TYPE-II DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS

[1]C. Charitha, [1]Amuluru Devi Chaitrasree, [1]Penmetsa Chidananda Varma

[2]Dr. C. Lakshmi

[1]B.Tech Students in Electronics and Communication Engineering, SASTRA Deemed to be University,
Thanjavur 613401,India

[2]Faculty in School of Electrical and Electronics Engineering, SASTRA Deemed to be University,
Thanjavur 613401,India

2022 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-6654-8035-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCCI54379.2022.9740844

Abstract—**Non-Insulin Dependent Diabetes Mellitus or Type2 Diabetes is one of the critical diseases and many people are suffering from it. Every year, approximately 2 to 5 million people are losing their lives as Diabetics. If Diabetes is predicted earlier, it can be controlled and also, deadly risks such as diabetes cardiac stroke, nephropathy and other disorders associated with it can be prevented. Therefore, early prediction of diabetes helps in maintaining good health. With the recent development in machine learning (ML), it is being applied to various aspects of the medical health. The Pima Indian Diabetes data set (PID), which was used in this paper, was acquired from the UCI repository. In this study, after undergoing a thorough data pre-processing and Feature engineering with feature importance models like Random Forest Importance and RFE, we used many Machine Learning models such as KNN, Logistic Regression, SVM, Random Forest , LightGBM and XGBoost for train-test splits like 60-40, 70-30 and 80-20 to predict Type-II diabetes mellitus . Among all models, the highest accuracy is obtained as 91.47% from lightGBM model for 80-20 train test split.**

Keywords-*Machine Learning, LightGBM, Diabetes Mellitus, Feature Engineering, Data Preprocessing*

I. INTRODUCTION

Diabetes is one of the critical diseases and many people are suffering from this. Hereditary diabetes, Lack of exercise, age, obesity, bad diet, high blood pressure, living style, etc., can cause Diabetes. Diabetics will have high risks like eye problem, heart disease, nerve damage, kidney disease, stroke, etc., Considering the current scenario, in some of the countries like India, Diabetes has become a very severe disease. Nearly 2-5 million patients every year lose their lives due to diabetes. It is said that by the year 2045 this will rise to 629 million. Diabetes Mellitus (DM) is generally classified as Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM) occurs due to Inability of human's body to generate sufficient amount of insulin. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM) occurs when body cells are not able to use insulin properly in human's body. Type-3 called as Gestational Diabetes results in the increase in blood sugar level in pregnant woman where early detection of diabetes is not done. Among these three types, Type-2 diabetes is the most common among the people than the other types as nearly 90% of diabetic people are of Type-2 diabetic according to Centre for Disease Control and Prevention (CDC).

Diabetes can easily be controlled if it willbe predicted early stages and risks associated with it can also be prevented with the early prediction. According to the global diabetes community, over 30 million have now been diagnosed with diabetes in India. The population in India is now more than 1000 million and the estimation over the actual number of diabetic people in India is around 40 million. It signifies that India has actually the highest number of diabetics than any other country in the world. IGT (Impaired Glucose Tolerance) is also becoming a mounting problem in India. It is actually thought that nearly 35 per cent of IGT sufferers go on developing type 2 Diabetes, so India is genuinely facing a crisis in healthcare. According to a recent study published by WHO, nearly 98 million people in India have type 2 Diabetes by 2030. Also, according to a survey 90% of the diabetic patients are of type 2 diabetic. In recent studies, Machine learning and Deep learning techniques have achieved reliable accuracy compared to existing works. Machine learning plays an essential role in healthcare field and is being increasingly applied to healthcare, where failure could be fatal. If a disease is diagnosed earlier, the more are the chances it can be prevented and triumphantly treated. So, it is important to get screened at early stages for diabetes and take steps to prevent it if you are identified to be at increased risk .So we mainly focused to work on detection of Type-II diabetes at an early stage using different ML algorithms.

II. LITERATURE REVIEW

There are several recent techniques that are being developed with evolution of Machine Learning technology for the diagnosis of diabetes.

The author in [1] has evaluated many ML algorithms for prediction of diabetes. The algorithms such as classification tree, SVM, ANN, logistic regression and KNN are implemented here. The performance of this system is appraised by different metrics namely specificity, Accuracy, recall, precision, FPR, negative prediction value, Rate of

978-1-6654-8035-2/22/$31.00 ©2022 IEEE

misclassification, F1 score and ROC curve.Logistic Regression gave best accuracy comparatively, 78% and rate of misclassification as 0.22 . The negative predictive values better precision came out to be as 73% and 82% using Logistic Regression and Naive Bayes respectively. And data is split using 10-fold cross-validation. And the author in [2] has used few ML algorithms namely LR, LDA, gradient boost classifier, Extra trees classifier, Ada boost classifier, Gaussian NB, Random Forest, Bagging, Decision tree, Perceptron, KNN, SVC for diabetes prediction.The resultant accuracies are achieved as 86% for DT, Gaussian Naive Bayes 93%, Linear Discriminant Analysis 94%, SVC 60%, RF 91%, Extra Trees classifier 91%, Ada Boost 93%, Perceptron Learning Algorithm 76%, Logistic Regression 96%, Gradient Boost Classifier 93%, Bagging 90%, KNN 90%. Among all the algorithms, LR gives the highest accuracy of 96%. With the accuracy of 98.8%, Ada Boost classifier became the best model on application of pipeline. After pipelining the results became better. The paper also used another data set namely diabetes dataset along with PIMA.

In this paper [3], titled Diabetes Prediction using MLT, algorithms such as DT, KNN, Gradient Boosting (GB), LR, RF and SVM are used and among them according to the results, RF achieved the highest classification accuracy namely 77% comparatively. In the paper [4], For conventional machine learning method, SVM,RF and for Deep Learning - fully Convolutional Neural Network (CNN) are the methods that are used. The accuracy obtained were DL (76.81%), SVM (65.38%), and RF (83.67%) .According to the results Random Forest was more effective for diabetes prediction compared to SVM and Deep Learning methods.

Coming to the paper [5], the dataset used in this paper is Luzhou’s hospital physical examination data from China and also PIMA Diabetes dataset. Minimum redundancy maximum relevance (mRMR), PCA and Random Forest algorithm are the algorithms used. Best performance is 0.8084 for Luzhou dataset, and the best result is 0.7721for Pima Indians, which indicates that ML can be used for prediction diabetes. In addition, we can find there is no notable difference among the algorithms used but RF is better comparatively.In this paper [6], KNN algorithm is used for the elimination of the unwanted data, thus reducing time for processing. However, the proposed classification approach is based on Decision Trees. On experimenting, proposed system has achieved high accuracy of 98.7% comparing to existing system using PID data set.

The author in [7] has proposed a system that uses principal component analysis and recursive feature elimination for diabetes prediction. They classified diabetes using ANN and DNN. On using DNN they got 82.67% accuracy and using ANN they got 78.62% accuracy.

III. Proposed Methodology

The motive of this paper is to find suitable ML model for Diabetes Prediction that provides better accuracy. This was done with different ML algorithms for Diabetes Prediction. The process we have gone through is briefly discussed below.

A. Description of the Dataset

From the UCI repository, the Pima Indian Diabetes Data set [8] was gathered. The data set has information of 768 patients and their corresponding features. This dataset consists of totally 9 attributes among which the last attribute is named as outcome, which indicates whether a person is diabetic or not where the outcome 1 indicates that the person is suffering from diabetes and 0 indicates that the person is non diabetic. And the remaining 8 attributes are independent factors corresponding to diabetes disease.

B. Preprocessing of Data

After analyzing all attributes and their significance in diabetes prediction, It is noted that many of the attributes have zero values in them and having zeroes in attributes such as Blood Pressure, BMI, Skin Thickness, age and Insulin is troublesome. And as the count of zero values in those columns is more, if we eliminate them we may lose large amount of data. So the zero values have been imputed with the mean of their respective columns. And the features are then scaled using Normalization technique to bring them to same scale, which is always advisable for applying Machine learning algorithms otherwise it fails to give the desired accuracy and to gain stability of the model as higher magnitude attributes affect the model output.

Then the data is checked for the outliers and notably the data consists of many outliers .As it also holds some information, only 2-5% of the outliers are removed from each column showing outliers. And removing some outliers is necessary as Machine Learning algorithms are sensitive to outliers. And the further process on the scaled and cleaned data is done by taking three different train test split cases which are 80%-20%, 70%-30% and 60%-40%.

C. Feature Selection

Feature selection techniques are mainly intended to reduce no. of independent variables to choose and select those that are believed to be really useful to a model in order to predict the desired one. Also it enables the ML algorithms to train fast as it decreases the complexity and makes it easier . In this study a correlation matrix is taken to check whether there is any notable correlation between any of the two attributes in order to eliminate one feature among them. But there is no notable correlation. Random Forest for feature importance and Recursive Feature Elimination with Random Forest Importance is the feature selection techniques used in this study and the collective results are taken.

1. Feature Importance - Random Forest(RF)

Using RF, the feature importance is measured as average impurity decrease which is computed from all the Decision

Trees in the forest.The feature importance from RF is calculated as a decrease in the impurity of the node which is weighted by the probability of reaching that node. Probability of it is calculated as the count of samples that reach that particular node, divided by total number of samples. And finally we can say that the larger the score the more important is the corresponding feature. The scores of different features is shown in Fig 1.

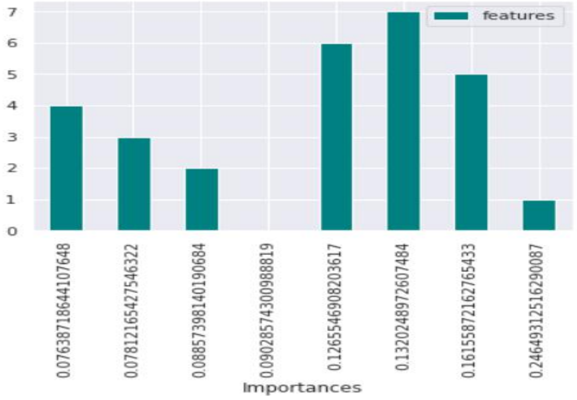

Figure 1. Feature Importance plot for Random Forest

2.FI - Recursive Feature Elimination-RFE

RFE is a wrapper method used in feature selection algorithms. In this method, a different ML algorithm will be chosen to use in the core of the method, here we have chosen Random Forest, which means it is wrapped up by RFE which will be used to select relevant details. RFE actually works by selecting a subset of features, and initiating all features in training data set then by deleting some features with least score till the desired number of features remain. This is actually achieved through fitting the chosen ML algorithm in the core of the model, Random Forest here, ranking the features by their importance, and leaving the least contributing features, finally by re-fitting model. The above process is repeated till we get specified number of features, 5 features here remain as shown in Fig 2.

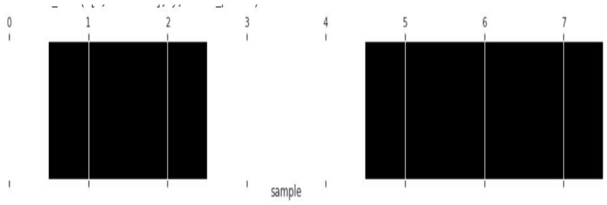

Figure 2. Feature Importance plot through Recursive Feature Elimination

According to the collective Information, Diabetes Pedigree Function, Glucose, Age, Blood Pressure and BMI are the features that are taken for the further process.

D. Machine Learning Algorithms

As the data and the features are ready, the next step is to apply machine learning techniques. In this paper, we used different ensemble and classification techniques present in machine learning to predict Diabetes at an early stage. The detailed information about the used techniques is as below.

1. K-Nearest Neighbor

k-NN is one of the supervised ML models.This model takes a set of input objects and output values(Labels). The model learns on the process of mapping inputs to their corresponding output with the help of training set. So that it could be able to make some predictions on the unknown data. It works by looking at k neighbor (closest labeled) data points for the given data point. And it's finally assigned to most similar category of the k closest points. This algorithm is applied with k cross fold technique in order to have optimal parameters, number of nearest neighbors in particular.

- Use only Testing data for finding accuracy for our KNN model.
- Now split again the Training data as k fold times and apply cross validation.
- Apply the KNN algorithm on the training data based on the results obtained with cross validation.
- Finally find the accuracy based on the mean obtained from the results of the k fold.

2. Random Forest (RF)

Random Forest is a famous ML algorithm, used for both Regression and Classification problems. It is an ensemble learning technique which comes under bagging, which is used to decrease the variance by generating extra data from training is a combination of many classifiers to improve the performance of the model. With a combination of many decision trees, the algorithm operates during training and assigns the mode of classification of individual trees as output.

- Firstly, take a set of random data from training set.
- Then, on association with the selected data, create decision trees.
- Give the number of your choice for the decision trees.
- Repeat step 1 and step 2 again.
- For the unknown data, find predictions of each DT, and then assign them to the category with majority votes.
- Finally, find resultant accuracy from chosen category.

3. Logistic Regression(LR)

LR is a familiar supervised ML classification algorithms, which is generally used to find the probability of the response, which is binary on the basis of atleast one predictor. They can be discrete/continuous. It is used to classify data into categories. And Logistic Regression usually classifies the data in binary form (0 and 1) which is used to distinguish whether a patient is affected or not affected with diabetes. The target of this model is to provide best fit for evaluating the relationship between predicted and target values.To predict probability of the output, sigmoid function is used, as it binds the output to either 0 or 1.

4. Support Vector Machine-SVM

It is a ML technique especially used for supervied learning and one of the most popular ML classification techniques. SVM generates a hyper plane which mainly performs the separation into two classes. Hyper plane will be used for classifying. SVM classify the entities and can also differentiate instances in specific classes. The hyperplane doess the segregation of any class to the closest training point.

- Firstly,we need to choose the hyper plane which performs division of the class better.
- The Margin, which is the distance between data points and plane needs to be calculated in order to find the better hyper plane.
- If the distance of seperation between the classes is high then there is a low chance for the miss conception and vice versa.
- So the class having high margin needs to be selected.

5. Light Gradient Boosting Machine-LightGBM

LightGBM, Light Gradient Boosting Machine, is a fast (hence the word 'Light'),distributed and performance wise high gradient boosting framework based on the decision trees used for the ranking,and many other ML tasks and classification to increase efficiency of model and reduces memory usage. As it is working on the basis of decision tree algorithms, with the best fit it performs leaf wise split whereas level wise or depth wise split will be chosen by other boosting algorithms. While growing on the same leaf in this framework, the leaf-wise algorithm will be able to reduce huge loss and henceforth it results in a much better accuracy which will be achieved rarely by any of the existing algorithms in the boosting category. Here this model has been used with suitable hyper parameter tuning.

6.XGBoost

XGBoost stands for eXtreme Gradient Boosting.This is a popular and efficient implementation of the gradient boosted trees algorithm.Actually, Gradient boosting is a supervised ML technique that attempts to accurately predict the target variable by the combination of an ensemble of estimates from a set of simpler and weaker models. And the beauty of this algorithm lies in its scalable ability which provides faster learning through distributed, parallel computing and offers an efficient usage of memory. In boosting technique, the trees are built such that each subsequent tree aims in reducing the former tree's errors. Each tree in this algorithm learns from its former ones and then updates the errors of their residuals. Hence, with an updated version from these residuals, the next growing tree in the sequence will learn. XGBoost algorithm emerged as most useful, robust and straightforward solution. It serves well its purpose in various machine learning competitions due to robust handling capability of a variety of distributions and tuning of a variety of hyper parameters.

E.Model

Model Building is the most important phase. Various ML algorithms discussed above are implemented for Type-II diabetes prediction at an early stage.

- Import PIMA data set and Import required libraries.
- Perform data preprocessing in order to cleanse the data and impute missing values.
- Divide the data as 80%-20%, 70%-30%, 60%-40% for Training and Testing respectively and perform the following process as three different cases.
- Select algorithms from K Nearest Neighbor, RF, Support Vector Machine, Gradient boosting, LR and Decision Trees.
- For the mentioned ML algorithm, build the classifier based on training data.
- For the mentioned ML algorithm, test the Classifier based on test set.
- Compare performance of experimental results obtained for each classifier.
- Conclude the best performing algorithm, after analyzing the results.

IV. RESULTS AND DISCUSSION

In the proposed work different steps were taken as shown above to achieve better accuracy and stability. The approach which was proposed uses different classification and ensemble methods of machine learning and they are implemented using python.In this work we see that Light Gradient Boosting Machine (LightGBM) achieved better accuracy compared to other algorithms especially for 80%-20% split. Overall some best Machine Learning techniques have been used for the prediction In order to achieve high accuracy in performance. The information in Table I below shows the result of these various supervised Machine Learning techniques for the various splits.

TABLE I. ACCURACY OF DIFFERENT MACHINE LEARNING ALGORITHMS FOR RESPECTIVE TRAIN-TEST SPLIT

MACHINE LEARNING ALGORITHMS	**Train - Test Split Cases**		
	60%-40%	***70%-30%***	***80%-20%***
Random Forest	78.74%	76.44%	77.95%

MACHINE LEARNING ALGORITHMS	**Train - Test Split Cases**		
	60%-40%	***70%-30%***	***80%-20%***
(RF)			
Logistic Regression	78.35%	76.96%	81.10%
XGBoost	75.59%	77.49%	81.10%
Support Vector Machine (SVM)	79.13%	78.01%	81.10%
K-Nearest Neighbors (KNN)	80.86%	79.66%	79.10%
LightGBM	84.61%	87.05%	91.47%

So from the accuracy table shown above, we can say that in all the train test split cases, LightGBM achieved highest testing accuracy. And the accuracy of the remaining algorithms in different train test split cases is provided.

V. Conclusion and Future Scope

The project was aimed to design and implement suitable ML model in order to predict Diabetes of Type 2 at an early stage. And our objective is achieved successfully after going through different Machine Learning models and with a conclusion that LightGBM model provided highest testing accuracy of 91.47%, with the help of suitable hyper parameter tuning. And the features we have chosen, the scaling mechanism we took, hyper parameters and the percentage of outliers we removed also affected our accuracy. Changing the above aspects may also leads to different results. The proposed approach uses various methods in which SVM, Random Forest(RF), XGBoost, KNN, LightGBM and Logistic Regression classifiers are used. And the highest classification accuracy of 91.47%, 87.04% and 84.61% has been achieved by LightGBM for 80%-20%, 70%-30% and 60%-40% Train-Test splits respectively. The obtained experimental results can be used to assist the health care in order to take an early prediction and also to make some early decisions to cure the diabetes and tp save numerous lives from the effects caused by Diabetes.

References

[1] A. Kumar Dwivedi, “Analysis of computational intelligence techniques for diabetes mellitus prediction,” Neural Comput. Appl., vol. 13, no. 3, pp. 1–9, 2017.

[2] Mujumdar, Aishwarya & Vaidehi, V.. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science. 165. 292-299. 10.1016/j.procs.2020.01.047.

[3] Mitushi Soni , Dr. Sunita Varma, 2020, Diabetes Prediction using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 09 (September 2020)

[4] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019,pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.

[5] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Front Genet*. 2018;9:515. Published 2018 Nov 6. doi:10.3389/fgene.2018.00515

[6] Kadhm, Mustafa & Ghindawi, Ikhlas & Enteesha, Duaa. (2018). An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. International Journal of Applied Engineering Research. 13.

[7] J. Vijayashree and J. Jayashree, “ An Expert System for the Diagnosis of Diabetic Patients using Deep Neural Networks and Recursive Feature Elimination,” International Journal of Civil Engineering and Technology, vol. 8, pp. 633-641, Dec. 2017.

[8] Pima Indian Diabetes Data Set, [Online]. Available: https: //archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes., accessed on May 01, 2018 Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.