Contents lists available at *Dergipark*

Journal of Scientific Reports-C

journal homepage: *https://dergipark.org.tr/en/pub/jsrc*

E-ISSN: 2717-8633 ***Sayı(Number) 6, Nisan(April) 2024***

ARAŞTIRMA MAKALESİ/RESEARCH ARTICLE

Geliş Tarihi(Receive Date): 25.09.2023 *Kabul Tarihi(Accepted Date): 29.01.2024*

A comparison of traditional and state-of-the-art machine learning algorithms for type 2 diabetes prediction

Naciye Nur Arslan[a*], Durmuş Özdemir[b]

[a]*Kütahya Dumlupınar University, Faculty of Engineering, Department of Software Engineering, 43000 Kütahya, Turkey,*
ORCID: 0000-0002-3208-7986
[b]*Kütahya Dumlupınar University, Engineering Faculty, Software Engineering department, 43100 Kütahya, Turkey.*
ORCID: 0000-0002-9543-4076

Abstract

This research investigates the use of machine learning algorithms for early detection of diabetes. Due to its global prevalence and significant impact on health, timely identification of diabetes is crucial for effective treatment. In this study, machine learning models including Gradient Boosting Machines, Extreme Gradient Boosting, Light gradient-boosting machine, Categorical Boosting, k-Nearest Neighbors, Random Forest, Ridge Classifier, Logistic Regression, Gaussian Naive Bayes, and Decision Tree are utilized to assess their capabilities in diabetes diagnosis. The primary aim is to train these models to distinguish between individuals with diabetes and those without, using relevant features from the dataset. Since the classes in the dataset are imbalanced, the SMOTE technique is applied to improve model performance. Categorical Boosting achieved the highest accuracy rate of 90.05%, making it the most successful model. By systematically evaluating the performance of these prominent machine learning models, valuable insights can be gathered regarding their ability to recognize complex patterns indicative of diabetes. As a result, healthcare professionals and researchers can leverage this newfound understanding to develop more accurate and effective diagnostic tools, enabling early intervention and subsequently improving the overall quality of life for individuals affected by diabetes.

Keywords: diabetes; machine learning; ensemble learning; boosting; bagging; catboost; xgboost; lightgbm.

[*] Corresponding author: naciye.arslan@dpu.edu.tr, Tel: +90 274 443 43 34

1. Introduction

Diabetes is a chronic disease characterized by high blood glucose (sugar) levels. Glucose is a vital energy source for the body's cells, but it requires insulin, a hormone produced by the pancreas, to enter the cells. In people with diabetes, their body either does not produce enough insulin or cannot use it effectively, resulting in elevated blood glucose levels [1]. There are two main types of diabetes: Type I and Type II diabetes. Type I diabetes usually begins in childhood and adolescence. Patients must inject insulin externally. Type II diabetes occurs with advancing age. In the initial stages, it can be controlled with diet and exercise, but in advanced cases, medication or insulin injection may be required [2]. The cause of diabetes is not fully understood. However, there are several reasons of diabetes, including a person's lifestyle and genetics.

Diabetes necessitates constant medical care to lower the risk of several health concerns, including neuropathy, kidney failure, cardiovascular issues, and vision impairment [3]. Early identification and treatment are essential to avoiding the disease's secondary complications as well as any long-term effects.

Research has focused on using various classification techniques for diabetes prediction [5]. These studies aim to develop accurate and reliable models for early diagnosis and intervention. Xie et al. attempted to create risk prognosis models for Type 2 diabetes by applying the Synthetic Minority Oversampling Technique (SMOTE) to eliminate class imbalance in the BRFSS 2014 (Behavioral Risk Factor Surveillance System) dataset. With 82.4% accuracy, their results showed that the neural network offered the best overall accuracy rate among all models [5]. Wei and colleagues conducted a comprehensive analysis of diabetes detection approaches using the Pima Indian dataset, focusing on Deep Neural Networks and support Vector Machines (SVM) [6]. In another study, Yahyaoui et al. used the Pima Indian dataset to examine the performance of SVM, Random Forest (RF), and Convolutional Neural Network models for diabetes prediction. They found that RF was the most effective model, with an accuracy rate of 83.67% [7].

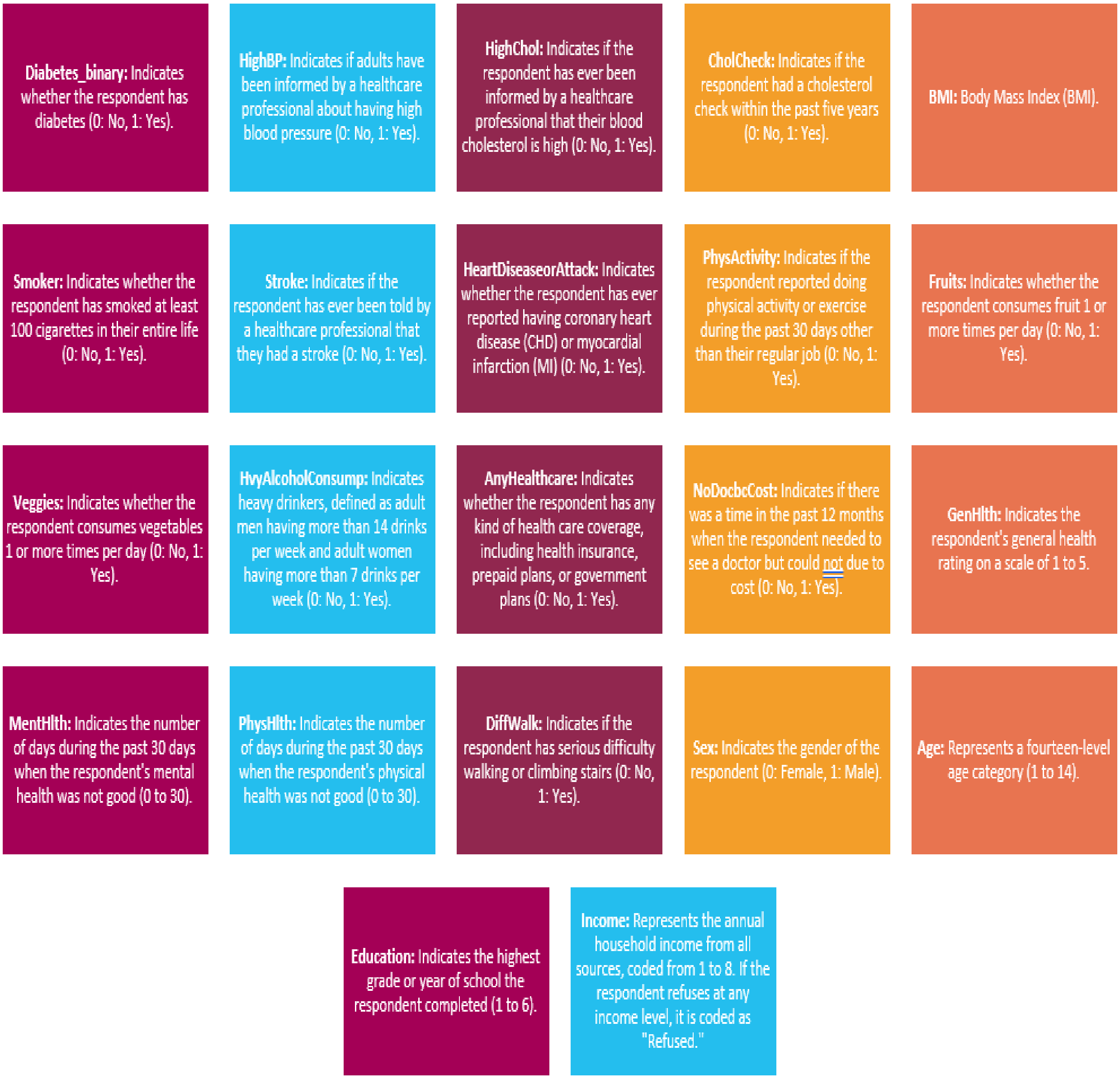
Our research seeks to obtain high accuracy rates, thoroughly evaluate different machine learning approaches for diabetes diagnosis, and assess the BRFSS 2015 dataset utilized for diabetes diagnosis. The aim is to identify which techniques are most suited for obtaining the most accurate and reliable outcomes in the diagnosis and prognosis of diabetes by contrasting the performance of different methods.

The introduction comprehensively introduced the overall objectives and methodology of the research. The following section provided information about the dataset used for diabetes diagnosis and discussed data preprocessing techniques in detail. Additionally, it explained machine learning models used for classification. Section 3 presented the experimental results of the machine learning models used in diabetes classification. The 4th and final section presented a summary of the findings, general conclusions, contributions of the study, and recommendations for future research.

2. Material and methods

2.1. Dataset

In this study, the data set named "Diabetes Health Indicators Data Set" obtained from Kaggle was used [8]. The dataset was derived from the BRFSS-2015 results, one of the most widespread and comprehensive public health surveys in the United States. Three CSV data sets were produced for diabetes prediction from BRFSS-2015 survey data. The dataset we use consists of 253,680 survey responses. The data set we use includes a total of 22 features such as diabetes status, high blood pressure, cholesterol, physical activity, smoking habit, gender, age, education level, income level, general health and mental health status. These features and their properties are shown in Figure 1. Investigating the importance of these features in diabetes prediction directly affects model performance.



Diabetes_binary: Indicates whether the respondent has diabetes (0: No, 1: Yes).

HighBP: Indicates if adults have been informed by a healthcare professional about having high blood pressure (0: No, 1: Yes).

HighChol: Indicates if the respondent has ever been informed by a healthcare professional that their blood cholesterol is high (0: No, 1: Yes).

CholCheck: Indicates if the respondent had a cholesterol check within the past five years (0: No, 1: Yes).

BMI: Body Mass Index (BMI).

Smoker: Indicates whether the respondent has smoked at least 100 cigarettes in their entire life (0: No, 1: Yes).

Stroke: Indicates if the respondent has ever been told by a healthcare professional that they had a stroke (0: No, 1: Yes).

HeartDiseaseorAttack: Indicates whether the respondent has ever reported having coronary heart disease (CHD) or myocardial infarction (MI) (0: No, 1: Yes).

PhysActivity: Indicates if the respondent reported doing physical activity or exercise during the past 30 days other than their regular job (0: No, 1: Yes).

Fruits: Indicates whether the respondent consumes fruit 1 or more times per day (0: No, 1: Yes).

Veggies: Indicates whether the respondent consumes vegetables 1 or more times per day (0: No, 1: Yes).

HvyAlcoholConsump: Indicates heavy drinkers, defined as adult men having more than 14 drinks per week and adult women having more than 7 drinks per week (0: No, 1: Yes).

AnyHealthcare: Indicates whether the respondent has any kind of health care coverage, including health insurance, prepaid plans, or government plans (0: No, 1: Yes).

NoDocbcCost: Indicates if there was a time in the past 12 months when the respondent needed to see a doctor but could not due to cost (0: No, 1: Yes).

GenHlth: Indicates the respondent's general health rating on a scale of 1 to 5.

MentHlth: Indicates the number of days during the past 30 days when the respondent's mental health was not good (0 to 30).

PhysHlth: Indicates the number of days during the past 30 days when the respondent's physical health was not good (0 to 30).

DiffWalk: Indicates if the respondent has serious difficulty walking or climbing stairs (0: No, 1: Yes).

Sex: Indicates the gender of the respondent (0: Female, 1: Male).

Age: Represents a fourteen-level age category (1 to 14).

Education: Indicates the highest grade or year of school the respondent completed (1 to 6).

Income: Represents the annual household income from all sources, coded from 1 to 8. If the respondent refuses at any income level, it is coded as "Refused."

Fig. 1. Dataset features.

The pie chart presented in Figure 2 points to the class imbalance problem in the data set we used. The majority of the data set consists of the "no diabetes" class with a rate of 84.71%, while 15.29% consists of the "diabetes" class. This class imbalance problem directly affects the performance of models to be used in diabetes prediction.

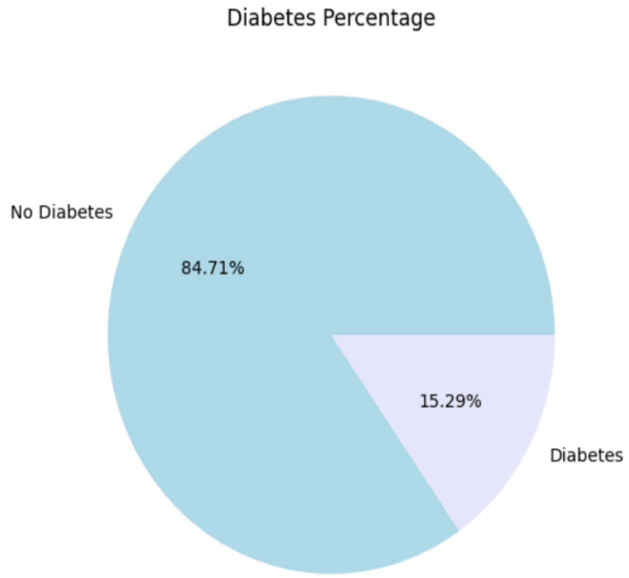


Fig. 2. Distributions between classes.

2.2. Data preprocessing

The presence of duplicate observations in the data was checked and removed. Duplicate observations refer to instances that have identical values across all columns. Removing duplicate observations is performed to prevent bias and ensure accurate data representation [9].

Fig. 3 displays a graph illustrating the correlation between the Diabetes feature and other attributes. Correlation is a statistical measure that quantifies the strength and direction of the relationship between two variables. Values close to 0 indicate a weak or nearly nonexistent correlation between the two variables. A value of 0 signifies no relationship between the two variables. As values move from 0 towards a positive value, they represent a positive relationship (an increase in one variable corresponds to an increase in the other). On the other hand, values moving away from 0 towards -1 indicate a negative relationship (an increase in one variable corresponds to a decrease in the other) [10]. In line with this, features with correlations close to zero were removed from the dataset to improve performance and reduce processing costs. Expressly, nine features named 'AnyHealthcare,' 'Sex,' 'Smoker,' 'NoDocbcCost,' 'CholCheck,' 'MentHlth,' 'Fruits,' 'Veggies,' and 'HvyAlcoholConsump' were excluded from the dataset.

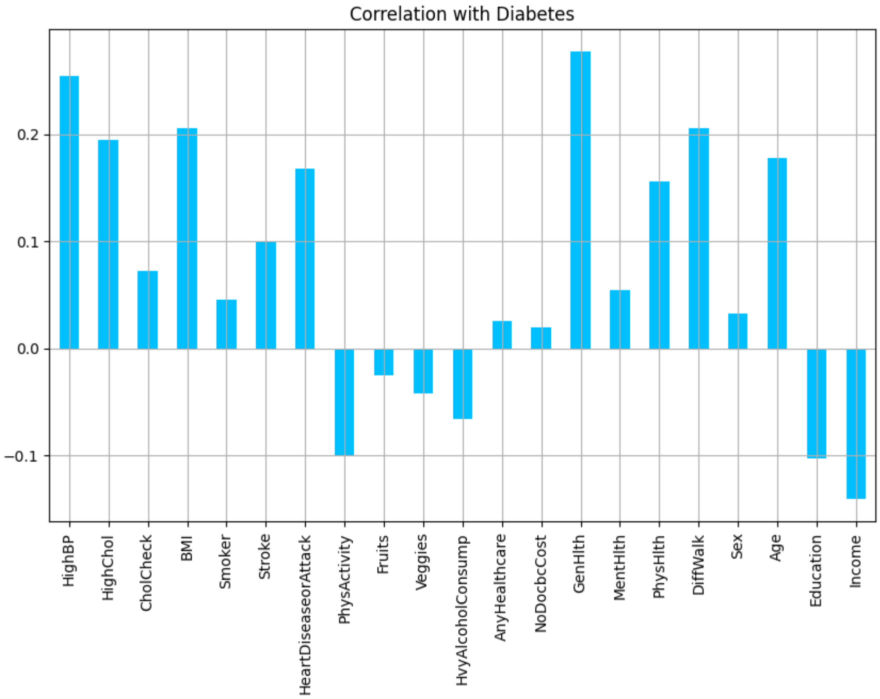


Fig. 3. Correlation with diabetes.

Due to the imbalanced distribution of the dataset, the SMOTE technique was applied. The primary objective of SMOTE is to balance the dataset by oversampling the minority class, thereby enhancing the performance of machine learning models [11].

The following are the SMOTE technique's application stages:

- The k-nearest neighbors are chosen for every instance of the minority class. We selected k = 5 neighbors for our study, and we used the Euclidean distance or another similarity metric to find these neighbors.
- A neighbor is chosen at random from the list of k-nearest neighbors for each minority sample. The delta, or difference, between the selected neighbor and the original sample, is then computed.
- A new synthetic sample is created by randomly selecting a multiplier between 0 and 1, based on the delta value, to determine the size of the difference between the original sample and the chosen neighbor.
- Finally, the first minority class is expanded to include the new synthetic sample.

By increasing the amount of minority class samples using this technique, our study intended to remove class imbalance in the data set. The dataset was divided into 80% for training and 20% for testing after SMOTE was used. The purpose of these preprocessing procedures was to guarantee data quality and get the data ready for modeling activities.

2.3. Machine learning models

Building a model with several learners as opposed to only one is known as ensemble learning [12]. Ensemble learning methods are machine learning approaches that include bagging and boosting. By combining several weak learners, these techniques provide a prediction or classification model that is more reliable. Although they employ different strategies, both techniques seek to increase model performance. Bagging is short for "bootstrap aggregation." The Bagging technique first creates multiple subsets from the original dataset. A weak learner model is built for each of these subsets. The models make predictions or classifications independently of each other [13]. In other words, the models work in parallel to each other. The final prediction result is obtained by averaging the predictions of each model. The most popular machine learning technique used in Bagging is Random Forest. These methods are used to enhance model performance while handling the characteristics and complexity of the dataset, demonstrating the power of Ensemble Learning.

Boosting, on the other hand, works by sequentially training weak learners and weighting each learner based on the errors of previous learners [14]. Initially, a subset is created from the original dataset, and a base model is built on this subset. Initially, all data points are given equal weights. Errors are calculated, and more weight is assigned to incorrect predictions. Another model is then built, and this model attempts to correct the errors from the previous model. This involves a sequential process [15]. The final model is a weighted average of all the models. Examples of boosting algorithms include GBM (Gradient Boosting Machines), XGBoost (Extreme Gradient Boosting), LightGBM (Light Gradient Boosting Machine), and CatBoost (Categorical Boosting). XGBoost is known for its faster and more scalable structure, which is especially effective in handling large datasets. LightGBM is recognized for its high efficiency and low memory usage, providing advantages in handling large datasets and low-memory systems. CatBoost stands out with its ability to handle categorical variables automatically. LightGBM adopts a leaf-wise growth strategy, requiring fewer steps but more computational power. On the other hand, both XGBoost and CatBoost adopt a level-wise growth strategy, requiring more steps but needing less computational power. In Fig. 4, the difference between bagging and boosting is visualized [15].

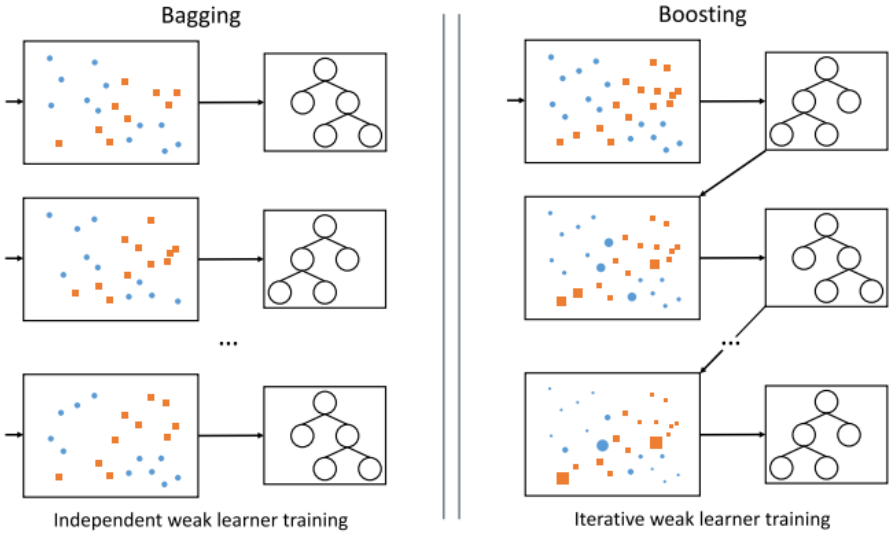

Bagging
Boosting
...
...
Independent weak learner training
Iterative weak learner training

Fig. 4. Bagging vs. Boosting [15].

In this study, predictions were made on Diabetes using ten state-of-art and traditional machine learning models (GBM, XGBoost, LightGBM, CatBoost, Random Forest, Decision Tree, K Nearest Neighbor, Logistic Regression, Gaussian Naïve Bayes, and Ridge Classifier), and the performances of these models were compared. Fig. 5 summarizes the study.

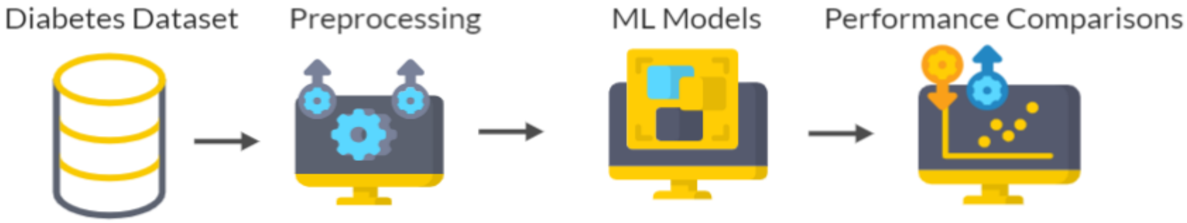


Fig. 5. Workflow of the study.

The ML algorithms used in our study are as follows:

Logistic Regression: Logistic regression is a statistical modeling technique used to predict the probability of a dependent variable. It was first introduced by David Cox in 1958 [16].

Gaussian Naive Bayes: Naive Bayes is a probability-based classification algorithm based on Bayes' theorem. The basic Naive Bayes algorithm became popular in the 1950s and later gained popularity for classification problems [17].

Decision Trees: Decision trees are a modeling technique based on tree structures for classifying data or solving regression problems. The foundation of decision trees dates back to the late 1960s [18].

K Nearest Neighbor (KNN): KNN, is a simple and effective algorithm used for classification and regression problems by examining the K nearest neighbors of a data point [19].

Ridge Classifier: The Ridge classifier is an extension of ridge regression and addresses classification problems [20].

Random Forests: Random Forests were introduced by Leo Breiman in a paper titled "Random Forests" in 2001. Random Forests are an algorithm that addresses complex classification problems by combining multiple decision trees [21].

Gradient Boosting Machines: Gradient boosting methods were developed in the late 1990s and gained popularity. They primarily combine weak learners to create a robust model [22].

XGBoost: XGBoost was introduced by Chen and Guestrin in 2014. XGBoost is a fast and effective gradient-boosting framework [23].

LightGBM: LightGBM, introduced by Microsoft in 2016, is a machine-learning library that provides fast and distributed gradient boosting for large datasets [24].

CatBoost: CatBoost, introduced by Yandex in 2017, is a gradient-boosting library that directly handles categorical variables and reduces overfitting [25].

Machine learning algorithms such as GBM, XGBoost, LightGBM, CatBoost, Random Forest, Decision Tree, K Nearest Neighbor, Logistic Regression, Gaussian Naïve Bayes, and Ridge Classifier represent various learning methods with different features and use case scenarios. Ensemble methods like GBM, XGBoost, LightGBM, CatBoost, and Random Forest create more robust and generalizable models by combining weak models. However, the critical differences among these algorithms emerge in terms of speed, scalability, categorical variable handling capabilities, and growth strategies. Decision Tree can classify or regress a dataset by creating a decision tree. K Nearest Neighbor performs classification or Regression based on the similarities of data points. Logistic Regression is used for binary classification problems, while Gaussian Naïve Bayes is a probability-based classification algorithm. Ridge Classifier, on the other hand, is a linear classifier that stands out, especially with regularization. During our study, we experimented with and compared the performance of these ten different machine-learning models for predicting diabetes.

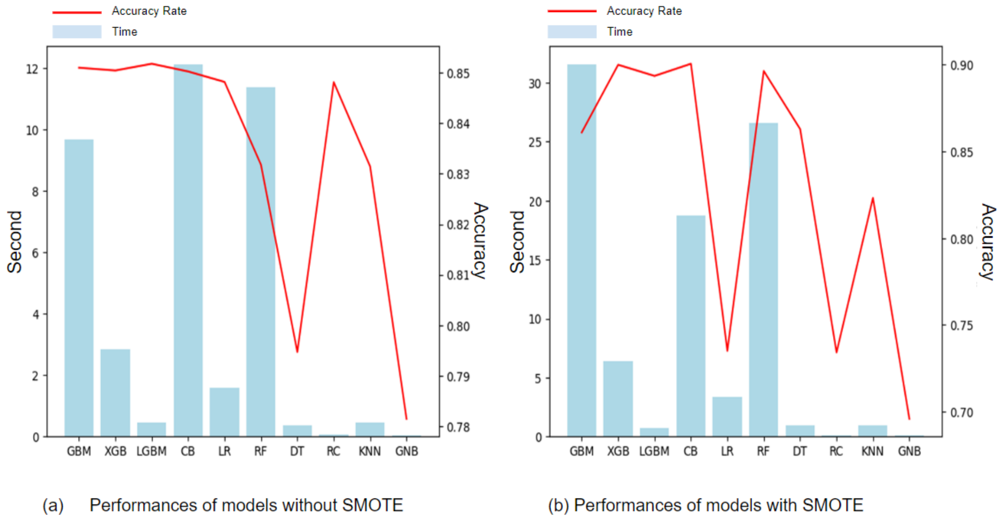
3. Experimental results

We conducted the experiments in two stages. Initially, we trained and tested the models on the dataset without applying the SMOTE technique. Subsequently, to enhance model performance, we applied the SMOTE technique and compared the accuracy rates of the models. We have presented both scenarios in Table 1. According to this table, we observed an improvement in the models' performance after applying SMOTE. In the comparisons conducted without SMOTE, the highest performance was achieved by the LightGBM model. However, when SMOTE was applied, the model with the highest performance had an accuracy rate of 0.9005, attributed to the CatBoost model.

Table 1. Performance comparisons of models.

Model without SMOTE	Accuracy Score	Model with SMOTE	Accuracy Score
LightGBM	0.8518	CatBoost	0.9005
GBM	0.8510	XGBoost	0.8999
XGBoost	0.8504	RF	0.8963
CatBoost	0.8502	LightGBM	0.8935
LR	0.8482	DT	0.8629
RC	0.8481	GBM	0.8610
RF	0.8318	KNN	0.8232
KNN	0.8314	LR	0.7351
DT	0.7947	RC	0.7342
GNB	0.7815	GNB	0.6959

In Fig. 6, the models' training times and accuracy rates are displayed graphically. According to this graph, when looking at the models trained and tested without applying SMOTE, the highest accuracy rate belongs to the LightGBM model, and the training speed of this model is higher than most of the other models. On the other hand, CatBoost has the highest accuracy rate for the models trained with SMOTE, but its training time is slower than the other models.



(a) Performances of models without SMOTE

(b) Performances of models with SMOTE

Fig. 6. Comparison of models' performance and training time.

The AUC (Area under the ROC Curve) – ROC (receiver operating characteristic curve) graph of the models applied to SMOTE is presented in Fig. 7. The AUC-ROC curve visualizes the sensitivity (true positive rate) and specificity (true negative rate) of the classification model against different threshold values.

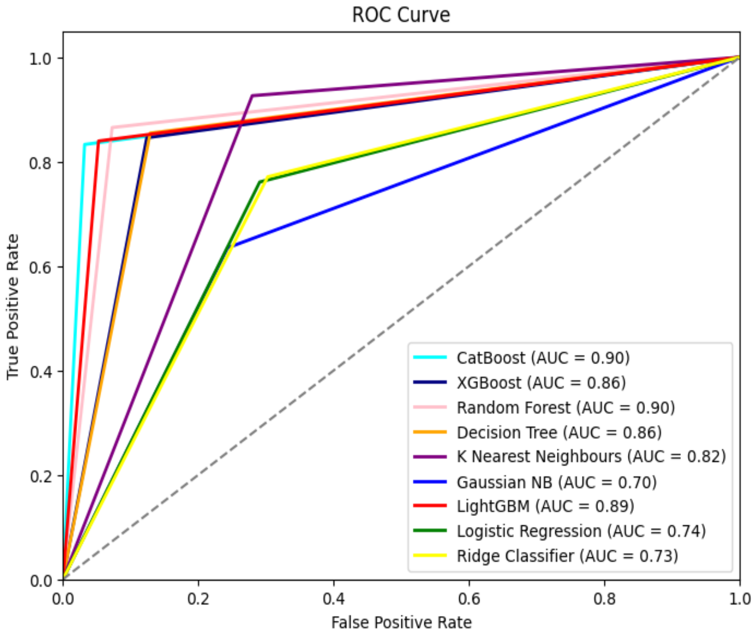


Fig. 7. AUC – ROC curve.

4. Conclusions

This study evaluated and compared the performance of 10 different machine learning models, including state-of-art models, to identify diabetes. Without applying the SMOTE technique to the dataset, the model that demonstrated the highest performance achieved an accuracy of 85.18%, attributed to the LightGBM model. Furthermore, it was noted that this model's training time was shorter than most of the other models. Due to the imbalanced nature of the dataset and the aim to enhance performance, applying the SMOTE technique to the data resulted in the CatBoost algorithm attaining the highest accuracy rate of 90.05%.

This comparative analysis of machine learning models provides valuable insights for selecting the most suitable model for addressing classification problems. Nevertheless, certain limitations and opportunities for future research should be acknowledged. Firstly, hyperparameter tuning and model optimization could enhance results. Expanding the dataset with more extensive and diverse sampling may improve the model's generalization capabilities. Collaborating with healthcare professionals and diabetes experts to interpret and validate features and outcomes from a clinical perspective is crucial. This interdisciplinary approach can unlock the potential of machine learning models to deliver real value in clinical applications.

In conclusion, this study successfully demonstrated the efficacy of machine learning models in diabetes identification. Future research endeavors should explore advanced methodologies and foster collaborations to achieve superior performance and increased applicability in clinical settings.

Acknowledgements

This study did not receive any dedicated funding from governmental, commercial, or non-profit organizations.

References

[1] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, "Epidemiology of diabetes and diabetes-related complications," *Phys. Ther.*, vol. 88, no. 11, pp. 1254-1264, Nov. 2008, doi: https://doi.org/10.2522/ptj.20080020.

[2] H. D.McIntyre, P. Catalano, C. Zhang, G. Desoye, E. R. Mathiesen, and P. Damm, "Gestational diabetes mellitus," *Natur. Rev. Dis. Prim.*, vol. 5, no. 1, pp. 47, Jul. 2019, doi: https://doi.org/10.1038/s41572-019-0098-8.

[3] İ. Akgül, Ö. Çağrı Yavuz, and U. Yavuz, "Deep Learning Based Models for Detection of Diabetic Retinopathy," *Tehnički glasnik,* vol. 17, no. 4, pp. 581-587, Dec. 2023, doi: https://doi.org/10.31803/tg-20220905123827.

[4] N. Yuvaraj and K.R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster," *Clust. Comp.*, vol. 22, no. 1, pp. 1-9, Jan. 2019, doi: https://doi.org/10.1007/s10586-017-1532-x.

[5] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Peer reviewed: building risk prediction models for type 2 diabetes using machine learning techniques," *Prev. Chro. Dis.*, vol. 16, Sep. 2019, doi: 10.5888/pcd16.190109.

[6] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," *in 2018 IEEE 4th World Forum on Int. of Things*, 2018, pp. 291-295, doi: 10.1109/WF-IoT.2018.8355130.

[7] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," *in 2019 1st Inter. Inform. and Soft. Eng. Conf.*, 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.

[8] Diabetes Health Indicators Dataset, https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset (accessed December 12, 2023).

[9] M. Crowther, W. Lim, and M. A. Crowther, "Systematic review and meta-analysis methodology," *The Jour. of the Amer. Soc. of Hemat.*, vol. 116, no. 17, pp. 3140-3146, Oct. 2010, doi: https://doi.org/10.1182/blood-2010-05-280883.

[10] S. Pandey, "Principles of correlation and regression analysis," *Jour. of the prac. of cardio. Sci.*, vol. 6, no. 1, pp. 7-11, Apr. 2020, doi: 10.4103/jpcs.jpcs_2_20.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Jour. of artif. Intel. Res.*, vol. 16, pp. 321-357, Jun. 2002, doi: https://doi.org/10.1613/jair.953.

[12] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. of Comp. Sci.*, vol. 14, pp. 241-258, Apr. 2020, doi: https://doi.org/10.1007/s11704-019-8208-z.

[13] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123-140, Aug. 1996, doi: https://doi.org/10.1007/BF00058655.

[14] R. E. Schapire, "A brief introduction to boosting," *Ijcai*, Vol. 99, No. 999, pp. 1401-1406, 1999.

[15] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Infor. Fus.*, vol. 64, pp. 205-237, Dec. 2020, doi: https://doi.org/10.1016/j.inffus.2020.07.007.

[16] D. R. Cox, "The regression analysis of binary sequences," *Jour. of the Roy. Stat. Soc. Ser. B: Stat. Meth.*, vol. 20, no. 2, pp. 215-232, Jul. 1958, doi: https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

[17] S. Jayachitra, and A. Prasanth, "Multi-feature analysis for automated brain stroke classification using weighted Gaussian naïve Bayes classifier," *Jour. of Circ. Sys. and Comp.*, vol. 30, no. 10, pp. 2150178, 2021, doi: https://doi.org/10.1142/S0218126621501784.

[18] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nat. Biotech.*, vol. 26, no. 9, pp. 1011-1013, Sep. 2008, doi: https://doi.org/10.1038/nbt0908-1011.

[19] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, pp. 6256, Apr. 2022, doi: https://doi.org/10.1038/s41598-022-10358-x.

[20] S. Priyadarshinee and M. Panda, "Cardiac disease prediction using smote and machine learning classifiers," *Jour. of Pharma. Neg. Res.*, vol. 13, no.8, pp. 856-862, Nov. 2022, doi: https://doi.org/10.47750/pnr.2022.13.S08.108.

[21] L. Breiman, "Random forests," *Mac. Learn.*, vol. 45, pp. 5-32, Oct. 2001, doi: https://doi.org/10.1023/A:1010933404324.

[22] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. in Neurorob.*, vol. 7, pp. 21, Dec. 2013, doi: https://doi.org/10.3389/fnbot.2013.00021.

[23] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, and T. Zhou, "Xgboost: extreme gradient boosting," R package version 0.4-2, vol. 1, no. 4, pp. 1-4, 2015.

[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, and T. Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. in Neur. Info. Process. Sys.*, 30, 2017.

[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv. in Neur. Info. Process. Sys.*, 31, 2018.