Proceedings of the Second International Conference on Automation, Computing and Renewable Systems (ICACRS-2023)
IEEE Xplore Part Number: CFP23CB5-ART; ISBN: 979-8-3503-4023-5

2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) | 979-8-3503-4023-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACRS58579.2023.10404311

Diabetes Prediction with Ensemble Learning Techniques in Machine Learning

1st S.Phani Praveen
Computer Science Department
P.V.P. Siddhartha Institute of Technology
Vijayawada, India
sppraveen@pvpsiddhartha.ac.in

2nd Vamsi Saripudi
Computer Science Department
P.V.P. Siddhartha Institute of Technology
Vijayawada, India
vamsi7514official@gmail.com

3rd Harshalokh.V
Computer Science Department
P.V.P. Siddhartha Institute of Technology
Vijayawada, India
harshalokhveeramachaneni@gmail.com

4th Sohitha.T
Computer Science Department
P.V.P. Siddhartha Institute of Technology
Vijayawada, India
tadikonda.sohitha3@gmail.com

5th Venkat Sai Karthik.S
Computer Science Department
P.V.P. Siddhartha Institute of Technology
Vijayawada, India
karthiksai131@gmail.com

6th Venkata Pavana Surya Sreekar.T
Computer Science Department
P.V.P. Siddhartha Institute of Technology
Vijayawada, India
tumulurisrikar@gmail.com

***Abstract*—Diabetes is a chronic disease caused by high blood glucose levels. Either the pancreas produces insufficient amounts of insulin or the body's cells stop responding to hormones. There is an increasing need for creative solutions in early prediction and healthcare management due to the rising global prevalence of diabetes. The article explores the field of predicting diabetes using machine learning techniques. The model covered in this article evaluates susceptibility to potential diabetes hazards in the future. With a focus on early identification and the reduction of the burden associated with complications related to diabetes, this study highlights the revolutionary potential of machine learning in healthcare. An ensembling approach for diabetes prediction is suggested in this paper. The best model is selected from a collection of diverse machine learning algorithms that includes Random Forest (RF), K-Nearest Neighbors (KNN), XGboost, and Catboost. Using metrics like accuracy, specificity, sensitivity, precision, false omission rate, and Area Under Curve (AUC), the performance of the developed model is assessed. The suggested model has a 96 percent accuracy rate after being trained on the PIMA Indian Diabetes dataset. This article adds to the continuing conversation about how artificial intelligence can be used in the healthcare industry to promote proactive health management and better patient outcomes.**

***Index Terms*—Ensemble, XGBoost, CatBoost, KNN, RF, Machine Learning**

I. Introduction

The latest studies by the World Health Organization (WHO) show an up-and-coming pattern in the number and fatalities of diabetic patients around the world. After analyzing these tendencies, the World Health Organization (WHO) predicts that diabetes will become one of the top 10 fastest growing illnesses worldwide by the year 2030. Diabetes is characterized as an assortment of metabolic sicknesses that outcomes in increased glucose levels in human blood. Diabetes is caused by two primary factors: inadequate secretion of insulin by the human body and impaired sensitivity of body cells to insulin. Insulin, an essential pancreatic enzyme, is responsible for regulating the levels of insulin in the circulation. Diabetes is classified into three distinct types: type-1, type-2, and gestational diabetes. [1], [8]. Type-1 diabetes is caused by immune cells attacking the pancreatic beta cells that produce insulin. Although it is difficult to prevent, it can be treated by administering insulin externally. On the other hand, Type-2 diabetes occurs when the insulin produced by the pancreas is not effectively used. [9]. The significance of this research lies in the fact that diabetes poses a substantial public health challenge, and timely detection is imperative in order to avert consequences. Hence, it is imperative to strengthen public consciousness of the hazards linked to diabetes and improve our capacity to forecast the probability of having diabetes. Diabetes arises from the dysfunction of the pancreas, resulting in an insufficiency of a vital hormone. This illness can lead to serious outcomes, such as the possibility of entering a coma, experiencing kidney and eye failure [13]. Furthermore, diabetes is associated with a range of problems, including cardiovascular dysfunction, cerebral vascular dysfunction, and others. Currently, clinical practice entails collecting the requisite data for diabetes detection through a range of tests. Several healthcare companies are currently incorporating machine learning techniques, such as predictive modeling, into their healthcare procedures. These techniques utilize complex algorithms to detect underlying processes and patterns that may be difficult for humans to perceive [19]. This facilitates researchers in the formulation of novel drugs and treatment strategies. Predictive modeling is based on the principles of data mining, machine learning, and statistics. Its purpose is to discover patterns in data and assess the probability of certain outcomes happening. Subsequently, the utilization of computational insight is tremendously valued for the prediction of diabetes as it helps in accurate predictions.
Diabetes prediction could be a vital use of machine learning and data science in healthcare. It involves using several

data sources and predictive modeling approaches. Predictive modeling approaches encompass classification techniques. Utilizing feature selection and engineering strategies improves the performance of the model [22]. Model evaluation employs many metrics such as accuracy, sensitivity, specificity, ROC-AUC, FOR, etc. Diabetes prediction models are useful in several clinical settings, such as identifying the disease at an early stage, tailoring treatment plans to individual patients, efficiently allocating resources, and facilitating research. The challenges encompass issues related to the accuracy of the data and the ability to understand and analyze it. Although faced with these difficulties, the prediction of diabetes is an essential element of preventive healthcare, facilitating timely detection and customized therapies.The goal of the research is to develop a thorough diabetes prediction model that will aid in estimating a person's likelihood of developing the disease. It is essential to mitigate further issues arising from diabetes, as it can give rise to several health complications if disregarded.There are various ML algorithms that are helpful for analyzing and synopsizing the information into important data. Machine learning includes different stages from preprocessing to testing and validation [14]. The PIMA Indian diabetes dataset is picked for this reason. To change over these information into suitable format, preprocessing is required. After the preprocessing is done the data is used to train the machine learning models (models like NN, RF, SVM, NB, AdaBoost, XGBoost, CatBoost). Now these models are tested for accuracy. In ensembling the prediction of each model is considered out of which, the class having maximum voting is chosen as the end prediction.

II. REALTED WORK

A.

In [1], the authors introduced a novel framework called "eDiaPredict," which employs a combination of machine learning techniques to predict the diabetes condition of patients. Initially, the dataset is subject to preprocessing, which includes addressing missing values and feature selection. The recursive feature elimination (RFE) method is employed, ensuring that only the most relevant features are retained for the prediction model. The results indicate that XGBoost stands out, achieving an impressive 92.21 percent accuracy when used in isolation. The paper also introduces an ensemble approach, which involves combining the top-performing models using a majority voting method. The authors provide an in-depth look at the specifics of this ensemble process, highlighting that XGBoost and Random Forest models are the best-performing combination. The evaluation of the eDiaPredict framework incorporates various performance metrics. This ensures that the results are evaluated without any discrepancies. The results demonstrate the effectiveness of the ensemble approach, with XGBoost and Random Forest achieving a remarkable 95 percent accuracy.However, the utilization of methods like Recursive Feature Elimination (RFE) has the risk of overfitting.

B.

In [2], the work centers around the creation of a diabetes prediction model utilizing machine learning techniques, namely logistic regression, to aid in the early identification of the disease. The logistic regression algorithm is used as the main method for constructing the prediction model. Two distinct feature selection techniques are utilized, which involve generating new features based on diagnostic measurements and employing univariate feature selection. The work investigates the utilization of ensemble approaches, that are normally used to amplify the execution of the prediction model. PIMA Indians Diabetes consists of 9 features, and Dataset 2, obtained from Vanderbilt, consists of 16 features, with one target variable indicating the presence of diabetes. Dataset 1 has newly generated features (NF1 through NF5) that are derived from diagnostic measurements pertaining to diabetes risk factors. This method improved the precision of predictions. Ensemble approaches, specifically Max Voting, Majority Voting and etc are evaluated on both datasets. The utilization of Max Voting demonstrated its superior efficiency, resulting in a substantial enhancement in performance. The accuracy of Dataset 1 increases to approximately 78 percent when the ensemble technique of Max Voting is applied. The accuracy of roughly 93 percent achieved by the ensemble techniques Max Voting and Stacking in Dataset 2 showcases the efficacy of ensemble approaches in improving prediction accuracy. The research highlights that the performance of prediction models depends on multiple factors beyond the choice of algorithm, emphasizing the significance of comprehensive data analysis and preprocessing in healthcare applications.

C.

In [3], the researchers presented a robust paradigm for predicting diabetes. The core of this methodology is based on rigorous data preparation. To enhance the dataset's quality, the authors utilized outlier rejection techniques to detect and eliminate data points that exhibited large deviations from the mean. To handle missing values in the dataset, they are imputed with the mean values. This ensures that important data is preserved and not lost in the process. The study investigates various machine learning models for predicting diabetes, such as KNN, DT, AdaBoost, RF, Naive Bayes, and Extreme Boosting. In order to guarantee the dependability of the results, a K cross-validation methodology is utilized. The performance metrics sensitivity, specificity, precision, FOR and DOR are employed to evaluate the models' potential to accurately detect positive and negative cases, as well as the usefulness of the diagnostic test. The authors emphasize the significance of choosing models with minimal correlation in order to attain improved ensemble performance. This study specifically examines the integration of Adaptive Boosting (AB) and Extreme Boosting (XB), demonstrating the significant efficacy of this combined strategy in predicting diabetes. The (AB+XB) model attained a peak efficiency of 95 percent. This work suggests potential future paths, such as creating a user-friendly web application using the trained model and exploring its

applicability and adaptability in forecasting various diseases in different medical settings.

D.

In [4], the study aims to evaluate and contrast the effectiveness of two machine learning algorithms, namely KNN and Naïve Bayes, in the study related to diabetes. The data utilized for this research came from the Pima Indians Diabetes Database, which comprises 768 records featuring eight variables and two outcome classifications (0 or 1). The authors utilize K-Fold Cross Validation, which involves dividing the dataset into training and testing data using different proportions ranging from 80 percent to 10 percent. The paper provides a comprehensive analysis of the accuracy, precision, and recall of both systems. The results indicate that in multiple experiments, Naïve Bayes outperforms KNN, achieving an average accuracy of 76.07 percent compared to KNN's 73.33 percent.The study findings indicate that when using the Pima Indians dataset, the Naïve Bayes algorithm is the most advantageous choice for predicting diabetes. To improve the quality of diabetes prediction, authors of this study propose using alternative methods, such as neural networks, and investigating optimization approaches like Particle Swarm Optimization in future studies. This paper conducts an extensive comparative analysis of KNN and Naïve Bayes, with the conclusion that the latter is more suitable for the prediction of diabetes.

E.

In [5], the study centers on the development of a model for assessing the risk of diabetes by utilizing data obtained from community follow-up. The study utilizes an extensive dataset of 252,176 subsequent records of individuals with diabetes, spanning from 2016 to 2023. The main goal is to investigate the correlation between important lifestyle markers derived from community follow-up data and the likelihood of developing diabetes. The authors utilized machine learning techniques, specifically the random forest classifier, to develop a model for assessing the risk of diabetes. Feature selection approaches are employed to find the most essential indications, resulting in a model with a high accuracy rate of 91.24 percent and an AUC of 97 percent. In order to enhance the applicability of the model in clinical settings, the study presents a diabetes risk score card that enables prompt identification by community follow-up doctors and self-assessment by patients, resulting in a precision rate of 95.15 percent. The proposed model exhibits the capacity for extensive risk screening on a wide scale, providing early warnings and enabling individual patient self-assessment. The proposed model and risk score card possess the capacity to support the early identification of diabetes in the patients, thereby tackling the escalating worldwide occurrence of diabetes. A future study could entail augmenting a range of features and incorporating time series analysis of subsequent data to improve the precision of the model and provide better assistance for patients in managing their own health.

F.

In [6], the study introduced a new technique called Average Weighted Objective Distance (AWOD) for diabetes diagnosis. The primary objective was to develop a predictive model that incorporates individual health issues and utilizes parameters that indicate such conditions. The study utilized two datasets to assess the effectiveness of the AWOD method compared to other machine learning approaches such as KNN, Support Vector Machines, Random Forest, and DL. The AWOD technique is grounded in the principle that individuals possess a wide range of health conditions and seeks to prioritize factors based on their influence on health outcomes. The suggested method underwent testing on two datasets: the PIMA dataset and the MD for Diabetes dataset. Each dataset consisted of 392 records. The findings indicated that the AWOD technique had superior performance compared to other machine learning methods, achieving accuracy rates of 93.22 percent. The prediction performance was assessed using various metrics. The study emphasizes the potential of the AWOD technique for accurately predicting type 2 diabetes, even in situations where individual health factors can differ considerably. Nevertheless, certain constraints, such as the intricacy of calculations and the configuration of parameters, could be resolved in future research.

G.

In [7], the primary objective is to employ data mining techniques for the early detection of diabetes. The study utilized a dataset obtained from the National Institute of Diabetes to predict the occurrence of diabetes by examining various diagnostic indicators. Principal component analysis was employed to perform data reduction by extracting important attributes from the dataset. The notable variables comprised glucose, BMI, diastolic blood pressure, and age. The dataset was analyzed using association rule mining to identify patterns and frequent items. The Apriori method was utilized to derive rules, and the investigation revealed robust correlations between diabetes and variables such as blood glucose, blood pressure, age, and BMI. The early prediction of diabetes involved the utilization of three distinct modeling techniques: Artificial Neural Networks (ANN), Random Forest, and K-means clustering. The Random Forest model attained a precision rate of 74.7 percent along with an AUROC value of 0.806. The Artificial Neural Network (ANN) model demonstrated superior performance compared to other approaches, with an accuracy of 75.7 percent. The K-means clustering achieved an accuracy rate of 73.6 percent and yielded an AUROC value of 0.816. The accuracy of this procedure was relatively lower compared to the other two methods used. The study highlighted the efficacy of machine learning and data mining methodologies in predicting diabetes. Further research could entail the inclusion of supplementary factors, such as sedentary lifestyle, familial predisposition to diabetes, and tobacco consumption, and the application of these methodologies to other areas of medicine.

III. PRELIMINARIES

Here, we present a concise overview of the many Machine Learning models and techniques used in the suggested framework.

A. K-Nearest Neighbors

K-Nearest Neighbors (K-NN) is a machine learning technique employed for the purposes of classification and regression. Utilizing distance metrics, this algorithm gauges the proximity of data points within the feature space. The K-NN algorithm assigns a class label to a data point through a majority vote among its K nearest neighbors. It uses the average of K data points' closest neighbors to determine the class to which that point belongs.

B. Random Forest

The Random Forest Classifier is a machine learning algorithm renowned for its exceptional accuracy and robustness. It is an approach in machine learning that combines the predictions of numerous decision trees to provide a final prediction. The technique employs bootstrap aggregation and feature randomization to mitigate the problem of overfitting [18]. The approach is beneficial for classification problems that involve a large number of variables and noise data.

C. XGBoost

XGBoost is a higly useful ML approach that is widely recognized for its outstanding performance in complex case studies. The algorithm constructs an ensemble of weak learners in a sequential manner, employing regularization approaches to avoid overfitting. The method employs tree pruning to restrict the depth of individual decision trees and offers efficient cross-validation capabilities. XGBoost is specifically designed to efficiently process tasks in parallel, effectively handle missing values, and include integrated functionality for selecting relevant features [23]. It is well-suited for processing huge datasets and utilizing multicore CPUs, and it has the capability to manage missing data.

D. CatBoost

CatBoost is a machine learning technique developed by Yandex that supports categorical features and uses gradient boosting. It is open-source and can be affected according to requirements without any legal constraints. Its effectiveness in predictive modeling tasks, especially with tabular data, is well established. CatBoost employs gradient boosting, built-in handling of categorical data, regularization methods, and ordered boosting to build trees. It is renowned for its ability to do computations swiftly, handle missing data automatically, and ensure robustness.

E. Stratified K-Fold Cross Validation

Stratified k-fold cross-validation is a machine learning method employed to assess the effectiveness of a model in datasets that are imbalanced or non-uniform. The process involves partitioning the dataset into k folds of equal size, doing model training and testing on each fold, and assessing the performance metrics. The procedure includes dividing the data, doing training and testing, subsequently obtain the comprehensive and mean performance of the model.It is beneficial for models that have an imbalanced class distribution. It helps in making informed judgments on model selection and fine-tuning of hyperparameters.

F. Ensemble Methods

Ensemble methods represent a powerful strategy within machine learning, where they amalgamate the forecasts of multiple models to produce a final prediction that is not only more resilient but also more precise . These techniques are extensively employed to enhance predicted accuracy, enhance model robustness, and minimize overfitting [10]. Ensemble approaches are highly efficient when individual models possess complimentary strengths and weaknesses.

1) Bagging: Bagging is an ensemble methodology in machine learning that entails generating numerous base models by training them on distinct subsets of training data, employing a method known as bootstrapping. The objective of this strategy is to minimize the variance and minimize the problem of overfitting, thereby enhancing the accuracy of predictions and the robustness of the model.

2) Boosting: Boosting is a method in ML that uses ensembling that amalgamate several models outcomes to make a accurate final decision. Boosting algorithms employ weak learners, sequential learning, weighted data, and weighted voting to perform classification or regression tasks.

3) Stacking: Stacking is an practice in machine learning that involves training a large number of models and aggregating their predictions using a meta-model to generate a final prediction. It utilizes a variety of various models to enhance forecast accuracy and the resilience of the model. The meta-model generates predictions by employing k-fold cross-validation.

IV. PROPOSED FRAMEWORK

This study involves the application of ensemble learning techniques, specifically KNN, RF, XGBoost, and CatBoost, to predict the risk of an individual developing diabetes using the PIMA dataset. The procedure begins with data preprocessing, involving the identification and imputation of missing values as well as the elimination of outliers by using the interquartile range. The models are trained and ensembled based on voting classifier and by averaging predicted probabilities. The model's robustness is assessed using thirty-fold cross-validation. Fig. 1 provides an elaborate elucidation of the procedure.

A. Data Preprocessing

The PIMA Indian Diabetes dataset has 768 records, with 268 individuals diagnosed with diabetes and 500 individuals without diabetes [24]. The information is openly accessible and seeks to offer a thorough comprehension of diabetes. The dataset has eight attributes:

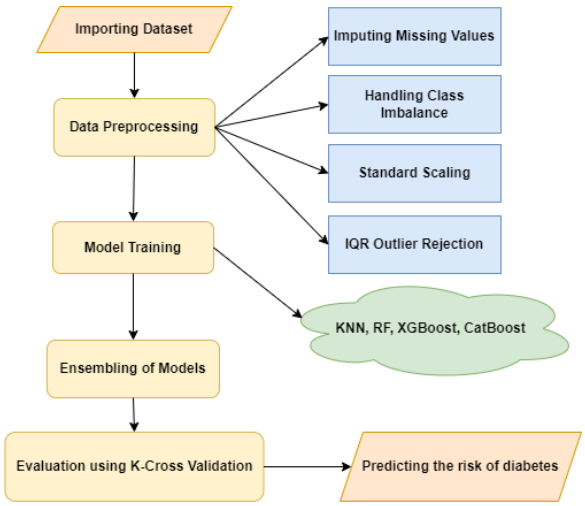

Fig. 1. Proposed System.

1) Pregnancies: Indicates the number of pregnancies experienced.

2) Glucose: Provides information on the concentration of glucose in the plasma after a two-hour oral glucose tolerance test.

3) BloodPressure: It is quantified using the unit millimeters of mercury (mm Hg).

4) SkinThickness: It refers to the measurement of the thickness of the skin fold on the triceps muscle, expressed in millimeters.

5) Insulin: Provides the measurement of 2-hour serum insulin in milli units per milliliter (mU/ml).

6) BMI: The calculation involves obtaining the individual's weight and height.

7) Age: It refers to the number of years that a person has lived.

8) DiabetesPedigreeFunction: The Diabetes Pedigree Function assesses the probability of developing diabetes based on the individual's age and their family history of diabetes.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Fig. 2. Sample records in the dataset.

Addressing missing data is an essential stage in the data preprocessing process for machine learning tasks. The presence of missing data can greatly influence the effectiveness of machine learning models, making it crucial to comprehend the methods for handling them [11]. Any feature, excluding from pregnancy, that has a value of zero should be regarded as a missing value. It has been noted that all features, except for age and pedigree function, have zero values.The box plot in Fig. 3 gives a visual summary of data distribution in the dataset.

It is observed that there is a class imbalance in the given dataset. The number of instances with label 0 is twice as large as that of 1. So performing mean imputation may induce a bias that is in favor of the majority class. To prevent this, the minority class instances are imputed with only minority class data, and the majority class instances are imputed with majority class data, which is commonly referred to as class-aware imputation.Standard Scaling is a higly employed data preprocessing method applied to datasets. This procedure entails converting the characteristics of a dataset in such a way that they possess a mean (average) with a value 0 and a standard deviation value of 1. This process makes it sure that all the features in the collection of data are on a congruous scale [20]. Normalizing the input features is crucial, especially when dealing with machine learning algorithms that are highly influenced by the magnitude of the features [15]. Standard scaling eliminates biases and offsets in the data, enhances the convergence of optimization techniques, and ensures an equal contribution of all features to model predictions.

$$z = (x - \mu)/\sigma \tag{1}$$

z is the standardized value.
x is the original data point.
μ is the mean of the feature.
σ is the standard deviation of the feature.

The IQR outlier rejection approach is employed to identify and address outliers within a dataset. This strategy is highly effective in identifying and potentially removing outliers. It is computed by subtracting the value of the first quartile of the feature from the of its third quartile in the dataset, and is performed for all the features.

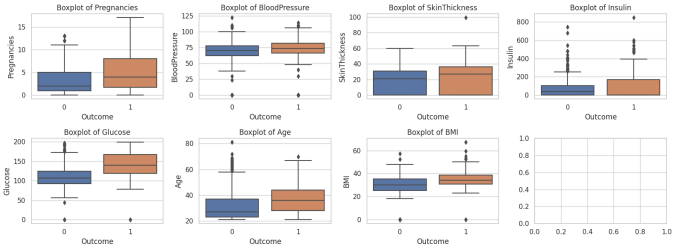

Fig. 3. Box plot of features.

The data correlation plot in Fig. 4 is used to understand the association between the features in the dataset. It helps understand how changes in one variable are related to changes in another.

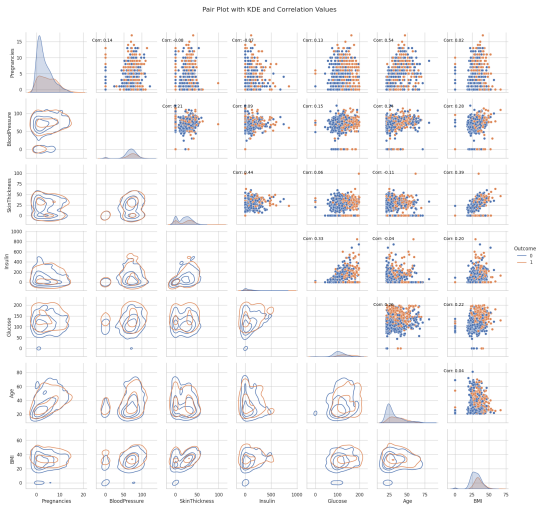

Fig. 4. Correlation plot of the dataset features.

B. Training of Models

The process of training and testing models is essential to the creation of the ML models. These procedures involve the division of data into training and test sets. The process of model training involves the careful selection of a suitable machine learning algorithm or model for the given job, followed by training the model and tuning hyperparameters to enhance performance [21]. The model's performance is evaluated using performance metrics that were mentioned in the previous sections of the study. After the model has undergone training, testing, and evaluation, it can be implemented in practical scenarios to generate predictions based on fresh data. Efficient model training and testing necessitate a careful combination of data preparation, model selection, hyperparameter tuning, and performance evaluation. This process may involve multiple iterations to refine the model and achieve optimal results.
The KNN model is trained with the parameter 'n_neighbors' set to 5. It utilizes the majority class among its five closest neighbors to make predictions. The RF model is prepard by setting the 'n_estimators' parameter to a value of hundred. Voting is done on the results of the 100 decision trees, and the class with highest number of votes is chosen. XGBoost is trained by setting the 'max_depth' parameter to 3 and 'n_estimators' to 15000. In CatBoost, the model is trained by setting the 'iterations' parameter to 900 with a 'depth' of 5. Among the models, CatBoost and XGBoost are better when contrasted with the other models that were used.

C. Model Ensembling

In this stage, the previously trained models are ensembled to get improved results. A soft voting classifier is employed as it combines the results of the selected models by calculating the mean of their occurence. For a classification task, each base model generates a probability distribution for the possible classes. The final prediction is determined by averaging these probabilities [12]. This strategy is generally more robust and frequently results in enhanced precision compared to hard voting, where each model's vote is considered a distinct class. KNN and RF are combined into an ensemble. By combining them, you are utilizing their complementary traits. The KNN algorithm utilizes the proximity of data points to produce predictions, enabling it to capture local patterns [17]. On the other hand, the Random Forest (RF) algorithm, which consists of a collection of decision trees, has the ability to capture intricate global patterns. Considering the strong individual performance of XGBoost and CatBoost, they are ensembled with other models to combine the predictions in order to attain the final results.

V. Results and Analysis

This section presents a detailed analysis of the outcomes obtained from the implemented framework. The aspects encompassed in this are the criteria for evaluation, the presentation of study results, the visual representation of data, and the discussion of performance.

A. Criteria for Evaluation

The models' performance is evaluated based on the following criteria:

1) Accuracy: Accuracy is defined as the proportion of correctly classified instances out of the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

2) Precision: It refers to the proportion of accurately predicted positive events to the total number of events that were identified as true.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

3) Sensitivity: The measure quantifies a model's capacity to accurately identify positive occurrences from the entire set of actual positive instances.

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

4) Specificity: This statistic measures the model's capacity to accurately detect true negative situations out of all the real negative instances.

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

5) False Omission Rate (FOR): The False Omission Rate (FOR) quantifies the tendency of a model to incorrectly categorize negative cases as positive.

$$FOR = 1 - Specificity \tag{6}$$

6) AUC-ROC (Receiver Operating Characteristic): The AUC-ROC score evaluates a model's capacity to distinguish between the correctly classified postive classes to wrongly classified positive classes, where a score above 0.5 signifies remarkable performance.

B. Study Results

Initially, models are trained separately and evaluated using metrics such as accuracy, precision, sensitivity, specificity and FOR. The models underwent 30-fold operations of cross-validations. Here the dataset is partitioned into 30 subsets and the training and testing process is repeated 30 times. This approach guarantees that the performance of a model is sufficiently trained and assessed, thereby detecting possible problems like overfitting or underfitting. The procedure involves data preprocessing, partitioning into 30 subsets, training and testing, collecting performance metrics, evaluating the model, fine-tuning parameters, and ultimately training the final model.

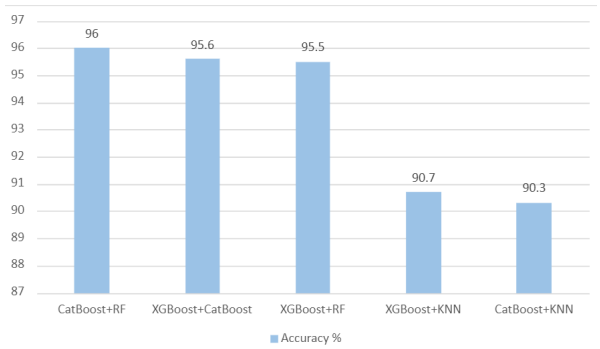

Fig. 5. A comparision of ensembled models performances.

This approach yielded a more robust assessment of a model's performance. CatBoost has demonstrated superior performance compared to other models, with an accuracy rate of 95%. The method of data preparation and the optimization of parameters had a significant impact on this outcome which were discussed in section IV-B. Given that CatBoost and XGBoost produced better results independently, they were ensembled with the other models in order to determine the model that exhibited the highest level of efficiency.

The Fig. 7 gives an illustration of peformance metris of ensembled models.The evaluation of several ensemble models revealed a spectrum of accuracies, providing insights into the prediction capacities of various combinations of machine learning algorithms [25]. The combination of CatBoost and RF yielded the most optimal results, with an accuracy rate of 96%. The XGBoost and CatBoost combination demonstrated a strong accuracy of 95.6%. The ensemble model consisting of CatBoost and K-Nearest Neighbors (KNN) achieved an impressive accuracy rate of 90.7%. The XGBoost and Random Forest (RF) ensemble obtained an accuracy of 95.5%, just behind the leading models. Meanwhile, the XGBoost and K-Nearest Neighbors (KNN) combination displayed a consistently accurate performance of 90.3%.

These results clarify the different degrees of performance exhibited by ensemble models and provide significant insights for making informed decisions about the most appropriate technique for predicting diabetes in patients .

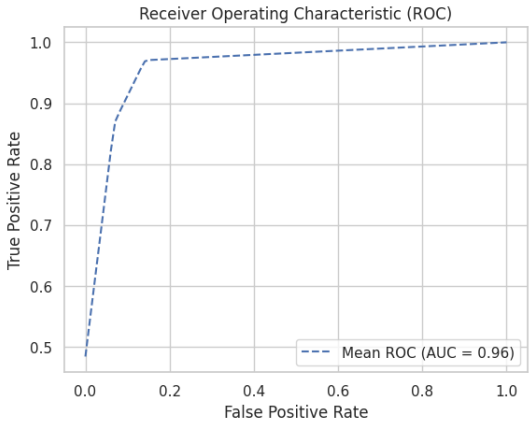

Fig. 6. ROC Curve for CatBoost and RF Ensemble.

The suggested framework demonstrates the appropriateness of the models based according to the performance metrics that were chosen for this research. The CatBoost algorithm, along with the random decision forests model, achieves the highest performance in properly predicting the risk of diabetes in patients with a 96% correctness. The Fig. 6 depicts the receiver operating characteristics of the CatBoost+RF model, which attained the greatest accuracy among all the ensembled models.

Additionally, these models have greater values for several performance measures. CatBoost utilizes gradient boosting, integrated handling of categorical data, regularization techniques, and ordered boosting to construct trees. The integration of models offers significant advantages by consolidating the capabilities of many selected models and directing them towards a specific job required to accomplish the ultimate objective of the work [16].Previous studies have demonstrated that achieving 100% effectiveness is unattainable for any model. However, doing so can undoubtedly be advantageous for individuals based on their specific needs. The RF model operates by randomly generating several decision trees and making predictions based on the collective voting of these trees.

Model	*Accuracy*	*Precision*	*Sensitivity*	*Specificity*	*Mean AUC*	*FOR*
CatBoost + RF	96	95.5	96.7	95.1	0.96	0.031
XGBoost + CatBoost	95.6	94.9	96.8	94.4	0.98	0.020
XGBoost + RF	95.5	95.4	95.9	95.1	0.95	0.040
CatBoost + KNN	90.7	93.9	87.2	94.2	0.91	0.115
XGBoost + KNN	90.3	93.6	86.6	93.9	0.90	0.121

Fig. 7. Performance metrics of ensembled models.

Ultimately, the final tree is depicted according to the voting outcomes. These ensemble models can be efficiently implemented in both distributed and memory-constrained situations.The findings of this study demonstrate that the suggested model surpasses current diabetes prediction models in the

perspective of accuracy and precision.Previous research have employed recursive feature removal, a technique that has the potential to result in overfitting. On the other hand, our method circumvents this problem by taking into account all features. In addition, a vital component of our study entails the utilization of "class-aware imputation," as elaborated upon in section IV-A, to safeguard against the introduction of bias into the data. In addition, the integration of very sophisticated machine learning algorithms has greatly improved the accuracy of our study. . The outcomes of our work indicate, suggested model holds great potential in enhancing the efficiency of diabetes prediction systems. Additionally, this model can be utilized with different datasets, hence improving the quality of diabetes diagnosis.

VI. Conclusion and future work

This study presents a predictive framework that aims to accurately anticipate the risk of diabetes by conducting a thorough review of important health parameters. Significantly, we found that our methodology attained exceptional outcomes, as the CatBoost algorithm combined with Random Forest (RF) showcased an exceptional individual accuracy rate of 96%. We anticipate that our framework will be further validated in the future using real-world clinical data to ensure its robustness in a variety of healthcare contexts. Enhancements can be made to the current framework by creating a user-friendly interface, thereby improving its usability.

References

[1] Ashmia Singh, Arwinder Dhillon, and Neeraj Kumar,M. Shamim Hossain, Ghulam Muhammad, Manoj Kumar, "eDiaPredict: An Ensemble-based Framework for Diabetes Prediction", doi: 10.1145/3415155, June 2021.

[2] Priyanka Rajendra, Shahram Latif, "Prediction of diabetes using logistic regression and ensemble techniques" doi: 10.1016/j.cmpbup.2021.100032 , 25 October 2021.

[3] MD.Kamrul Hasan, MD. Ashraful Alam, Dola Das, Eklas Hossain, Mahmudul Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers" , doi: 10.1109/ACCESS.2020.2989857, April 23, 2020.

[4] Muhammad Exell Febriana, Fransiskus Xaverius Ferdinana, "Diabetes prediction using supervised machine learning", doi: 10.1016/j.procs.2022.12.107, 2022.

[5] Liangjun Jiang, Zhenhua Xia .., "Diabetes risk prediction model based on community follow-up data using machine learning ", doi: 10.1016/j.pmedr.2023.102358 , 20 August 2023.

[6] Pratya Nuankaew, Supansa Chaising, Punnarumol Temdee, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction", doi: 10.1109/ACCESS.2021.3117269, October, 2021.

[7] Talha Mahboob Alama, Muhammad Atif Iqbala, Yasir Alia, "A model for early prediction of diabetes", doi: 10.1016/j.imu.2019.100204, July 2019.

[8] Roosa Peramaki, Mika Gisslerb, "The risk of developing type 2 diabetes after gestational diabetes: A registry study from Finland", doi: 10.1016/j.deman.2022.100124, 2022.

[9] Srinivasu, P. N., Shafi, J., Krishna, T. B., Sujatha, C. N., Praveen, S. P., & Ijaz, M. F. (2022). Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data. Diagnostics, 12(12), 3067.

[10] Joy Dhar, Nigus Asres Ayele, "Multi-Tier Ensemble Learning Model With Neighborhood Component Analysis to Predict Health Diseases", doi: 10.1109/ACCESS.2021.3117963, October,2021.

[11] Martin Dodek, Eva Miklovicova, Marian Tarnik, "Correlation Method for Identification of a Nonparametric Model of Type 1 Diabetes", doi: 10.1109/ACCESS.2022.3212435, October 2022.

[12] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfain , Jongtae Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension" , doi: 10.1109/ACCESS.2019.2945129, October, 2019.

[13] American Dibetes Association, "Diagnosis and classification of Diabetes Mellitus", doi: 10.2337/dc11-S062, January, 2011.

[14] Phani Praveen, S., Hasan Ali, M., Musa Jaber, M., Buddhi, D., Prakash, C., Rani, D. R., & Thirugnanam, T. (2023). IoT-Enabled Healthcare Data Analysis in Virtual Hospital Systems Using Industry 4.0 Smart Manufacturing. International Journal of Pattern Recognition and Artificial Intelligence, 37(02), 2356002.

[15] Virginie Felizardo, Diogo Machado, Nuno M. Garcia, "Hypoglycaemia Prediction Models With Auto Explanation", doi: 10.1109/ACCESS.2021.3117340, October, 2021.

[16] S. P. Praveen, S. Sindhura, A. Madhuri and D. A. Karras, "A Novel Effective Framework for Medical Images Secure Storage Using Advanced Cipher Text Algorithm in Cloud Computing," 2021 IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 2021, pp. 1-4, doi: 10.1109/IST50367.2021.9651475.

[17] Jie Zhang, Fang Wang, "Prediction of Gestational Diabetes Mellitus under Cascade and Ensemble Learning Algorithm", doi: 10.1155/2022/3212738, 2022.

[18] Tadao Ooka , Hisashi Johno, Kazunori Nakamoto, "Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan", doi:10.1136/bmjnph-2020-000200, 2021.

[19] Swamy, S. R., Praveen, S. P., Ahmed, S., Srinivasu, P. N., & Alhumam, A. (2023). Multi-features disease analysis based smart diagnosis for covid-19. Computer Systems Science and Engineering, 45(1), 869-886.

[20] Ruihu Wang, "AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review ", doi: 10.1016/j.phpro.2012.03.160 , 2012.

[21] Ziyue Yu, Wuman Luo, Rita Tse, Giovanni Pau, "DMNet: A Personalized Risk Assessment Framework for Elderly People With Type 2 Diabetes", doi:10.1109/JBHI.2022.3233622, 2023.

[22] Praveen, S. P., Jyothi, V. E., Anuradha, C., VenuGopal, K., Shariff, V., & Sindhura, S. (2022). Chronic Kidney Disease Prediction Using ML-Based Neuro-Fuzzy Model. International Journal of Image and Graphics, 2340013.

[23] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", doi: 10.1145/2939672.2939785, 2016.

[24] Dataset Retrieved from : https://data.world/data-society/pima-indians-diabetes-database

[25] Ensemble Methods retrieved from : https://www.toptal.com/machine-learning/ensemble-methods-machine-learning