EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Assessment of CatBoost for Diabetes Prevention in Comparison to XGBoost: AI model capable of predicting the onset of diabetes

Jagadeesh. Selvaraj[1,*], G. Gifta Jerith[2], Karthikeyan R.[3], Senthil K.[4]

[1]Dept. of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
[2]Dept. of CSE(Artificial Intelligence and Machine Learning), School of Engineering, Malla Reddy University, Hyderabad, India
[3]Department of CSE(AI&ML),Vardhaman College of Engineering, Hyderabad, India
[4]Dept. of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India

Abstract

The metabolic disease known as diabetes is defined by consistently elevated blood sugar levels. An increase in hunger, thirst, and urine production are symptoms of high blood sugar. Untreated diabetes may lead to a variety of complications. Acute complications of diabetes include hyperosmolarity, hyperglycemia, diabetic ketoacidosis, and perhaps death. The most devastating long-term effects are cardiovascular disease, cerebrovascular accident, chronic kidney disease, foot ulcers, and vision loss. The World Diabetes Organization estimates that 463 million people were diagnosed with diabetes in 2019. This population will increase by 578 million by 2030 and by 700 million by 2045, if forecasts pan out. The ability to quickly and accurately diagnose sickness is one of its current medical uses. Therefore, we might potentially reduce death rates via the use of machine learning by creating an AI model that can anticipate when diabetes will start. We will compare the CatBoost and XGBoost algorithms to find the one that is most suited for this purpose. Finally, using a number of health markers from the dataset, the study contrasted XGBoost and CatBoost, two models that may predict diabetes. We train and build our recommended system using Python on a real-world dataset taken from Kaggle. We evaluate our algorithms using precision, recall, F1score, and accuracy metrics, among other performance evaluation parameters. While XGBoost achieved an F1 Score of 91.75, an accuracy rate of 93.33%, a precision of 90.38%, and a recall of 90.63%. The accuracy, precision, recall, and F1 score for CatBoost are 96.09%, 93.38%, 91.38% and 92.13%, respectively. It's the most effective ensemble method, according to CatBoost.

Keywords: Diabetics, CatBoost, XGBoost, Artificial Intelligence, Ensemble Model, Boosting Algorithms

Received on 24 04 2024, accepted on 31 01 2025, published on 10 02 2025

doi: 10.4108/eetiot.5880

*Corresponding author. Email: jagadeesh15.sj@gmail.com

1. Introduction

One characteristic of type 2 diabetes is the body's inability to manufacture insulin. In the case of this chronic illness, the pancreas is the organ directly affected [1]. Maintaining steady blood glucose levels is mostly the job of the hormone insulin. People with hypertension, high cholesterol levels, obesity, and insufficient physical activity are more likely to acquire diabetes [2]. One of the many possible adverse effects is an increase in the amount of time one has to go to the bathroom [3]. Untreated diabetes may lead to renal failure and diabetic retinopathy, an eye disorder that can affect the skin, nerves, and eyes.

One kind of diabetes is insulin-dependent diabetes mellitus (IDDM). Since insulin production is inadequate on a physiological level, people with type 1 diabetes must inject themselves with insulin. When insulin is unable to control blood sugar levels, a disease known as type 2 diabetes mellitus (NIDDM) sets up. This kind of diabetes develops when cells in the body are unable to properly use insulin. Type 3 gestational diabetes develops in pregnant women when the problem is not recognized in a timely manner and blood sugar levels reach dangerously high levels. Diabetes is associated with complications in the long run. Numerous additional health issues might also occur in a diabetic person.

By 2030, with healthcare systems already struggling under the weight of an increasingly severe pandemic, the International Diabetes Federation predicts that diabetes will have surpassed all other causes of death [4]. Obesity, poor diet, smoking, excess body fat, and insufficient physical activity are some of the risk factors linked to the development of diabetes, according to previous studies [5-9].

Only medical experts can determine which kind of diabetes a patient has. Due to the length of time it required to determine the diagnosis, many patients who undergo evaluations are suffering from severe illnesses. Therefore, we will compare CatBoost with XGBoost to see whether algorithm is more effective for diabetes prediction. With the help of the Kaggle Community, we want to analyze the Pima Indian Diabetes dataset as part of this project. There is a single dependent variable and eight independent variables in the dataset. This research considers eight diabetes-related factors—age, skin thickness, blood pressure, glucose, insulin, body mass index (BMI), function of the diabetic lineage, and pregnancy—to validate a diabetes diagnosis. There are a lot of moving parts in a diabetes diagnosis, but getting the right one quickly is crucial.

2. Literature Review

In order to increase the dataset's ability to forecast illnesses, Li et al. [10] used ensemble learning methods to build a diabetes predictor model. In terms of accuracy, XGBoost is the way to go; it reached 80.20 percent. An improved feature combination classifier based on the XGBoost model was proposed by the authors to improve healthcare-related disease prediction. When assessing ensemble learning approaches for diabetes prediction, Mahabub [11] considered a number of clinical characteristics. A variety of methods were used, including AdaBoost, gradient boost, XGBoost, random forest, and others. Using multilayer perceptron technology, they attained an impressive accuracy rate of 84.42%. If Mushtaq et al. [12] wanted to make better use of the dataset for diabetes prediction, they would use an ensemble approach based on vote classification. For this study, the researchers used a two-step model selection procedure to create the model. In terms of accuracy, the voting classifier stands head and shoulders above the others at 81.50%.

Multiple boosting techniques were used by Beschi Raja et al. [13]. The accuracy rate of the gradient boosting strategy was the highest of all the classifiers at 89.70%. The proposed strategy's efficacy has been assessed using a number of statistical measures. Using the boosting method, A model for the prediction of diabetes was developed by Khan et al. [14]. The authors searched into many classifiers for predictive analytics, such as artificial neural networks (ANNs), naive Bayes, deep learning, gradient boosting, j48, and hybrid k-nearest neighbour (kNN). The gradient boosting approach outperformed the other classifiers.

In order to analyze diabetes predictions, Lai et al. [15] came up with a detailed method. Minimizing the loss of classification prediction probabilities was achieved by using gradient boosting machine approaches with hyperparameter change, especially for class balance. Singh et al. [16] created the eDiaPredict framework, which employs an ensemble approach to forecast whether a patient will have diabetes. Included in the proposed method are the following algorithms: support vector machine, neural network, XGBoost, decision tree, and random forest. We can verify eDiaPredict's efficacy by running it on the PIMA diabetes dataset from India. We get a sensitivity level of 90.32%, a precision level of 88%, and an accuracy of 95% by merging XGBoost with random forest. According to Hasan et al. [17], a diabetes prediction framework is suggested that uses XGBoost, kNN, AdaBoost, decision trees, Naive Bayes, random forest, and multilayer perceptron. A dataset was used to evaluate a weighted ensemble of ML models, with the goal of improving prediction accuracy.

The proposed ensemble model was much more effective than the alternatives, with an area under the curve (AUC) of 0.950 and a specificity of 0.934. While it was accurate 84.32% of the time and 88.84% of the time, its sensitivity was just 78%.

3. Research Methdology

In Figure 1 we can see the proposed structure for the experimental investigation and its flow of procedures. It details the steps to take in a certain sequence to use an ensemble learning method based on boosting methods to

improve diabetes prediction accuracy. The diabetic dataset was contributed to this study by members of the Kaggle community. After upsampling and normalizing were finished, two boosting algorithms were created. We trained the models on 80% of the dataset and tested and validated their performance on the remaining 20%. When building the model, hyperparameter tweaking was used to get better results.

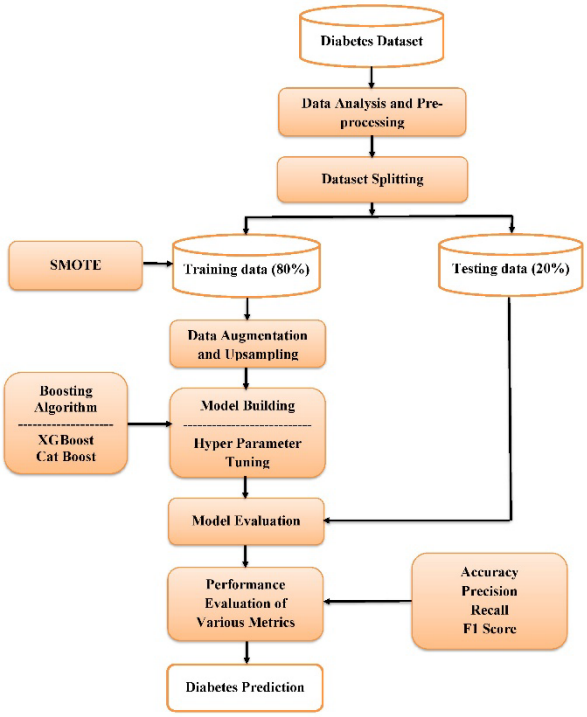


Figure 1. Research approach suggested

3.1. Boosting Algorithms

Boosting is one method for improving weak classifiers. Ensemble learning has many real-world uses [18]. Ensemble learning has grown in popularity in the healthcare sector because to its impressive performance in predicting, detecting, diagnosing, and prognosing a diverse array of diseases. In this particular experiment, we assessed two ensemble learning-based boosting algorithms for diabetes prediction:

The XGBoost: XGBoost is able to independently compute similarity scores because it employs a network of weak learners, which is analogous to decision trees [19]. By using regularization and gradient descent, When training, XGBoost successfully prevents the overfitting problem. Adjustments are made to the gradient descent and regularization strategy to tackle the problem of overfitting during training.

In comparison to other boosting algorithms, CatBoost—Because it does not investigate data preparation, CatBoost, a compressed variant of categorical boosting, is faster [20]. A high cardinality of categorical variables is its intended use. In order to convert variables with poor cardiacity, the one-hot encoding approach is used.

3.2. Data Pertaining to Attributes

Quite a large dataset, with 768 occurrences and 9 attributes in all. The last attribute will serve as the success metric, with the preceding eight traits serving as independent variables or predicates. All of the qualities, together with their descriptions, range values, and measurements, are shown in Table 1.

Table 1. Provide details about the characteristics of the dataset.

	Attributes	Description	Measurements	Value Range
1	Pregnancy	Number of pregnancies reported by participants	Numeric	0–17
2	Glucose	The subject's plasma glucose tolerance	mg/dL	0–199
3	Blood pressure	The diastolic blood pressure of the subject	mmHg	0–122
4	Skin thickness	The measurement of the triceps fascia's thickness in the participant	mm	0–99
5	Insulin	Two-hour serum insulin level of the participant	(mu U/mL)	0–846
6	Body mass index	A participant's body fat percentage as a function of their height and weight	kg/m2	0–67
7	Diabetes pedigree function	Diabetes risk according to the individual's family medical history	p-value	0.07–2.42
8	Age	The participant's age	Years	21–81
9	Diabetes	Class attribute	0 = no diabetes, 1 = diabetes	0 or 1

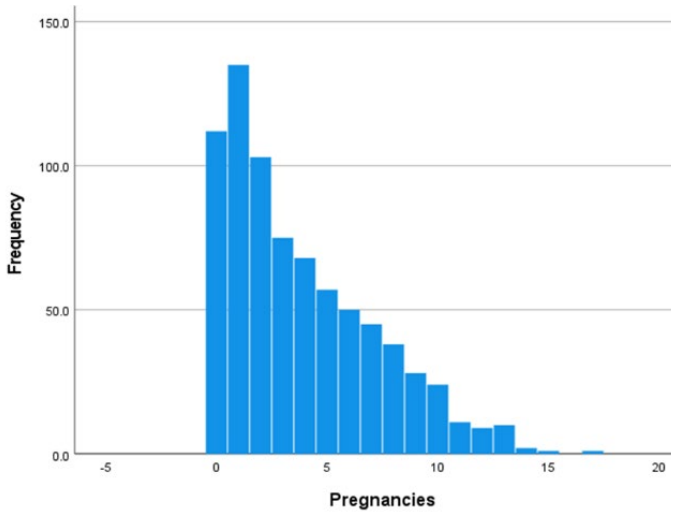
3.3. Dataset Description

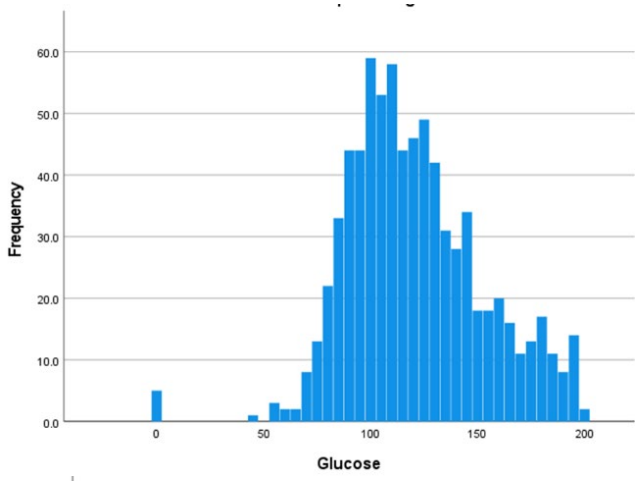
When individuals need to describe data samples or have information summarized for interpretation purposes, descriptive statistics are a lifesaver. Table 2 shows the data for each attribute, including the total number of records, minimum and maximum values, average, and standard deviation (std). Suppose we look at the Pregnancy attribute. Results for pregnancy range from 17 to 0, and the data set contains 786 items with an average of 3.84 and a standard deviation of 3.36. We have also measured the other attributes using statistical approaches. The data's distribution and features may be better understood with the help of these measures.

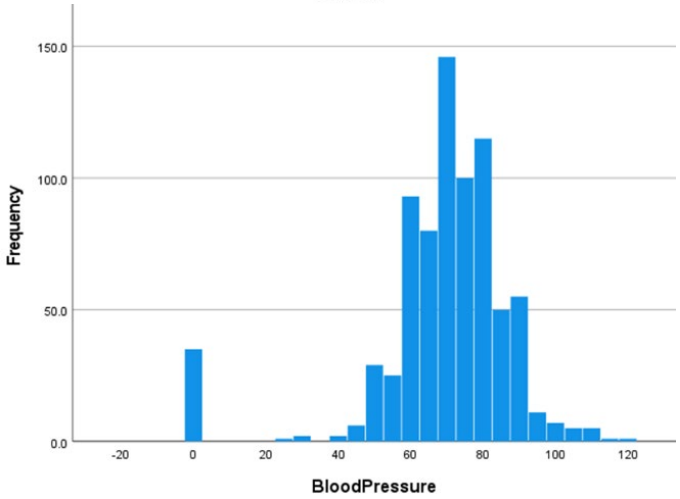
3.4. Attribute Histrogram

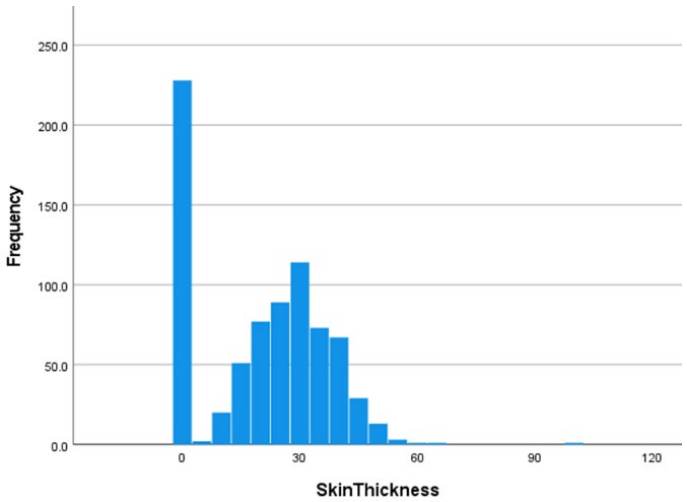
A histogram may help you understand how the data in a dataset is distributed. This will reveal whether the data is normally distributed, left-skewed, or right-biased. The distribution of all characteristics is shown by the normally distributed histograms, which can be seen in Figure 2. If you're having trouble seeing trends or outliers in your dataset, this visualization could help you better understand its distribution. We may see the input attributes and their values side by side on the graph. In Figure 3, we can see the distribution of traits for both people with and without diabetes. In this example, a patient without diabetes is represented by the number 0, while a patient with diabetes is represented by the number 1, both in green.

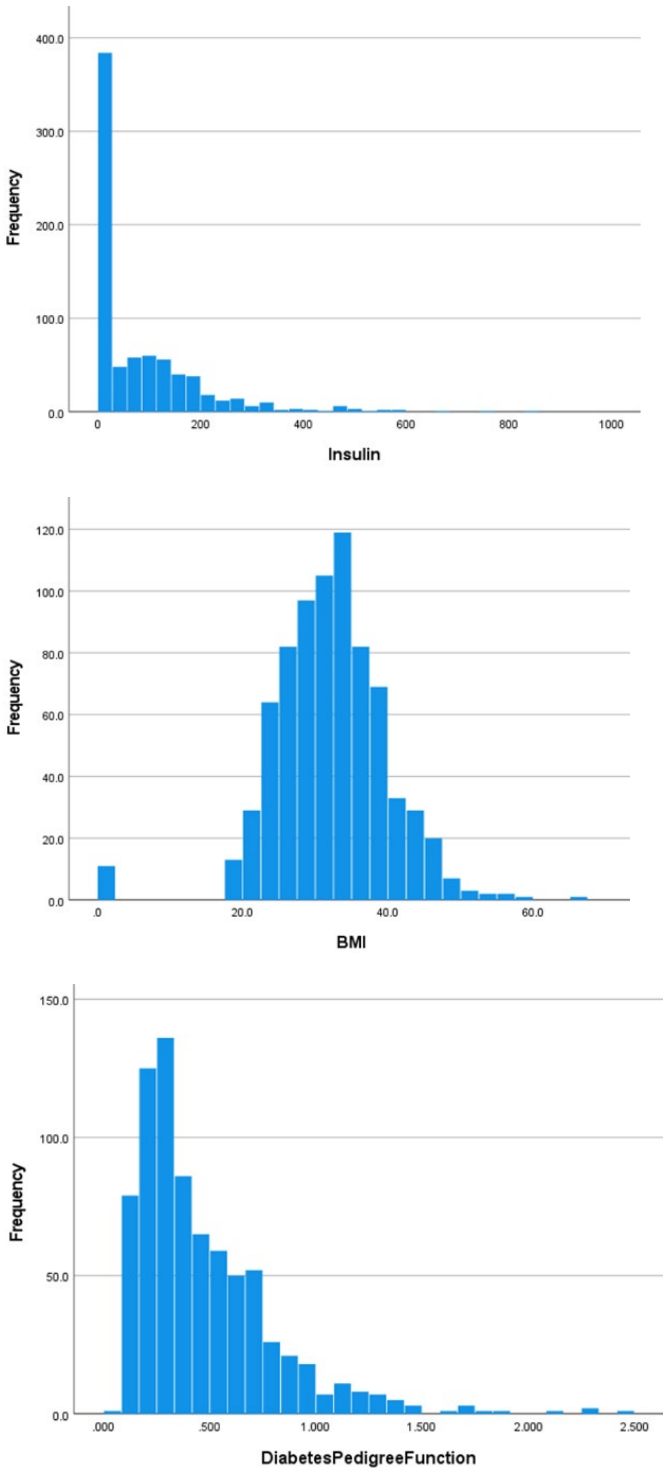
Graphical Representation of the distribution using Histogram











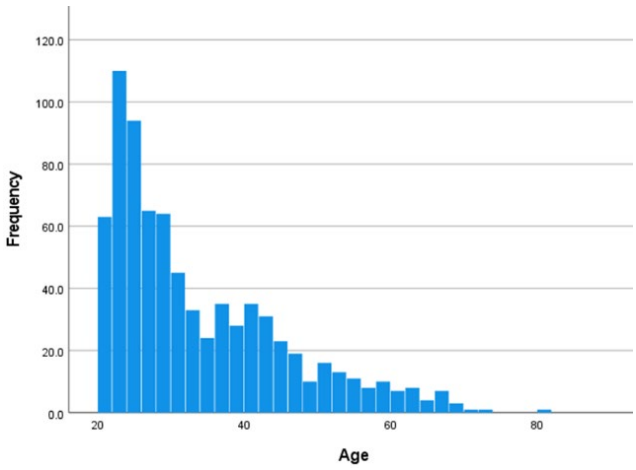


Figure 2. Histogram of attributes

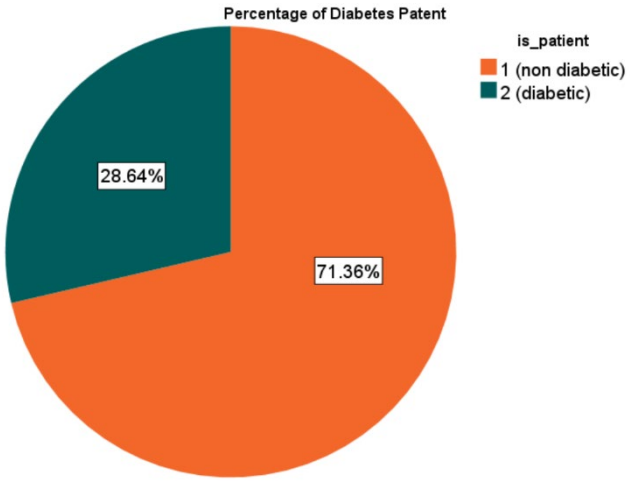


Figure 3. Percentage of diabetes

Here Figure 3 represents the percentage of diabetes among Pima Indians participants.

4. A Trial, Analysis and Commentary

The results and experimental details of the boosting algorithm diabetes prediction are covered in this section. Results obtained via the use of the suggested framework are meticulously documented and assessed. Accuracy, recall, precision, and F1-score are only a few of the criteria used in a comprehensive evaluation of the boosting methods under consideration. We can learn a lot about the efficacy of the algorithms used to improve performance and the accuracy of diabetes predictions from these measurements.

4.1. Data preprocessing

If you want to create a reliable system, you must prepare the data before using machine learning boosting methods. The data imputation approach was first used for handling missing variables. After missing data was detected using the isnull() function, it was filled in using the mean and mode imputation approach using the SimpleImputer() method. This method filled in blanks caused by missing data by using the median, mode, or mean of the corresponding column.
Using data cleaning procedures, we were able to remediate any duplicate, inconsistent, or corrupted data. By eliminating or fixing any duplicate entries, incorrect values, or broken data points, these techniques guaranteed the dataset's trustworthiness and integrity. By optimizing and preparing the dataset for future machine learning algorithms, these data pretreatment strategies enhanced the analysis's reliability and quality.

4.2. Data upsampling

Alphabetically unbalanced datasets provide subpar performance from ML and DL algorithms [21]. A substantial bias existed in the dataset that favored the negative class, to the positive category of "1-diabetic," from "0-non-diabetic," in this investigation. Out of 786 records, only 268 were originally kept for the positive class. Out of 500 records, the class was negative. Following the split, the training dataset consisted of 614 records; 218 of these records pertained to individuals with diabetes, whereas 396 pertained to controls. A SMOTE-based balancing procedure was used on the training set. The min-max approach was used to normalize all attribute values in datasets to a range of 0 to 1. The procedure was carried out in accordance with Eq. (1), where x is the value of the attribute x_{min}, x_{max} are its lowest and maximum values, respectively.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

4.3. Optimising hyperparameters

Because they control how the training algorithm operates and greatly affect how the model's performance is assessed, hyperparameters are an important variable to change [22].
The outcomes of the hyperparameter tweaking, which included both random and grid searches, are shown in Table 2. Our experiments showed that the aforementioned parameter values for each strategy yielded the most effective outcomes.

Table 2. Optimising hyperparameters for boosting algorithms

CatBoost	Two values, 0.010 and 0.004, make up the learning rate. We have 4 for the "depth" parameter and 1.0 for the "leaf_reg" parameter. There are 32 possible options for the "min_child_samples" parameter, which ranges from 1 to 32. Both the random state and the number of iterations are set to 3000 and 42, respectively.
XGBoost	With these parameters set: learning rate to 0.01, estimators to 1000, maximum depth to 4, minimum child weight to 8, subsample to 0.6, regularisation alpha to 0.005, seed to 27.

We used precision, F1score, recall, and classification accuracy to evaluate the XGBoost and CatBoost Algorithms. The equations that represent these metrics are,

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = \frac{2\ x\ Recall\ x\ Precision}{Recall+Precision} \quad (4)$$

where TP means that the result is positive and the model's forecast is positive. The result is negative, despite the fact that FP represents the model's optimistic prediction. A negative outcome relative to the model's expectations is denoted by TN. The FN shows that the result is positive, even though the model expected a negative outcome. Stratified 8:2 train-test splits are used in this work to implement the holdout validation approach using Boosting algorithms. Using the SMOTE synthetic oversampling approach, we can examine the results of several classifiers on the combined dataset in Table 3. Table 3 shows that CatBoost had the best overall performance with the following metrics: accuracy (91.08%), recall (86.38%), precision (88.38%), and F1 score (87.38%).

Table 3. Classification Accuracy of CatBoost and XGBoost Model

		XGBoost				CatBoost			
Attempt	**Training Percentage**	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**Accuracy**	**Recall**	**Precision**	**F1-Score**
1	80%	97.93%	95%	94%	96%	98.55%	94%	95%	95%
2	70%	97.92%	93%	95%	94%	96.23%	90%	93%	91%
3	60%	94.02%	89%	91%	90%	95.62%	89%	93%	91%
4	50%	93.96%	90%	91%	91%	95.81%	91%	93%	91%
5	40%	93.52%	90%	91%	90%	95.56%	91%	93%	92%
6	30%	90.07%	87%	87%	89%	96.58%	93%	95%	92%
7	20%	88.12%	94%	85%	95%	95.48%	91%	93%	92%
8	10%	91.1%	87%	89%	89%	94.89%	92%	92%	93%
Average		93.33%	90.63%	90.38%	91.75%	96.09%	91.38%	93.38%	92.13%

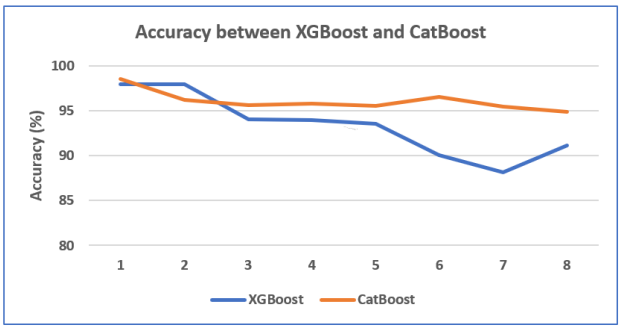


Figure 4. Accuracy between XGBoost and CatBoost

The figure 4 displays that the accuracy value between XGBoost and CatBoost fluctuates in relation to the dataset's percentage distribution.

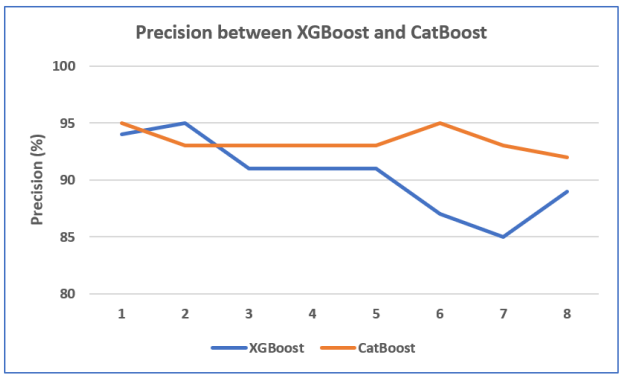


Figure 5. Precision between XGBoost and CatBoost

The accuracy value between XGBoost and CatBoost varies when the dataset's percentage distribution changes (Figure 5).

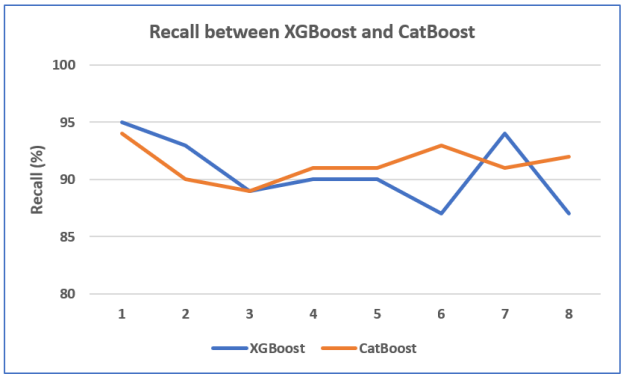


Figure 6. Recall between XGBoost and CatBoost

Figure 6 shows that the recall value between XGBoost and CatBoost varies with the percentage distribution of the dataset.

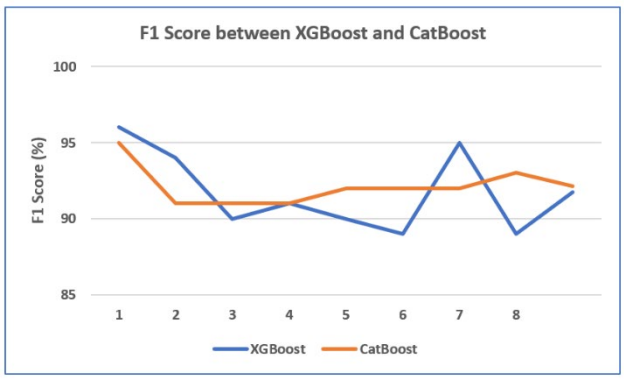


Figure 7. F1 Score between XGBoost and CatBoost

Figure 7 shows that the F1 Score value between XGBoost and CatBoost varies with the percentage distribution of the dataset.

5. Conclusion and suggestions

For supervised machine learning diabetes prediction using a variety of health factors in the dataset, the study ultimately contrasted the XGBoost method with CatBoost approaches. In our trials and assessments, XGBoost obtains an average value of 93.33% accuracy, 90.38% precision, 90.63% recall, and an F1 Score of 91.75, whereas CatBoost earns an average value of 96.09% accuracy, 93.38% precision, 91.38% recall, and 92.13 F1 Score. Predicting diabetes in the Pima Indians population using the CatBoost algorithm is, therefore, the clear winner. Improving diabetes prediction and diagnosis might be possible with further study into deep learning methods. Additionally, additional boosting algorithms, such Bagging, might enhance the precision and accuracy of future investigations. The potential for improved healthcare solutions is enhanced by these ML and deep learning breakthroughs.

Conflict of Interest

The writers have all denied any involvement in a potential conflict of interest. We would want to make it clear that we did not accept any kind of financial support or payment to conduct this study.

Acknowledgements

SJ and GJ drafted and edited the paper. The tables and graphs in the experimental findings section were contributed by SJ, who also conducted the experiments.

References

[1] Kharroubi, A.T., Darwish, H.M.: Diabetes mellitus: The epidemic of the century. World J. Diabetes 6, 850–867 (2015)

[2] Wu, Y., Ding, Y., Tanaka, Y., Zhang, W.: Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. Int. J. Med. Sci. 11, 1185–1200 (2014)

[3] Papatheodorou, K., Banach, M., Edmonds, M., Papanas, N., Papazoglou, D.: Complications of diabetes. J. Diabetes Res. 2015, 1–6 (2015).

[4] International Diabetes Federation. Available online: https://diabetesatlas.org/ (accessed on 1 September 2022).

[5] Baliunas, D.O.; Taylor, B.J.; Irving, H.; Roerecke, M.; Patra, J.; Mohapatra, S.; Rehm, J. Alcohol as a risk factor for type 2 diabetes: A systematic review and meta-analysis. Diabetes Care 2009, 32, 2123–2132.

[6] Vazquez, G.; Duval, S.; Jacobs, D.R., Jr.; Silventoinen, K. Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: A meta-analysis. Epidemiol. Rev. 2007, 29, 115–128.

[7] Odegaard, A.O.; Koh,W.-P.; Butler, L.M.; Duval, S.; Gross, M.D.; Yu, M.C.; Yuan, J.-M.; Pereira, M.A. Dietary patterns and incident type 2 diabetes in chinese men and women: The singapore chinese health study. Diabetes Care 2011, 34, 880–885.

[8] Smith, A.D.; Crippa, A.; Woodcock, J.; Brage, S. Physical activity and incident type 2 diabetes mellitus: A systematic review and dose–response meta-analysis of prospective cohort studies. Diabetologia 2016, 59, 2527–2545.

[9] Pan, A.; Wang, Y.; Talaei, M.; Hu, F.B.; Wu, T. Relation of active, passive, and quitting smoking with incident type 2 diabetes: A systematic review and meta-analysis. Lancet Diabetes Endocrinol. 2015, 3, 958–967.

[10] Li, M., Fu, X., and Li, D. (2020). Diabetes prediction based on XGBoost algorithm. IOP Conf. Ser. Mater. Sci. Eng. 768 (7), 072093. doi:10.1088/1757-899x/768/7/072093

[11] Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Appl. Sci. 1 (12), 1667–1712. doi:10.1007/s42452-019-1759-7

[12] Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., and Husnain, M.(2022). Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. Mob. Inf. Syst. 2022, 1–16. doi:10.1155/2022/6521532

[13] Beschi Raja, J., Anitha, R., Sujatha, R., Roopa, V., and Sam Peter, S. (2019). Diabetics prediction using gradient boosted classifier. Int. J. Eng. Adv. Technol. 9 (1), 3181–3183. doi:10.35940/ijeat.a9898.109119.

[14] Khan, A. A., Qayyum, H., Liaqat, R., Ahmad, F., Nawaz, A., and Younis, B. (2021). "Optimised prediction model for type 2 diabetes mellitus using gradient boosting algorithm," in Proceedings of the 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, July 2021, 1–6.

[15] Lai, H., Huang, H., Keshavjee, K., Guergachi, A., and Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr. Disord. 19 (1), 101–109. doi:10.1186/s12902-019-0436-6

[16] Singh, A., Dhillon, A., Kumar, N., Hossain, M. S., Muhammad, G., and Kumar, M. (2021). eDiaPredict: an ensemble-based framework for diabetes prediction. ACM Trans. Multimedia Comput. Commun. Appl. 17 (2), 1–26. doi:10.1145/3415155

[17] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., and Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 8, 76516–76531. doi:10.1109/access.2020.2989857

[18] Ganie, S. M., Pramanik, P. K. D.,Malik, M. B., Nayyar, A., and Kwak, K. S. (2023). An improved ensemble learning approach for heart disease prediction using boosting algorithms. Comput. Syst. Sci. Eng. 46 (3), 3993–4006. doi:10.32604/csse.2023.035244.

[19] Santhanam, R., Uzir, N., Raman, S., and Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. Int. J. Control Theory Appl. 9 (40), 651–662.

[20] Hancock, J. T., and Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. J. Big Data 7 (1), 94. doi:10.1186/s40537-020-00369-8.

[21] Ganie, S. M., Malik, M. B., and Arif, T. (2022b). "Machine learning techniques for diagnosis of type 2 diabetes using lifestyle data," in Proceedings of the International Conference on Innovative Computing and Communications, New Delhi, India, August 2021, 487–497.

[22] Ganie SM, Pramanik PKD, Bashir Malik M, Mallik S and Qin H (2023), An ensemble learning approach for diabetes prediction using boosting techniques. Front. Genet. 14:1252159. doi: 10.3389/fgene.2023.1252159.