research, vol. 4.

[47] Amina Azrar, Muhammad, Yasir Ali, Khurram Zaheer.2018. Data Mining Models Comparison for Diabetes Prediction from International Journal of Advanced Computer Science and Applications

[48] Minyechil Alehegn, Rahul Raghvendra Joshi, Preeti Mulay.2019. Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach from International Journal Of Scientific & Technology Research Volume 8, Issue 09, ISSN 2277-8616

[49] Musavir Hassan, Muheet Ahmad Butt and Majid Zaman Baba.2017. Logistic Regression Versus Neural Networks: The Best Accuracy in Prediction of Diabetes Disease from Asian Journal of Computer Science and Technology, ISSN: 2249-0701

[50] J. Bagyamani, K. Saravanapriya.2019. Data Mining Classification Techniques for the Diagnosis of Diabetes Mellitus – A Review from International Journal of Computational Intelligence and Informatics, Vol. 8: No. 4.

[51] Muhammad Noman Sohail , Ren Jiadong, Musa Muhammad Uba, Muhammad Irshad, Wasim Iqbal, JehangirArshad & AntonyVerghese John.2019.A hybrid Forecast Cost Beneft Classifcation of diabetes mellitus prevalence based on epidemiological study on Real-life patient's data,Scientific Report .

[52] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag .2015.Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors,Orginal Report.

[53] Krati Saxena1 , Dr. Zubair Khan , Shefali Singh.2014. Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm from International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4.

[54] N.Deepika, Dr.S.Poonkuzhali.2018. Design of Hybrid Classifier for Prediction of Diabetes through Feature Relevance Analysis from International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 10

[55] Harleen, Dr. Pankaj Bhambri.2016. A Prediction Technique in Data Mining for Diabetes Mellitus from Journal of Management Sciences and Technology, ISSN -2347-5005

[56] M.Mounika, S.D.Suganya, B.Vijayashanthi, S.KrishnaAnand.2015.Predictive Analysis of Diabetic Treatment Using Classification Algorithm from International Journal of Computer Science and Information Technologies

[57] DrPrakash Kuppuswamy , DrRajan John, DrShanmugasundaram Marappan.2019.Performance Evaluation of Data Mining Algorithm on Electronic Health Record of Diabetic Patients from nternational Journal of Engineering Science Invention (IJESI)

[58] J. Pradeep Kandhasamy, S. Balamurali .2015. Performance Analysis of Classifier Models to Predict Diabetes Mellitus, Procedia Computer Science

[59] Jack W. Smith,JE Everhart, WC Dicksont, WC Knowler, RS Johannes.1988.SCAMC.

[60] Tejas N. Joshi, Prof. Pramila M. Chawan.2018.Diabetes Prediction Using Machine Learning Techniques from Journal of Engineering Research and Application ,ISSN: 2248-9622, Vol. 8, Issue 1

[61] S. R. Priyanka Shetty, Sujata Joshi.2016.A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique from I.J. Information Technology and Computer Science

[62] Vaishali , Nisha Pandey.2018. Diabetes Prediction using Linear Regression, Decision Tree & Least Square Support Vector Machine from International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801

[63] Deeksha Kaul, 2 Harika Raju, 3 B.K. Tripathy.2017. Comparative Analysis of Pure and Hybrid Machine Learning Algorithms for Risk Prediction of Diabetes Mellitus from Helix Scientific Explorer.

[64] Karnika Dwivedi, Dr. Hari Om Sharan.2018. Review on Prediction of Diabetes Mellitus using Data Mining Technique from International Journal of Engineering and Technical Research (IJETR).

[65] Md. Aminul Islam, Nusrat Jahan.2017. Prediction of Onset Diabetes using Machine Learning Techniques from International Journal of Computer Applications.

[66] Jatin N Bagrecha, Chaithra G S , Jeevitha S.2019. Diabetes Disease Prediction using Neural Network from International Journal for Research in Applied Science & Engineering Technology (IJRASET)

[67] Sujaritha.M, Monica Murugesan, Bhuvana MK, Saleekha.2019. Risk Analysis of Diabetes using IoT and Deep Learning from International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[68] Basharat Naqvi , Arshad Ali , Muhammad Adnan Hashmi and Muhammad Atif .2018. Prediction Techniques for Diagnosis of Diabetic Disease: A Comparative Study from International Journal of Computer Science and Network Security.