2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT) | 978-1-6654-7416-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/CISCT55310.2022.10046630

Performance analysis of machine learning models for prediction of diabetes

Anuj Mangal
Department of Computer Engineering and Applications,
GLA University, Mathura, India
anuj.mangal@gla.ac.in

Vinod Jain
Department of Computer Engineering and Applications,
GLA University, Mathura, India
vinod.jain@gla.ac.in
(ORCID-0000-0003-0260-7319)

***Abstract*—Machine Learning is an emerging technology in artificial intelligence and computer science. Lot of researchers are applying Machine Learning (ML) models for solving various problems. Disease prediction using machine learning is a very hot area of research for many researchers. Diabetes is one of the major medical problems in common people. Because of improper life style and bad eating habits of people, this disease is increasing with a very high speed and affecting a lots of people throughout the world. It is very difficult to predict the chances of diabetes in human being with the help of existing methods and medical tests. Machine learning algorithms can be trained on a data set of the people or patients and then the trained model can be used in early prediction of diabetes in a person based on his or her medical symptoms. This paper applied the power of ML algorithms and implement two most commonly used ML models for diabetes prediction. After the experiment we are able to predict the diabetes by achieving 99% accuracy using random forest machine learning algorithm.**

Keywords—Predictive Analysis; Machine Learning; Diabetes Prediction; Random Forest; Logistic Regression

I. Introduction

The field of artificial intelligence [1] is growing day by day. It's some of the application areas are handwriting recognition, speech recognition, pattern identification, share market prediction, real state forecasting, disease prediction, and many others [2]. During the past two decades researchers are working on exploring new application areas for machine learning algorithms [3][4]. In medical science there are many usage of machine learning algorithms such as cancer prediction, diabetes prediction, stroke prediction etc.

The diabetes is mainly and broadly classified as of two types. Type 1 diabetes occurs when immune system of the human body damages the beta cells of the pancreas. In type 1 either no insulin is produced inside the body or very less amount of insulin is produced in the human body.

In type 2 diabetes, then human body is not able to interact with insulin or the cells in the human pancreas are not able to produce the sufficient amount of insulin so that glucose level in the human body can be managed. Because of the insufficient amount of insulin level, the blood glucose level starts increasing and the human body get suffered with diabetes. The diabetes is responsible for lots of deaths all over the world in last two years.

Without proper treatment the diabetes may badly affect the other parts of the human body. It may raise other complications in the human body. The other human body functions that may affect because of diabetes are heart disease, stroke, ulcers, failure of kidney, loss of vision by the eyes which may lead to blindness. Even the Covid -19 also badly affect the health of the people affected from diabetes. It also reduces the immunity power of the human body. The early prediction and early cure of the diabetes can control this diabetes. The diabetic patients die after affecting by the Covid – 19 virus.

The main cause of diabetes is the lack of insulin in the human body. When a patient gets diabetic then its main symptoms includes repeated urination, feeling a lots of hunger, feeling lots of thirsty and drinking water. The medical cure of the diabetes is almost not possible. The disease can only be controlled by following a good life style. The diabetes is responsible of failure for many important organism of the human body. The kidney may be badly affected because of diabetes it may affect human eyes badly. The heart failure may also occur because of diabetes.

In Machine Learning (ML), a statistical model is developed that is capable of learning from its experiences and thus reducing the errors and improving its performance [1][5]. There are many machine learning algorithms available which are based upon statistical tools. The most commonly used machine learning algorithms which gives better results in predictive analysis are Logistic Regression, Decision Trees, Nave Bayes, Random Forest, XGBoost, ADABoost etc. [6][7]. Many of these algorithms are based upon statistics, Neural Networks, classification and association techniques, support vectors etc. The author applied] ML models for prediction of very serious diseases such as cancer, Urinary diseases and Covid-19 using machine learning algorithms.

Predictive analysis includes the techniques such as machine learning, data mining and many other statistical or AI based approaches which are used to take decisions or classifications [8][9]. Data mining models are developed which are based upon statistical techniques and are found very useful in health care system to predict diseases.

Authorized licensed use limited to: Khulna Univ of Engineering & Technology - KUET. Downloaded on March 30,2024 at 11:36:11 UTC from IEEE Xplore. Restrictions apply.

Now a day, diabetes is the most common disease among the people. It is very common and non-communicable disease which affect the working of pancreas. If a person is suffering from diabetes, then either its pancreas is not generating enough insulin to dissolve all the glucose or the body is not utilizing the insulin hormone. Bad life style and bad eating habits are the two main causes of diabetes. It is also caused by obesity, high blood pressure.

In order to early prediction of diabetes, machine learning algorithms are the best AI techniques. It can predict the diabetes using its mathematical models and can improve the lifestyle of many people all over the world. Machine learning algorithms are not only capable of early prediction of diabetes but can also reduce the cost and burden on the health care system. These mathematical models can early predict the diabetes without going for any blood test or other type of diagnoses tests.

The life of many people can be safe if we can predict this disease with 100% of accuracy. Many researchers already working for diabetes predictions.

II. RELATED WORK

K. VijiyaKumar et al. in [1] applied ML models for prediction of diabetes. The Random Forest ML model was implemented and its performance was tested for diabetes. L. V. R. Kumari et al. [2] apply several ML algorithms for diabetes prediction. The paper summarises the prediction accuracy of ML algorithms for diabetes. C. Charitha et al. [3] predict the type 2 diabetes using machine learning algorithms. P. Cıhan and H. Coşkun [4] apply many ML models for diabetes prediction. The performance accuracy of different algorithms is found different and then the paper conclude some of the best ML algorithms for diabetes prediction.

N. Fazakis et al. [5] applied the concept of ML for prediction of diabetes. The major contribution was to predict type-2 diabetes using ML. predict the type 2 diabetes using ML models. The paper successfully predict the types 2 diabetes using ML. P. Nuankaew et al. [6] also applied the models of ML for type 2 diabetes prediction and able to apply ML for successful prediction of it. M. A. R. Refat et al. [7] applied deep learning models for prediction of the same diseases. Deep learning applied the concept of Neural Network and searcher deeper than the ML models for making predictions. S. K. Reddy et al. [8] also applied the ML models for diabetes prediction and acquire high prediction accuracy than the peer researchers in this area. V. Mounika et al. [9] successfully predict the type 2 diabetes. Most of these research are based upon some standard data sets of many medical symptoms for early prediction of this common disease [10] [11] [12] [13] [14] [15] [16] [17]. Even after a lot of research in prediction of diabetes, we are failing to stop this disease. So there is a need to do more research in prediction of diabetes.

III. PROPOSED WORK

This work uses the power of ML for prediction of diabetes. Figure 4 is depicting the proposed work for it. The python programming language is used in implementation of the ML models. After analysing the literature, we selected Logistic Regression (LR) and Random Forest (RF) classifier for diabetes prediction. In this work, the data about the diabetes patients is downloaded from Kaggle.Com. The data set is divided into training and testing part according to 80-20 rule. The eighty percent data of the data set is used to train the ML models and remaining twenty percent of the data is used to test the prediction accuracy of the ML models. Finally using the python programming language, the two models are implemented and their prediction accuracy is measured.

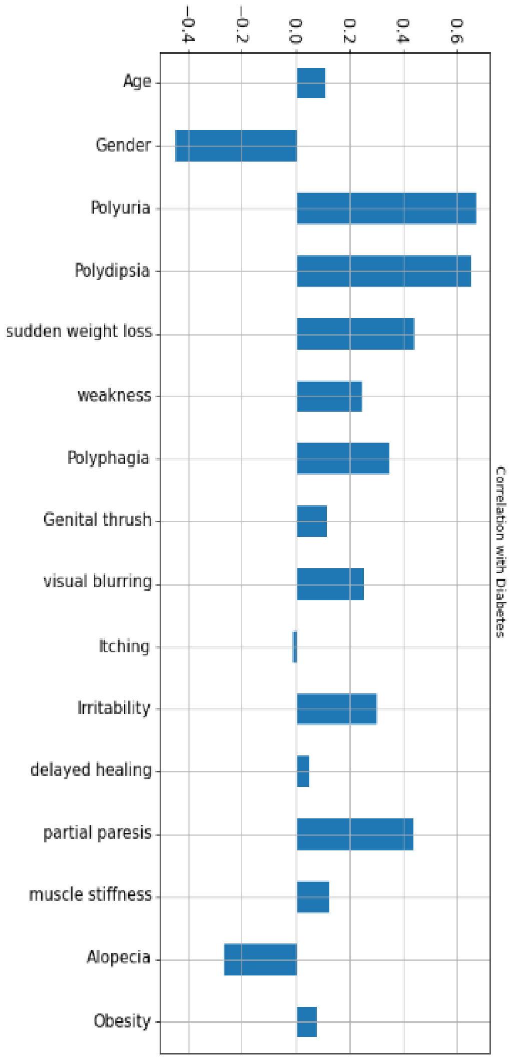

Fig. 1. Correlation between features of the data set.

A. *Machine Learning Models*

In this work two machine learning models (which are showing better results for diabetes prediction in the existing research work) are applied for diabetes prediction.

B. *Random Forest Classifier*

The RF algorithm creates a number of decision trees on different sub samples. This technique to create multiple trees

Authorized licensed use limited to: Khulna Univ of Engineering & Technology - KUET. Downloaded on March 30,2024 at 11:36:11 UTC from IEEE Xplore. Restrictions apply.

is to improving the prediction accuracy by controlling the over fitting problem.

C. Logistic Regression

It is a statistical tool to classify the probability of a target variable. The target variable should belong two one of any two possible cases. It is a supervised machine learning algorithm and give better results for binary classification problems.

D. Data set

The data set is freely available on Kaggle website. The data set contains the data about 520 individuals of age group 20 years to 65 years. The data of seventeen diabetic characteristics such as age, weakness, obesity etc. These characteristics are the risk factors which are responsible for diabetes. The data is collected by questionnaires of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

The data set is analysed by finding correlation between features of the data set. Figure 1 is showing the correlation between all features which are responsible for the diabetes.

Figure 2 is showing a bar graph of the distribution of the target variable. It depicts that how many instances in the data set are having positive diabetes and how many instances have negative diabetes. Out of 520 instances in the data set, 200 are found negative and 320 instances are found positive. Figure 3 is showing the pie chart of the target variable distribution.

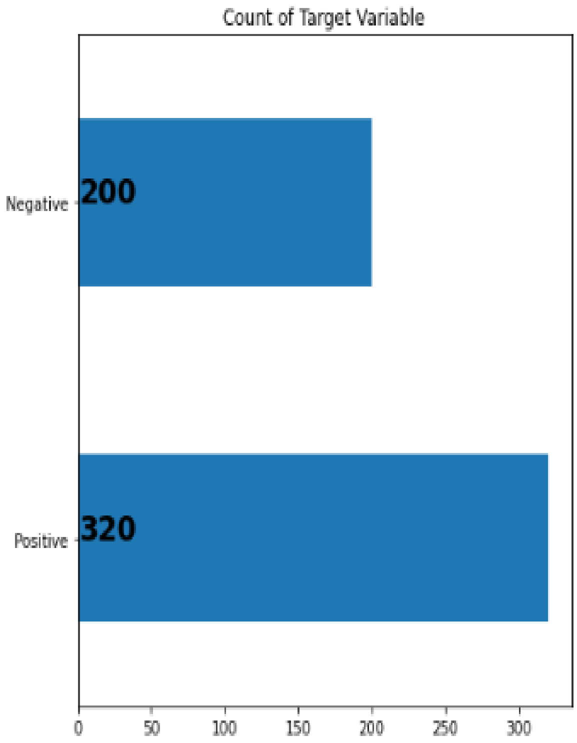

Fig. 2. Bar graph of the target variable in the data set

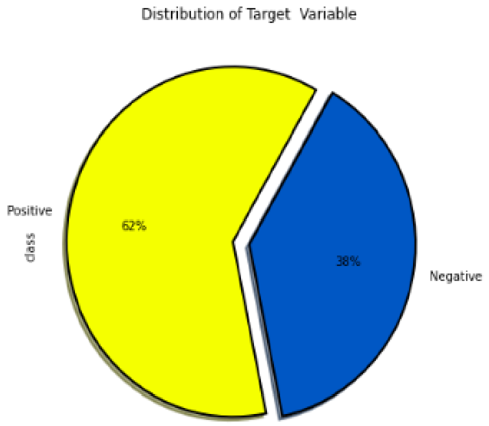

Fig. 3. Pie graph of the target variable in the data set

E. K-Fold Validation

In this work, we applied K-Fold validation for machine learning models. In K-Fold validation, the Ml model is validated K times where K is an integer. In this work, we take the value of K=10. Here the data set is divided in K parts and then training and testing will be conducted K times in a sequence.

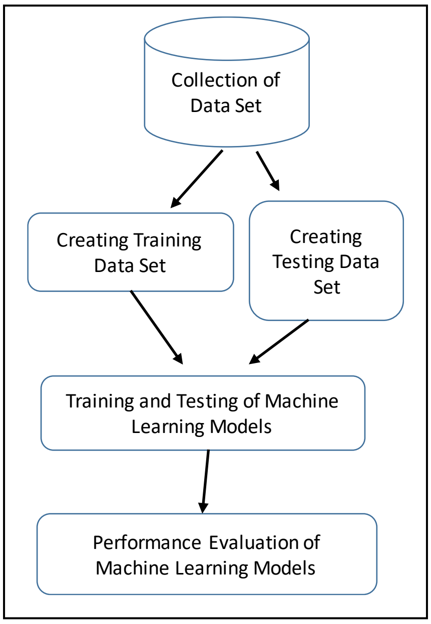
Collection of Data Set
Creating Training Data Set
Creating Testing Data Set
Training and Testing of Machine Learning Models
Performance Evaluation of Machine Learning Models

Fig. 4. Proposed model for diabetes prediction using ML

F. Decision Metric

In this work, the prediction accuracy is taken as the decision metric to evaluate the performance of the ML models. It is the ration between the two quantities i.e. the correct predictions made divided by total predictions made.

Authorized licensed use limited to: Khulna Univ of Engineering & Technology - KUET. Downloaded on March 30,2024 at 11:36:11 UTC from IEEE Xplore. Restrictions apply.

If the ML model made hundred predictions and out of which seven are incorrect and ninety-three are correct then its prediction accuracy is ninety-three percent.

IV. RESULTS AND ANALYSIS

In this work, RF and LR classifiers are proposed for prediction of diabetes. K-Fold validation is used to improve the performance for the prediction. The results of the python program are depicted in Table -1. The accuracy if LR model is found ninety-four percent and accuracy of RF model was around ninety-nine percent. From these results, the accuracy of RF model was almost five percent better than the RF model. The comparison of the results is shown in figure 5 using a bar chart graph.

TABLE I : Comparison of Prediction Accuracy

Sr. No.	**Machine Learning Algorithm**	**Accuracy of existing research S. Samet et al. 2021 [12]**	**Accuracy using Proposed models**
1	Logistic Regression	93.27%	94.23%
2	Random Forest	95.19%	99.03%

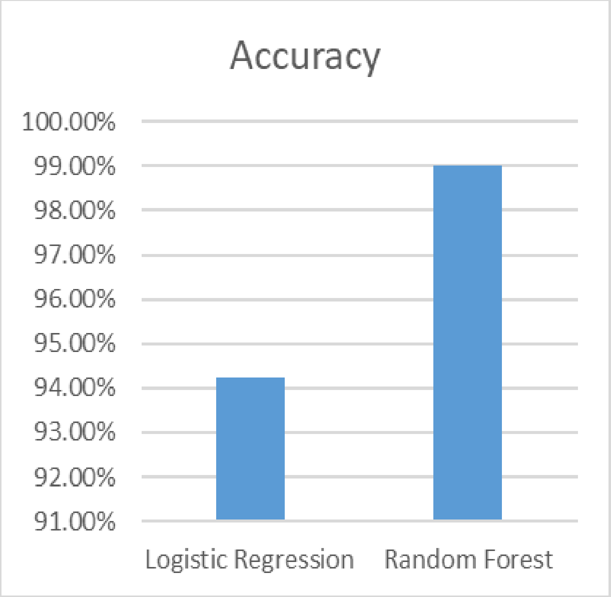

Fig. 5. Bar graph showing prediction accuracy

V. CONCLUSION AND FUTURE SCOPE

In this work, the prediction accuracy of machine learning models is evaluated. K-Fold validation is applied that improves the predication accuracy of ML models. The accuracy of RF model was found approximately ninety-nine percent which is far better than the other researchers [2][3]. In future RF model with K-Fold can also be applied for prediction of other diseases. The proposed K-Fold model can also be applied in future on other datasets for the diabetes.

REFERENCES

[1] K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," *2019 IEEE ICSCAN*, 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878802.

[2] L. V. R. Kumari, P. Shreya, M. Begum, T. P. Krishna and M. Prathibha, "Machine Learning based Diabetes Detection," *2021 ICCES*, 2021, pp. 1-5, doi: /10.1109/ICCES51350.2021.9489058.

[3] C. Charitha, A. Devi Chaitrasree, P. C. Varma and C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," *2022 ICCCI*, 2022, pp. 1-5, doi: 10.1109/ICCCI54379.2022.9740844.

[4] P. Cıhan and H. Coşkun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," *2021 SIU*, 2021, pp. 1-4, doi: 10.1109/SIU53274.2021.9477824.

[5] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," in *IEEE Access*, vol. 9, pp. 103737-103757, 2021, doi: 10.1109/ACCESS.2021.3098691.

[6] P. Nuankaew, S. Chaising and P. Temdee, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction," in *IEEE Access*, vol. 9, pp. 137015-137028, 2021, doi: 10.1109/ACCESS.2021.3117269.

[7] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin and M. K. Islam, "A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach," *2021 ISPCC*, 2021, pp. 654-659, doi: 10.1109/ISPCC53510.2021.9609364.

[8] S. K. Reddy, T. Krishnaveni, G. Nikitha and E. Vijaykanth, "Diabetes Prediction Using Different Machine Learning Algorithms," *2021 ICIRCA*, 2021, pp. 1261-1265, doi: 10.1109/ICIRCA51532.2021.9544593.

[9] V. Mounika, D. S. Neeli, G. S. Sree, P. Mourya and M. A. Babu, "Prediction of Type-2 Diabetes using Machine Learning Algorithms," *2021 ICAIS*, 2021, pp. 127-131, doi: 10.1109/ICAIS50930.2021.9395985.

[10] A. K. Uttam, "Transfer Learning-Based Approach for Identification of COVID-19," *2021 I-SMAC*, 2021, pp. 397-401, doi: 10.1109/I-SMAC52330.2021.9640956.

[11] A. K. Uttam, "Analysis of Uneven Stroke Prediction Dataset using Machine Learning," *2022 6th ICICCS*, 2022, pp. 1209-1213, doi: 10.1109/ICICCS53718.2022.9788309.

[12] Uttam, Atul Kumar, "Urinary System Diseases Prediction Using Supervised Machine Learning-Based Model: XGBoost and Random Forest", ICAISE 2022, pp 179—185, doi: 10.1007/978-981-16-8542-2_14

[13] A. K. Uttam, "Grape Leaf Disease Prediction Using Deep Learning," *2022 ICAAIC*, 2022, pp. 369-373, doi: 10.1109/ICAAIC53929.2022.9792739

[14] Uttam, A.K., Mangal, A., "Application of extreme gradient boosting ensemble model for sleep quality prediction on personalized wearable device data", IJAST, 2020, 29(5), pp. 3755–3762

[15] Mangal, A., Uttam, A.K., "Sleep prediction by various supervised machine learning model", IJAST, 2020, 29(5), pp. 3786–3792

[16] Hyuna Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", CA: A Cancer Journal for Clinicians, Volume71, Issue3, May/June 2021, Pages 209-249

[17] P. Saxena, S. Saha and S. K. Devi, "Analysis and Prediction of Diabetes Using Machine Models," *2022 MECON*, 2022, pp. 315-319, doi: 10.1109/MECON53876.2022.9751854.

Authorized licensed use limited to: Khulna Univ of Engineering & Technology - KUET. Downloaded on March 30,2024 at 11:36:11 UTC from IEEE Xplore. Restrictions apply.