International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

A Framework for Predicting Diabetes Using Ensembling of Machine Learning Classifiers

T.Rajasenbagam[1], A.Geetha[2]

Assistant Professor, Department of Computer Science & Engineering, Government College of Technology, Coimbatore, India[1]

Assistant Professor, Department of Computer Science & Engineering, Easwari Engineering College, Chennai, India[2]

ABSTRACT: Diabetes is a frequent illness in the modern world, affecting both industrialised and poor countries. It has the potential to damage blood vessels in the kidneys, eyes, heart, and nerves. Diabetes is also linked to the risk of a variety of ailments, including heart attack, stroke, and renal failure. It further has an effect on the blood vessels in the eyes, which may result to permanent eyesight loss. The objective of this project is to develop a system that reliably predicts diabetes in a patient by combining the findings of multiple machine learning algorithms for earlier detection of diabetes. By creating models from patient datasets, machine learning technologies improve prediction accuracy. In this paper, we predict diabetes using Machine Learning Classification and ensemble techniques on a dataset. It has been found that the ensemble technique provides a higher accuracy demonstrating that the model is capable of accurately predicting diabetes.

KEYWORDS: Machine Learning, K-fold cross-validation**,** Random Forest, XGBOOST, ADABOOST.

I. INTRODUCTION

Diabetes is a well-known term in today's world. The pancreas-produced insulin hormone in the body permits glucose to flow from food into the circulation of the blood. Diabetes arises when the pancreas fails to function, resulting in coma, kidney and retinal failure, harmful damage to pancreatic beta cells, heart problems, brain vessels dysfunction, peripheral vascular diseases, sexual dysfunction, ligament failure, loss of weight, ulcer, and harmful impacts on the immune system. Diabetes among adults (over the age of 18) has climbed from 4.7% to 8.5% from 1980 to 2014, corresponding to research on diabetes patients, and continues to rise in both third and second world countries. Although there is no long-term treatment for diabetes, it can be treated and prevented if an accurate forecast is achievable. Diabetes prediction is a difficult issue since the distribution of classes for all attributes is not linearly separable.

1.1 MACHINE LEARNING

A reliable framework for diabetes prediction was proposed which used outlier rejection, data standardisation, feature selection, K-fold cross-validation, and various Machine Learning (ML) classifiers (Random Forest, AdaBoost, XGBoost), and Multilayer Perceptron (MLP). In this literature, the weighted assembly of different ML models is also proposed to improve diabetes prediction, where the weights are calculated using the corresponding Area Under ROC Curve (AUC) of the ML model.

II. RELATED WORK

The ensembling classifier described by the author in [1] is the best performing classifier, outperforming the state-of-the-art findings by 2:00% in AUC and having the following performance metrics: sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC as 0:789, 0:934, 0:092, 66:234, and 0:950 respectively. With respect to other approaches, our suggested framework for the prediction of diabetes performs quite well. The achieved AUC value of just 0.9, which is less than 1, is the approach's disadvantage.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

In [2], the author suggested the Random Forest algorithm for the prediction of diabetes. The system was developed utilising the Random Forest algorithm and machine learning techniques to perform early diabetes prediction for a patient with a greater degree of accuracy. The results indicated that the prediction system is able to forecast the diabetes disease effectively, efficiently, and most significantly, quickly. The suggested model yields the best results for diabetic prediction. The categorization accuracy with this method is roughly 77%.

In [3], the author discussed forecasting the start of diabetes using an ensemble supervised learning approach. Five commonly used classifiers were utilised for the ensembles, and their outputs were combined using a meta-classifier. The findings are given and contrasted with those of earlier studies that made use of the same dataset. It is demonstrated that the proposed strategy can predict the start of diabetes with a low level of accuracy.

The author of [4] gave Diabetes Prediction with the goal of using Machine Learning Techniques is to predict diabetes using three different supervised machine learning techniques, including SVM, Logistic Regression, and ANN. This experiment suggests a useful method for identifying diabetes early on. According to prior study, the classification process has not significantly improved. Therefore, a system is needed to tackle the problems described based on prior research, as Diabetes Prediction is a key topic in computers. In many circumstances, an algorithm performs well in terms of speed but poorly in terms of data categorization accuracy.

III. METHODOLOGY

A new pipeline for diabetes prediction from the PIMA Indians Diabetes dataset is used in the proposed system which are shown in Figure 1, where outlier rejection, data standardisation, feature selection, K-fold cross-validation, various machine learning classifiers, and Multi-layer Perceptron were used.

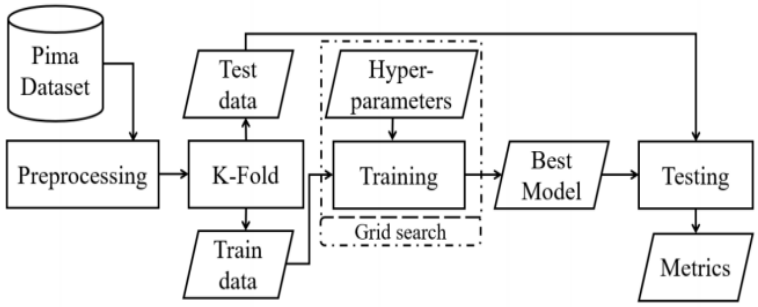

Figure 1: Overall architecture of the proposed method

3.1 DATASET

On a publicly available dataset of 2000 female diabetic patients from the Pima Indian population near Phoenix, Arizona, the ML models were trained and tested. This dataset includes 1316 non-diabetic individuals (negative) and 684 patients with diabetes (positive), each with eight unique features.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

3.2 FEATURES

The features of the dataset are shown in the Table 1

Attribute	Description	Type
Pregnant (F1)	Number of times Pregnant	Numeric
Glucose (F2)	Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test	Numeric
Pressure (F3)	Diastolic Blood Pressure (mm Hg)	Numeric
Triceps (F4)	Triceps Skin Fold Thickness (mm)	Numeric
Insulin (F5)	2-Hour serum Insulin	Numeric
BMI (F6)	Body Mass Index	Numeric
Pedigree (F7)	Diabetes Pedigree Function	Numeric
Age (F8)	Age in years	Numeric

Table 1: Features of the Dataset

3.3 DATA PREPROCESSING

The data is pre-processed to gain a better understanding of the data and make it ready for further analysis before any machine learning technique is applied. Outlier rejection, filling in for missing values, and feature selection are all included in the pre-processing stage.

3.4 CROSS FOLD VALIDATION

One of the most popular methods for model selection and classifier error estimation is the K-fold Cross-validation (KCV) methodology. K folds have been established in the PID dataset. In the inner loop, where the grid search technique was applied, the hyperparameters are trained and adjusted using K 1 folds. The model was assessed using the test data and the best hyperparameters in the outer loop (K times). Because the positive and negative samples in the PID dataset are unbalanced, the original percentage of samples for each class has been preserved using the stratified KCV. Figure 2 shows the partitioning of the PID dataset for KCV for both the hyperparameters tuning and evaluation.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

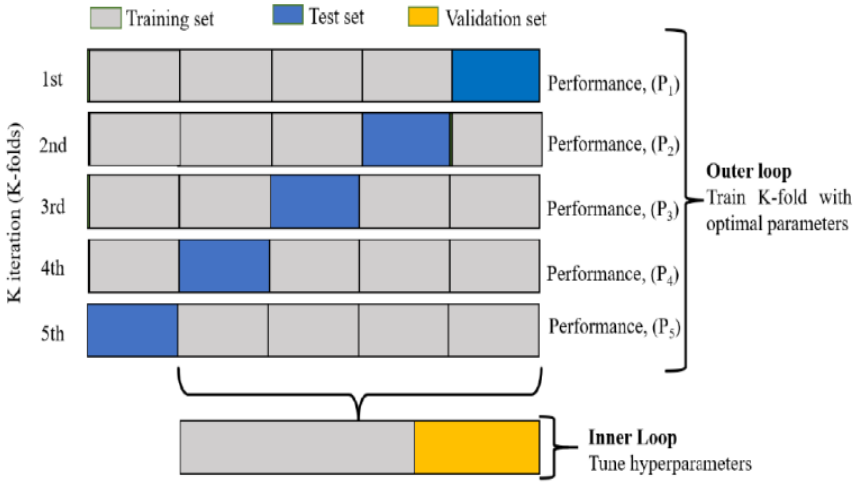

Figure 2: Partitioning of the PID dataset

3.5 HYPERPARAMETER TUNING

The process of fine-tuning the parameters that are present as tuples as we create machine learning models is known as hyperparameter tuning. These parameters are never learned by machine learning algorithms.

3.6 CLASSIFICATION

On the basis of the variables, datasets are clustered. Each clustered dataset is subjected to the classifiers in order to measure their performance. Based on the results mentioned previously, the models with the highest accuracy and lowest rate of error are determined to be the best performers. On the dataset, the classifier's performance is assessed for error reduction.

3.7 RANDOM FOREST

Random Forest is a classifier that uses many decision trees on different subsets of the input dataset and averages the results to increase the dataset's predicted accuracy. Instead than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

Algorithm : The Steps of Implementing Random Forest (RF)

Input: The n-dimensional data, $X \in \mathcal{R}^n$ and target outcome, $Y \in \mathcal{R}$

Output: The posterior probability, $P \in [0,\ 1]$ of unseen test data, x, where $\sum_{i=1}^{C} P_i = 1$ and $C = 2$ (diabetes present (C_1) or not (C_2))

1 **for** $b = 1$ *to* N ($n\_Bagging$) **do**

2 Draw a bootstrap sample, (X_b, Y_b) from given $(X \in \mathcal{R}^n, Y \in \mathcal{R})$

3 Grow a random-forest tree T_b using X_b and Y_b by repeating recursively using the following steps until the minimum node size is n_{min}.

1) Randomly select m variables from the given n variables
2) Pick the best variable or split-point among the m variables
3) Split the node into two daughter nodes

Output the ensemble of trees will be $\{T_b\}_1^N$

4 The posterior probability, $\hat{P}_{RF}^N(x) = Voting\{\hat{P}_k(x)\}_1^N$, where $\hat{P}_k(x)$ is the class prediction of the k_{th} random-forest.

3.8 XGBOOST

Gradient Boosted decision trees are implemented using XGBoost technology. In C++, this library was created. It is a kind of software library that was primarily created to increase model performance and speed. In recent years, it has dominated applied machine learning. Many Kaggle Competitions are dominated by XGBoost models. Decision trees are generated sequentially in this approach. Weights are significant in XGBoost. Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes.

Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes. Increased weight is applied to factors that the tree incorrectly anticipated, and these variables are subsequently fed into the second decision tree. These distinct classifiers/predictors are then combined to produce a robust and accurate model. It can be used to solve problems including regression, classification, ranking, and custom prediction. Supported gradient boosting methods include three primary types:

- Regularised Gradient Boosting;
- Gradient Boosting;
- Stochastic Gradient Boosting.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

Algorithm : The Steps of Implementing XGboost (XB)

Input: The n-dimensional data, $X \in \mathcal{R}^n$ and target outcome, $Y \in \mathcal{R}$
Output: The posterior probability, $P \in [0, 1]$ of unseen test data, x, where $\sum_{i=1}^{C} P_i = 1$ and $C = 2$ (diabetes present (C_1) or not (C_2))
1 Initialize the model with constant value:
$F_o(x) = \text{argmin}_\gamma \sum_{i=1}^{N} L(Y, \gamma)$ [32], where $L(Y, F(x))$ is the differentiable loss function and N is the number of sample
2 **for** $m = 1$ *to* M (*n_Iterations*) **do**
3 Compute pseudo-residuals, $r_{im} = -[\frac{\delta L(Y, F(X_i))}{\delta F(X_i)}]$, where $i = 1, 2, \ldots, N$
4 Fit a base tree, h_m using training set (X_i, r_{im}) for $i = 1, 2, \ldots, N$
5 Compute multiplier γ_m by
$\gamma_m = \text{argmin}_\gamma \sum_{i=1}^{n} L(Y_i, F_{m-1}(X_i) + \gamma h_m(X_i))$
6 Update the model by $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$
7 $F_m(x)$ is the desired posterior probability, $P \in [0, 1]$

3.9 ADABOOST

Machine learning ensemble methods use the Boosting methodology known as the AdaBoost algorithm, also known as Adaptive Boosting. The weights are redistributed to each instance, with higher weights given to instances that were mistakenly identified, hence the name "adaptive boosting." For supervised learning, boosting is used to lower bias and variation. It operates under the premise that students are developed in stages. Each student after the first is developed from a prior learner, with the exception of the first. Simply said, weak students are transformed into strong ones. Although there is a small difference in how it functions, the adaboost algorithm still operates on the same fundamentals as boosting.

3.10 MULTILAYER PERCEPTRON (MLP)

A neural network is made up of processing units, or neurons, which communicate with one another through unidirectional connections with varying weights. a feed-forward neural network, also known as an MLP, that has numerous hidden layers and an input-output layer.

3.11 ENSEMBLE MODELS

In order to minimise variance, bias, or enhance predictions, ensemble is a machine learning methodology whose methods are meta algorithms that integrate many machine learning techniques into one ideal predictive model. Comparing the predicted performance of this method to that of a single model, it is more accurate. The ML models in our system are combined to improve the accuracy of the diabetes prediction. Four ensemble models are produced by combining the ML models. The combination of AB and RF outperforms the other three combinations to get the best results for diabetes prediction.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

3.12 Experimental results

Eight characteristics are used in the development of the prediction models, and the modelling approaches' accuracy is calculated. Utilising the confusion matrix, the model is evaluated. Confusion matrix produces TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) as the total number of outcomes. Accuracy=(TN+TP) / (TN+TP+FN+FP) is the formula used to calculate the models' accuracy. Figure 4 demonstrates the output of MLP-Average Performance of All Folds and Figure 5 shows output of ensembling of random forest, adaboost and xg boost. Table 2 shows the performance obtained.

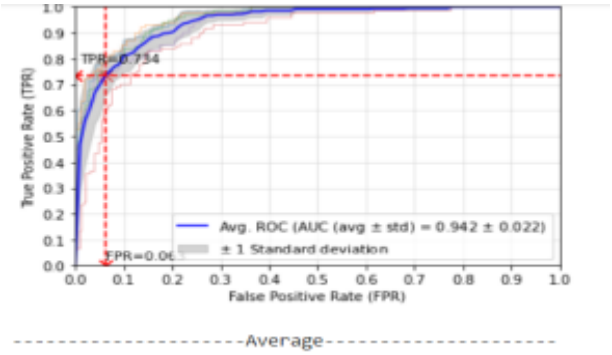

Figure 4: Output of MLP- Average Performance of All Folds

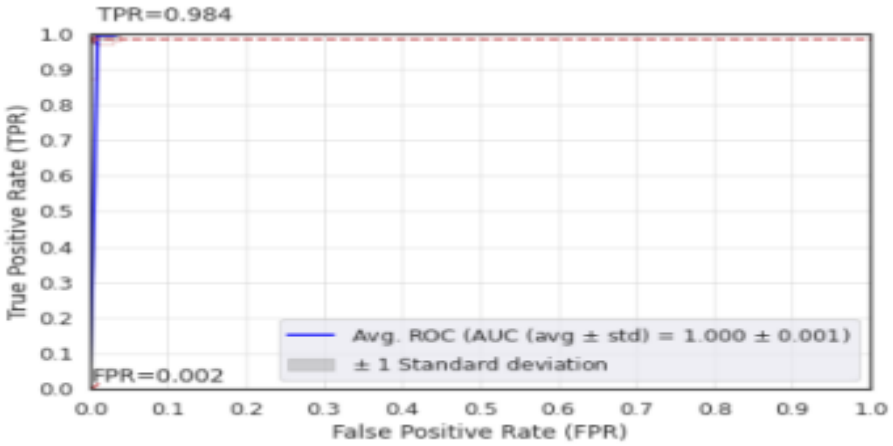

Figure 5: Output of ensembling of Random forest, Adaboost and xg boost

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

SL.NO	MODELS	ACCURACY	AUC	SENSITIVITY	SPECIFICITY
1	Random Forest	99.6%	1.00	99.6%	99.6%
2	AdaBoost	96.2%	0.998	90.1%	99%
3	XG Boost	99%	0.99	97%	99%
4	MLP	87.5%	0.94	73%	93%
5	Random Forest + AdaBoost	99.6%	1.00	99.6%	99.6%
6.	AdaBoost + XG Boost	99%	0.99	97%	99%
7.	RF+ XG Boost AdaBoost	99.4%	1.00	98%	99%

Table 2. Comparison of results of various models

IV. CONCLUSION

Pre-processing is essential for accurate and reliable prediction in the suggested technique, which uses the proposed ensemble model using the PID dataset to predict diabetes. The suggested pre-processing method, where outlier rejection and filling in missing values were of utmost importance, increased the quality of the dataset. A pre-processing like this can enhance the PID dataset's attribute distribution's kurtosis and skewness. While PCA and ICA only consider inter-attribute redundancy, correlation-based attribute selection might increase the correlation between an attribute and the desired conclusion. When using a classifier that uses trees as its basis, data standardisation cannot guarantee performance improvement. Using the 5-fold cross validation robustness assessment of the XB, MLP, and suggested ensemble classifier. The learning capacity of several classifiers, which were optimised using a grid search approach in our proposed framework, may be driven by their hyperparameters. AUC is a preferable weight to use when creating a general ensembling classifier since it gives more weight to models with higher AUC. When inter-class redundancy is substantial (not linearly separable), as it is in the PID dataset, random tree-based classifiers are well suited to classify the data. The comparison outcomes reveal that our suggested framework has done better than other frameworks on AUC, which has showed tremendous potentiality for diabetes prediction using the PID dataset. The optimum combination for predicting diabetes is the assembly of two boosting type classifiers (AB and XB), since the basis classifiers should have a low correlation with one another. When our suggested pre-processing (P + Q and correlation-based (feature selection) is used, the greater accuracy in diabetes prediction from the PID dataset utilising the optimal combination (AB +RF) may be attained.

REFERENCES

[1] MD. Kamrul Hasan 1, MD. Ashraful Alam1, Dola Das2, Eklas Hossain and Mahmudul Hasan, April 2020, “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers”, EKLAS HOSSAIN,vol.8, pp.76516-76531.

[2] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, 2019, “Random Forest Algorithm for the Prediction of Diabetes”, Vol.7, pp.11-15.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 4.988

| Volume 4, Issue 1, January 2021 |

[3] Nonso Nnamoko, Abir Hussain, David England, 2018, “Diabetes Onset: an Ensemble Supervised Learning Approach”, Vol.6, pp.22-26.
[4] Tejas N. Joshi, Prof. Pramila M. Chawan, January 2018, "Diabetes Prediction Using Machine Learning Techniques", Vol. 8 ,pp.09-13.
[5] Ridam Pal, Dr.Jayanta Poray, and Mainak Sen. Application of Machine Learning Algorithms on Diabetic Retinopathy. 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).
[6] Deeraj Shetty, Kishor Rit, Sohail Shaikh, and Nikita Patil.Diabetes Disease Prediction Using Data Mining. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).
[7] Rukhsar Syed, Rajeev Kumar Gupta, and Nikhlesh Pathik. An Advance Tree Adaptive Data Classification for the Diabetes Disease Prediction. 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE).
[8] K. VijiyaKumar, B. Lavanya, I. Nirmala,and S. Sofia Caroline. Random Forest Algorithm for the Prediction of Diabetes. 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN).
[9] Md. Kamrul Hasan,Md. Ashraful Alam, Dola Das, Eklas Hossain and Mahmudul Hasan. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers.
[10] Muhammad Azeem Sarwar,Nasir Kamal,WajeehaHamid, Munam Ali Shah.Prediction of Diabetes Using Machine Learning Algorithms in Healthcare.2018 24th International Conference on Automation and Computing (ICAC).