International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 10 | October - 2022 Impact Factor: 7.185 ISSN: 2582-3930

A comparative analysis of early stage Diabetes Mellitus prediction using machine learning approach

SUJITH KUMAR THIRUVEEDULA[1],CHILUKURI LAKSHMI MANI SHANKAR[2,]
KRISHNA PRADEEP REDDY S[3],VEJELLA CHARITH VARMA[4]
Department of Computational Intelligence
ts6346@srmist.edu.in,cq1096@srmist.edu.in,kr9532@srmist.edu.in,vv2808@srmist.edu.in
SRM Institute of Science and Technology,Chengalpattu,India.

Abstract

Diabetes mellitus (DM) is a serious chronic illness and causes a severe impact on the complete human body. Diabetes that goes undiagnosed increases the risk of heart attack, renal disease, eye problems, and organ failure. Many diabetic patients are unaware that they have diabetes, and its complications are rising day by day. As Diabetes mellitus is not curable, it must be detected early to live a healthy life. The purpose of this research is to use relevant features to create a machine learning model and choose the best classifier to get the best results when compared to clinical results. The Diabetes dataset is taken from Kaggle. The dataset is preprocessed to improve the model's learning capabilities. To identify individuals as diabetes or non-diabetic, several machine learning approaches such as Logistic Regression, KNeighborsClassifier, Naive Bayes, Random Forest classifier, Decision tree classifier, AdaBoost, and XGBoost, Light GBM is used. To choose the best model, the performance of machine learning models is compared by using accuracy, recall, precision, and F1-Score as the evaluation metrics.Finally got an accuracy of 99% in DecisionTree.

Keywords

Diabetes mellitus,Feature selection,Machine Learning,Decision Tree,Precision.

1. INRODUCTION

Diabetes mellitus is a chronic disease, It refers to a group of metabolic conditions characterised by elevated blood sugar levels due to either inefficient insulin production or it can be due to body cells response to insulin poorly. Insulin is the hormone which regulates blood glucose level.

Too much sugar circulates in blood due to this chronic condition.The symptoms may include increased hunger,thirst,vision problems,tiredness. Prolonged Diabetes mellitus should be treated on time,if not it may lead to serious health issues on various organs like kidney,heart,brain,eye.Physical inactivity, hereditary issues, being overweight, and having insulin-resistant blood cells are all factors that contribute to diabetes mellitus.

2. LITERATURE REVIEW

For diabetes prediction, Hang Lai presented the Logistic Regression and Gradient Boosting Machine (GBM) approaches. The decision tree and random forest models are compared to this model. This model has an accuracy of 84.7 percent and a sensitivity of 71.6 percent when using the GBM model, and an accuracy of 84 percent and a sensitivity of 73.4 percent when using Logistic Regression. According to this study, GBM and Logistic Regression Models outperform Random Forest and Decision Tree Models[1].Toshita Sharma have discussed SVM,Decision Tree,Navie Bayes. Deep learning performs better on image datasets, hence visuals would be preferable for diabetes diagnosis. To compare their performance on the data, most researchers have implemented various algorithms in both machine and deep learning, while some have merged two or three approaches[2].Swapna used Deep Learning techniques on ElectroCardioGram signals by SVM algorithm with a high accuracy of 95.7 percent[15] Sisodia used machine learning techniques such as DT,NB,SVM algorithms with an accuracy of 76.3 percent[16],Han Wu used data mining techniques for developing type 2 diabetes and he got the greater accuracy of 95.42%[17].According to meng for identifying diabetes there are 12 risk factors and he used logistic regression ,artificial neural network and decision tree .Decision tree gave the best result[18] .Choubey used a hybrid algorithm with a genetic algorithm .They found that hybrid algorithm performed best[19].Boshra Farajollahi used six algorithms Logistic regression, Support Vector Machine (SVM), xgboost, Random Forest, Decision Tree and Adaboost.comparing all the algorithms four showed the better accuracy .In those Adaboost and logistic regression showed the best[4]. Harleen used the J48 algorithm, which had a 73.8 percent accuracy the Naive Bayes algorithm, which had a 76.3 percent accuracy, the discriminant analysis

algorithm, which had a 76.3 percent accuracy[13], the KNN Algorithm, which had a 71.1 percent accuracy, the SVM with Linear Kernel function algorithm, which had an accuracy of 74.1 percent, and the SVM with RBF Kernel function algorithm, which had an accuracy of According to ravi, the data preparation technique has a high level of accuracy[14].N. Sneha used five algorithms KNN,SVM,Randomforest,Navie Bayes,Decision Tree.Decision tree and Random Forest have the accuracy of 73.48% and 75.39%,SVM has the Highest accuracy of 77.73%,NB has an accuracy of 73.48%.The suggested algorithm is SVM[6].K.VijiyaKumar suggested the Random Forest algorithm for diabetes prediction in accordance with organisational procedures that can detect diabetes in a patient soon and consistently using machine learning techniques. The proposed methodology delivers superior diabetic prediction results, confirming that the prediction system can estimate diabetes disease adequately, reliably, and, most importantly, promptly[7].

3. DESCRIPTION OF CLASSIFICATION ALGORITHMS

3.1 K-Nearest neighbours:-

K-Nearest Neighbor (KNN) is a basic Machine Learning algorithm that handles classification and regression. It's also called a lazy learner algorithm since it doesn't understand the learning algorithm right away; instead, it saves the data and uses it to identify later.For the prediction of a new data point, the algorithm determines the closest data points in the training data set (its nearest neighbours). The number of nearest neighbours, K, is always a positive integer in this case. A neighbor's value is chosen from a list of classes. This algorithm can be used as a classifier or regression model.

3.2 Logistic regression:

The supervised learning approach includes logistic regression.It is based on the probability approach. It's used to calculate the probability of a binary response given one or more predictors. They can have a continuous or discrete nature. We adopt logistic regression as we want to identify or segregate some instances into segments. It aims to classify data only in binary form, such that, in 0 and 1 quantities, which leads to a situation in which a patient's diabetes status is labeled as positive or negative. The major aim of logistic regression is to find the best

fit, which describes the connection between the target and regression models. Logistic regression is a computation tool for predicting outcomes.

3.3 Decision tree:

The decision tree is a simple classification technique. A decision tree is a strategy for partitioning a given dataset into two or more test data periodically. It's a learning algorithm that would be guided. When the objective property is classified, this method is taken. The procedure is based on input features is expressed by a design with a tree-like structure called a decision tree. Any types, graphs, texts, discrete, continuous, and so on are among the input variables.

3.4 Naive Bayes:

A classification algorithm is a forecasting model centered on the Bayes theorem and the requirement for predicted independence. The naïve Bayesian procedure is being used to obtain the dataset as statistics, do evaluation, and calculate one of most likely designation by using Bayes Theorem. It develops a test statistic from input data and serves in the supervised segmentation of a sample of large datasets. It's a consistent segmentation solution that works well with enormous datasets.

3.5 ADA BOOST:

AdaBoost, also described as Adaptive Boosting, is a Machine Learning method used for the Ensemble Method. One of most common algorithm to use with AdaBoost is decision trees including one level, or decision trees for only one split. These trees are also termed as Decision Stumps. This approach yields a model by providing all of the data points the same weight. It therefore gives inaccurately labeled points a higher weight. All points with higher weights are given equal importance in the next model. This would continue to train models till it receives a low error.

3.6 XGBOOST:

Extreme gradient boosting (XGBoost) is very well known as gradient boosting methodology that enhance the efficiency and accuracy of tree-based (sequential decision trees) machine learning algorithms.This is the

important often widely used algorithm in applied machine learning. XGBoost is classed as a boosting mechanism in Ensemble Learning. To maximize prediction accuracy, ensemble learning integrates several models into a collection of predictors. By applying a load to the models in the boosting technique, the inaccuracy generated by previous models is considered to be sorted by successive models.

3.7 Random Forest:

It's an ensemble learning method that can be used for classification and regression. It's a well-known ensemble learning method. By lowering variation,the random forest increases the performance of decision trees. It proceeds by training a vast number of decision trees and then delivering the mode of the classes, categorisation, or overall average prediction (regression) of the individual trees as a class.

3.8 Light GBM:

LightGBM is a type of machine learning algorithm. Decision tree models are required to generate ensembles. Trees are introduced to the ensemble one by one and fitted to rectify prediction mistakes generated by previous models.LightGBM is a gradient boosting framework based on decision trees that optimizes model efficiency by using less memory.

4. METHODOLOGY

4.1 Dataset description

Diabetes dataset is taken from kaggle.Total participants are 2000.This dataset has 9 columns Pregnancies,Age,Glucose,BloodPressure,BMI,Insulin,DiabetesPedigreeFunction,SkinThickess, Outcome. Here ‘Outcome’ is the target feature.

4.2 Preprocessing

This diabetes dataset has the value '0' in the following columns Glucose, BloodPressure, SkinThickness, Insulin, BMI. Glucose has 13, BloodPressure has 90, Insulin has 956, SkinThickness has 573, BMI has 28 missing values. The rows with value '0' are removed. Now the shape of dataset is 1035x9. We can't have '0' as a reading for Glucose,BloodPressure,SkinThickness,Insulin and BMI. Since removal of '0' will help to get a very good accurate model. If we try to balance the data,under sampling will eliminate potentially helpful information that might be beneficial in the development of rule classifiers and also selected samples may cause biasing. Random over sampling may cause overfitting.

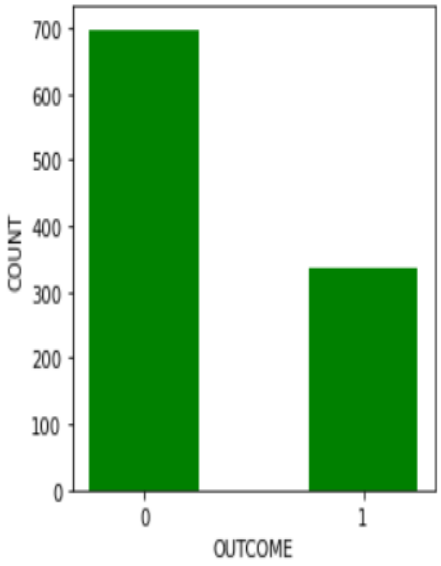

Fig 1: OUTCOME VS COUNT

From the Fig 1, we can see that non diabetic people are twice as diabetic people.Balancing this data will lead to loss of many valuable datas from dataset,due to this dataset becomes small, models are unable to learn properly,ultimately accuracy becomes low.

5.RESULT

5.1 EFFICIENCY COMPARISON OF MODELS

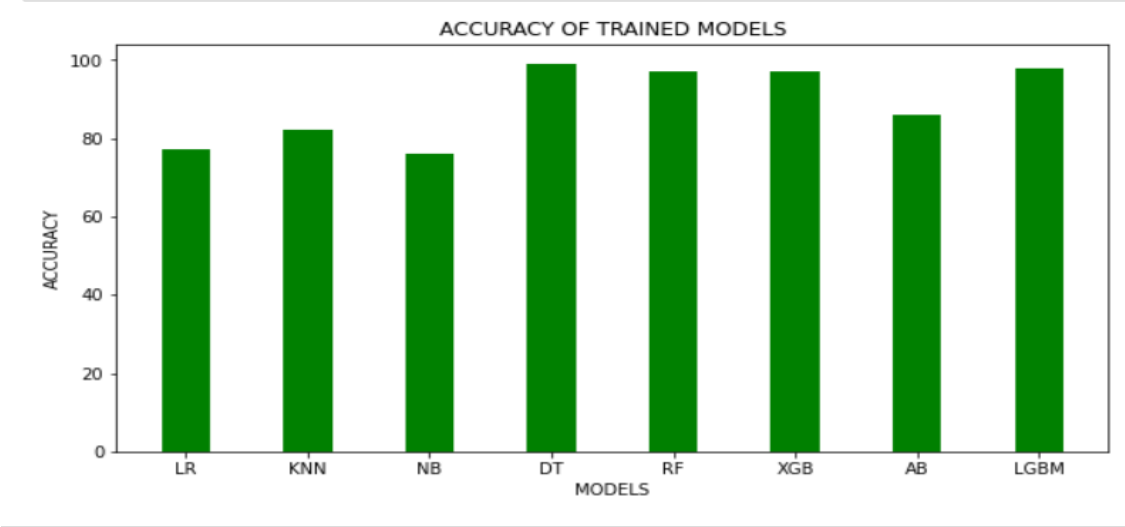

Fig 2: ACCURACY VS MODELS

Fig 2 shows the analysis of efficiency of 8 machine learning models based on accuracy.Decision Tree is the most accurate one , second highest is LightGBM, then Random forest,XGBoost have same accuray of 97%.Adaboost has 86% and KNN has 82%. Logistic Regression and Naive Bayes has accuracy less than 80%.

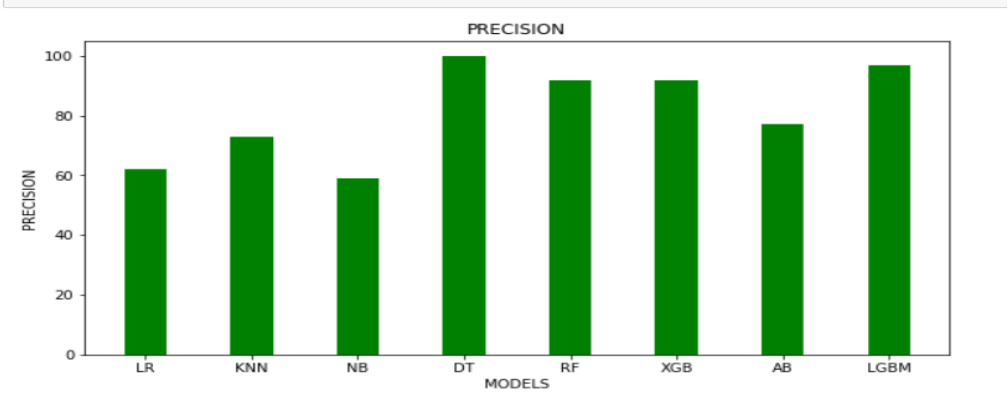

Fig 3: PRECISION VS MODELS

Fig 3 shows the analysis of efficiency of 8 machine learning models based on precision. Decision Tree has the highest precision, second highest is LightGBM, then Random forest,XGBoost have same precision of 92%.Adaboost has 77% and KNN has 73%. Logistic Regression has 62% and Naive Bayes has precision less than 60%.

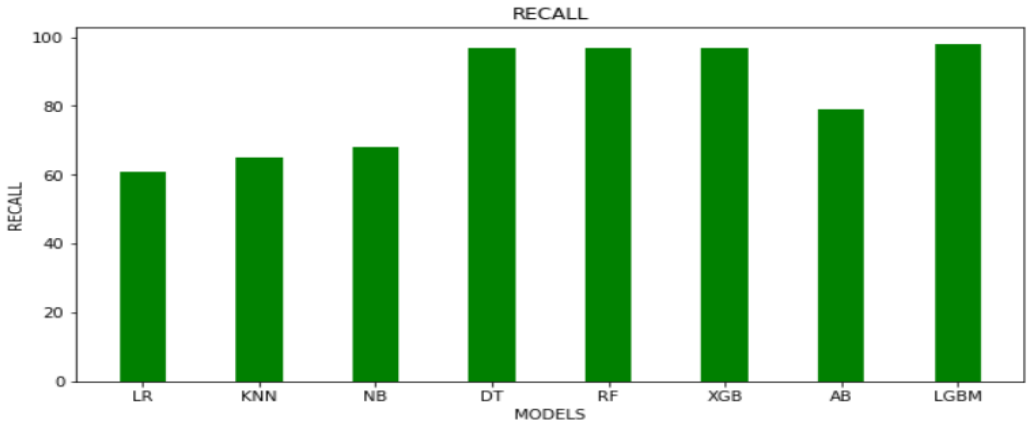

Fig 4: RECALL VS MODELS

Fig 4 shows the analysis of efficiency of 8 machine learning models based on recall. LightGBM has the highest recall value.Decision Tree,Random Forest,XGBoost has the same and second highest recall value of 97%,then Random forest. Adaboost has 79% and KNN,Logistic Regression,Naive Bayes has recall less than 70%.

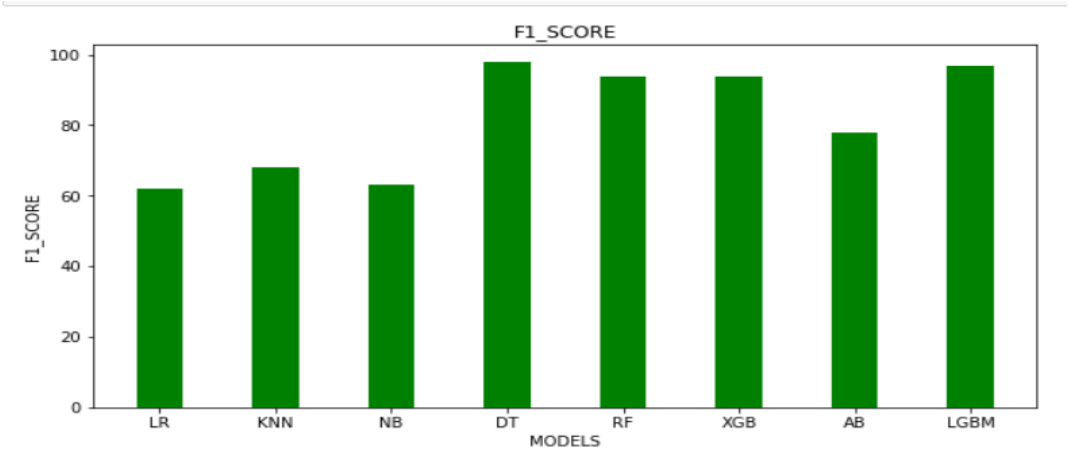

Fig 5: F1-SCORE VS MODELS

Fig 5 shows the analysis of efficiency of 8 machine learning models based on F1-SCORE. Decision Tree has the highest F1-SCORE of 98%, second highest is LightGBM, then Random forest and XGBoost have same F1-SCORE of 94%. Adaboost has 78% and KNN,Logistic Regression and Naive Bayes has accuracy less than 70%.

TABLE 1 : PRECISION,RECALL,F1-SCORE,ACCURACY OF CLASSIFIERS

MODEL	LR	KNN	NB	DT	RF	XGB	AB	LGBM
PRECISION	0.62	0.73	0.59	1.00	0.92	0.92	0.77	0.97
RECALL	0.61	0.65	0.68	0.97	0.97	0.97	0.79	0.98
F1 SCORE	0.62	0.68	0.63	0.98	0.94	0.94	0.78	0.97
ACCURACY	0.77	0.82	0.76	0.99	0.97	0.97	0.86	0.98

6. CONCLUSION:-

Diabetes diagnosis and detection at an early stage are essential in the real world..In our proposed model, we used 1035 instances from 2000 instances. The missing values were removed during data cleaning. In this study, 8 machine learning algorithms were completely studied and their efficiency were compared. 99% accuracy was achieved by using decision tree algorithm.The findings of this study will aid health officials to make a decision quick to treat diabetes and save lives.

7. ACKNOWLEDGEMENT

We would like to thank our faculty advisor Dr.S.Selvakumarasamy(Assistant Professor) and our Faculty members Dr.G.Maragatham(Associate Professor),Ms.A.Jackulin Mahariba(Assistant Professor),Department of Computational Intelligence,SRMIST for helping us in this study and to complete this work.

REFERENCES

[1] Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC endocrine disorders*, *19*(1), 1-9.

[2] Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1), 1-16.

[3] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*, 706-716.

[4] Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F., & Sayadi, M. J. (2021). Diabetes diagnosis using machine learning. *Frontiers in Health Informatics*, *10*(1), 65.

[5] 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.

[6]Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, *6*(1), 1-19.

[7] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[8]Mahariba, A. J., & Patel, S. (2021). Smart Band: An Integrated Device for Emergency Management. *arXiv preprint arXiv:2107.14100*.

[9] https://www.kaggle.com/johndasilva/diabetes

[10] Eswari, T., Sampath, P., Lavanya, S. (2015) “Predictive methodology for diabetic data analysis in big data.” Procedia Computer Science 50:203-208.

[11] G. Krishnaveni*, T. Sudha,” A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques” in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT2017), vol. 3, Issue 1, pp. 5-11, 2017.

[12] Aiswarya I., S. Jeyalatha and Ronak S., “Diagnosis Of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015

[13] Harleen, B. (2016). A Prediction Technique in Data Mining for Diabetes Mellitus. *Journal of Management Sciences and Technology*, *4*(1).

[14]G. Krishnaveni*, T. Sudha,” A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques” in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT2017), vol. 3, Issue 1, pp. 5-11, 2017.

[15]] Swapna, G., Vinayakumar R., Soman K. P. (2018) “Diabetes detection using deep learning algorithms.” ICT Express 4 (4): 243-246.

[16]Sisodia, D., Sisodia, D. S. (2018) “Prediction of diabetes using classification algorithms.” Procedia computer science 132: 1578-1585.

[17]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.

[18] 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): 93-9

[19]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.

[20] 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