INVESTİGATION OF THE EFFECT OF FEATURE SELECTİON AND HYPERPARAMETER OPTİMİZATION METHOD ON MACHİNE LEARNİNG ALGORİTHMS IN THE DIAGNOSİS OF DIABETES

Conference Paper · July 2023

CITATIONS
0

READS
31

2 authors:

Erkan Akkur
TOBB University of Economics and Technology
9 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Fuat Türk
Kirikkale University
17 PUBLICATIONS **68** CITATIONS
SEE PROFILE

INVESTİGATION OF THE EFFECT OF FEATURE SELECTİON AND HYPERPARAMETER OPTİMİZATION METHOD ON MACHİNE LEARNİNG ALGORİTHMS IN THE DIAGNOSİS OF DIABETES

Erkan AKKUR[1], Fuat TURK[2]
0000-0001-5573-5096-0000-0001-8159-360X
eakkur@gmail.com-fturk@kku.edu.tr
[1]Turkish Medicines and Medical Devices Agency, Ankara, Turkey
[2]Department of Computer Engineering of Kırıkkale University, Kırıkkale, Turkey

Abstract

Diabetes is one of the most important diseases threatening human health today. Early diagnosis is important for the prevention of larger diseases that this disease can cause. In recent years, machine learning algorithms have been used frequently in the diagnosis of this disease, as in the early diagnosis of many diseases. The use of feature selection and hyperparameter optimization methods significantly affects the success rates of machine learning algorithms. In this study, the effects of hyperparameter optimization and feature selection methods on machine learning algorithms are investigated. AdaBoost and Gradient Boost are used as machine learning algorithms. Anova was used as feature selection method and Grid Search is used as hyperparameter optimization method. Various experiments are carried out to confirm the feasibility of the proposed method. These experiments are carried out in two stages. In the first step, machine learning algorithms are tested before using feature selection and hyperparameter optimization. Then, machine learning algorithms were tested after using feature selection and hyperparameter optimization. During the experiments, the 5-fold cross-validation technique is used to resolve the bias of the models. Accuracy, precision, recall, F1-score, AUC-ROC performance criteria are used to evaluate the effectiveness of the method. The results obtained after applying the suggested feature selection and hyperparameter algorithm are promising. As a result, the Gradient Boost method for diagnosing diabetes showed a good prediction rate with 89.19% accuracy, 85.71% precision, 82.24% recall, 84.25% F1-score and 0.88 AUC-ROC.

Keywords: Diabetes, machine learning, feature selection, hyperparameter optimization, grid search.

1. Introduction

Diabetes is defined as metabolic disorder characterized by high blood sugar. It may lead to serious problems such as cardiovascular diseases, visual impairment, and kidney diseases. According to the global statistics, the

incidence rate of this disease is increasing day by day. It is estimated that there are approximately 500 million diabetics in worldwide (Maniruzzaman, Rahman, Ahammed & Abedin, 2020). Thus, early detection is very critical in terms of diminishing the incidence rate and the problems that this disease can cause. Thanks to advances in machine learning (ML) and artificial intelligence, early-stage disease detection and diagnosis is more efficient than manual diabetes diagnosis. Computer-aided expert systems based on ML have a very important place in the early diagnosis of this disease (Chaki, Ganesh, Cidham, & Theertan, 2022). The feature selection (FS) methods and hyperparameter optimization (HO) are two significant issues that affect the classification performance of ML algorithms. FS methods are defined as the process of selecting and finding the most useful features in data sets. These methods reduce the size of the feature set in data sets and increase the speed and success rates of ML algorithms (Miao & Niu, 2016). ML algorithms contain many hyperparameters. Hyperparameters are parameters that the model cannot learn and must be provided before the training process. Adjusting the appropriate hyperparameters improves the performance of ML algorithms. HO is the determination of the most suitable hyperparameter combination for ML algorithms (Claesen & De Moor, 2015). As a result, the use of FS and HO increases the success rates of ML algorithms.In this study, a classification system based on ML algorithms is proposed for the prediction of diabetes. In addition, the effect of FS and HO on ML algorithms is investigated.

2. Related Works

When the literature studies on the subject are analyzed, it is seen that many classification models of diabetes mellitus have been proposed using machine learning algorithms. Some studies for the prediction of diabetes are as follows. Mohan and Jain (2020) utilized the Support Vector Machine (SVM) algorithm with different core functions for the prediction of diabetes disease. The proposed method was tested on the PIMA Indians Diabetes dataset and the radial-based core function SVM technique achieved the highest accuracy of 0.82. Sivaranjani et al. (2021) tried to predict diabetes by using feature selection method-based machine learning algorithms. They preferred the Principal Component Analysis method as the feature selection method. For classification, random forest and support vector machine algorithms are used. The Random Forest algorithm achieved 83% accuracy. Khaleel and Al-Bakry (2023) compared the performance of many ML algorithms to predict whether diabetes is present and showed that the Logistic Regression algorithm is more efficient than other algorithms in predicting diabetes. Mercaldo et al. (2017) used six different classification algorithms for the prediction of diabetes. Among these algorithms, the Hoeffding Tree algorithm reached the highest result with 77% accuracy and 77.5% recall rate. Birjais et al. (2019) tried to predict diabetes on the PIMA Indians dataset using different machine learning algorithms. Gradient boosting algorithm reached the highest result with 86% accuracy.

3. Materials and methods

This section presents the proposed methodology for the prediction of diabetes disease. The suggested methodology is shown in Figure 1. In the first step of the proposed method, the diabetes disease dataset is passed through a series of preprocessing steps. Next, the proposed method consists of two parts. In the first part, the features found from the data set are given as input data to the models without using any feature selection and hyperparameter optimization. In the second part, the FS method is applied to determine the most suitable features. After this stage, suitable features are given as input to the ML algorithm. In the next stage, HO is performed with grid search in order to increase the efficiency of the models. Finally, the effects of FS and HO on the models were examined by comparing the two proposed sections with each other.

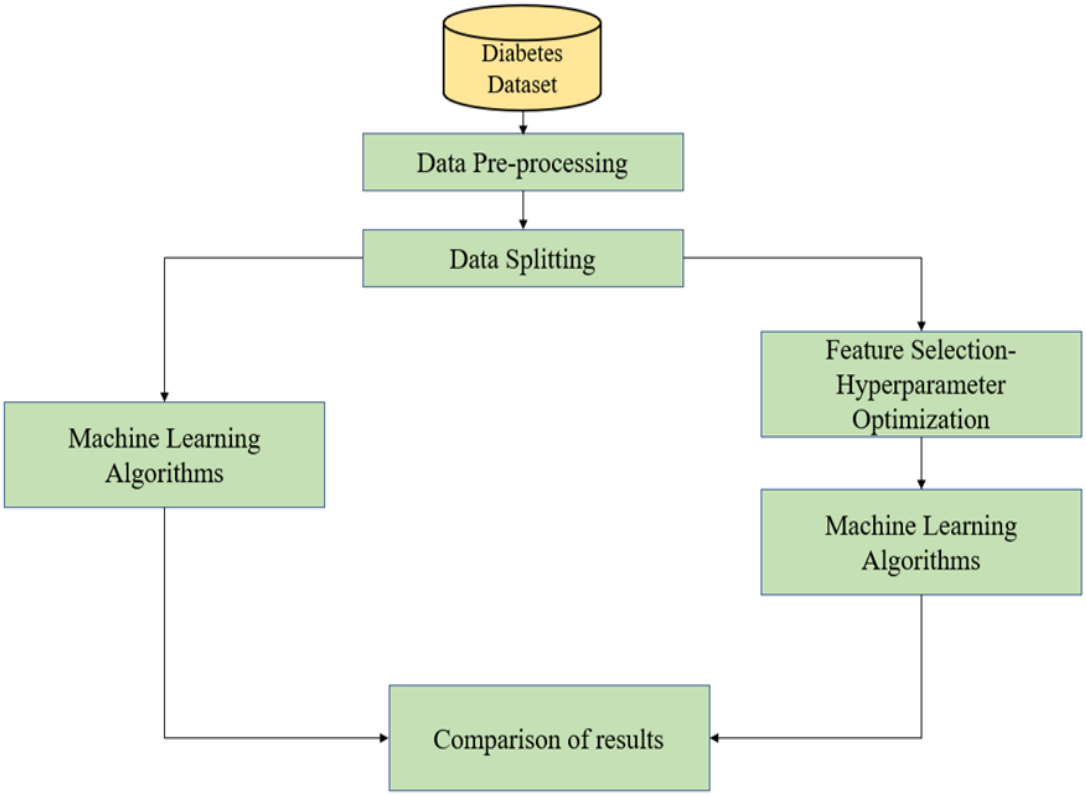

Figure 1: The suggested methodology for prediction of diabetes

Diabetes Dataset

In this study, the PIMA Indians Diabetes dataset, one of the datasets available for free use in the UCI Machine Learning Repository, was used. There are 768 samples in total in the data set. Diabetes is seen in 268 of them, while this disease is not seen in 500 of them (Pima Indians Diabetes Database, Kaggle). A statistical summary of the data set is given in Table 1.

Table 1: The statistical summary of PIMA Indians Diabetes Dataset

Attributes	Mean	Std.	Minimum	Maximum	Zero Values
Pregnancies	3.84	3.37	0	17	0
Glucose	120.89	31.97	0	199	5
Blood Pressure	69.1	19.35	0	122	35
Skin Thickness	20.53	15.95	0	99	227
Insulin	79.8	115.24	0	846	374
Body Mass Index	32	7.88	0.07	67.1	11
Diabetes Pedigree Function	0.47	0.33	21	2.42	0
Age	33.24	11.76	0	81	0

Data pre-processing

Data preprocessing is the first step in making diabetes dataset raw data available for the classification process. (Joshi & Patel, 2021). When the data set was examined, it was seen that there was no missing value in the data set and some features had zero values. In general, the range of Glucose, Insulin, BMI, and blood pressure can never start from zero. Therefore, the assignment operation is necessary to fill in the missing values. It is a technique that replaces missing data with some substitute values to preserve most of the data/information in the dataset. Values with missing data in the dataset were filled using the mean and median values of the features. In the next step, feature scaling was applied to the datasets whose missing data were completed. Feature scaling is one of the key issues in preprocessing before it can be fitted into ML algorithms. This process can make a weak machine learning algorithm better. Min-max scaling technique was used for feature scaling in this study (Ahsan et al., 2021).

Machine learning algorithms

Adaptive Boosting (AdaBoost) and Gradient Boosting ML algorithms are used to predict diabetes. AdaBoost is a boosting algorithm that generally uses decision trees for modelling. It is based on the principle of creating multiple sequential models, each of which corrects the errors in the final model. AdaBoost weights the incorrectly estimated observations, and the next model tries to predict these values correctly. Gradient Boosting

algorithm is formed by combining decision trees with machine learning algorithms. This algorithm basically aims to reduce the errors to the lowest level by creating a decision tree (Sai et al. 2023).

Feature Selection

Feature selection methods aim to reduce the feature size by selecting the most useful and important features in a data set. This process significantly increases the success rates of ML algorithms. ANOVA F-test classification processes are preferred if the features in the data are numerical, and the target is categorical. ANOVA F-test is preferred as the FS in this study since the PIMA Indians diabetes dataset also complies with these definitions. ANOVA is a parametric statistical hypothesis to determine whether the means from two or more data samples (usually three or more) come from the same distribution. known as test. An F-statistic or F-test is a class of statistical tests that calculates the variance of two different samples, or the ratio between values of variance explained by a statistical test, such as ANOVA, and values of variance, such as unexplained variance (Maniruzzaman et al., 2023).

Hyperparameter Optimization

It is very important to choose the right hyperparameter model in order to design a powerful machine learning algorithm. The performances of machine learning algorithms increase with HO. Hyperparameters can be defined as parameters that algorithms cannot learn beforehand and must be provided before the training process. HO is to find the most suitable hyperparameter combination according to the success criteria determined for a machine algorithm. In this study, the hyperparameter optimization method was chosen. Grid search approach is one of the most used HO methods. It is a technique used in HO to find the values that give the best performance in a given model. In this technique, the model is trained with all the parameter combinations given by the user, and it is an important process as the best parameters found affect the performance of the whole model. However, overfitting may occur during the optimization process. The overfitting problem can be reduced by applying the Cross validation (CV) method. The CV technique trains a model with a dataset and tests it with several datasets. A CV vs Grid search (GSCV) is used to determine the best combination of learning. Then, the set of parameter combinations with the highest accuracy is selected for each algorithm. After the best parameter set is selected, the estimation process of the data begins. Using the K-fold CV technique, the dataset is divided into training and testing parts. The 5-fold CV method is used for ten different training and test sets. The 5-fold CV method is used for each data set to determine the mean of the diabetes estimate. By using the grid search and CV model together, HO is obtained because of various experiments (Belete & Huchaiah, 2022).

4. Experimental Results

In this study, 5-fold cross validation method is used to divide the data set into training and test sets. Accuracy, precision, recall and F1-score metrics are used to evaluate the classification performance of machine learning algorithms, respectively. The classification results obtained without using the FS method and HO are shown in Table 2.

Table 2: The 5-fold cross validation for all models before feature selection and hyperparameter optimization

Models	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
AdaBoosting Classifier	75.39	66.52	59.32	62.72
Gradient Boosting Classifier	76.17	67.63	60.82	64.04

Figure 2 depicts the importance degrees of the features after applying the ANOVA F-test method to the diabetes dataset. In this method, high-value features have more importance for the classification process. As can be seen in Figure 2, the "Skin Thickness" and "Blood Pressure" attributes have lower importance than the other attributes. Therefore, these two features are excluded from the dataset.

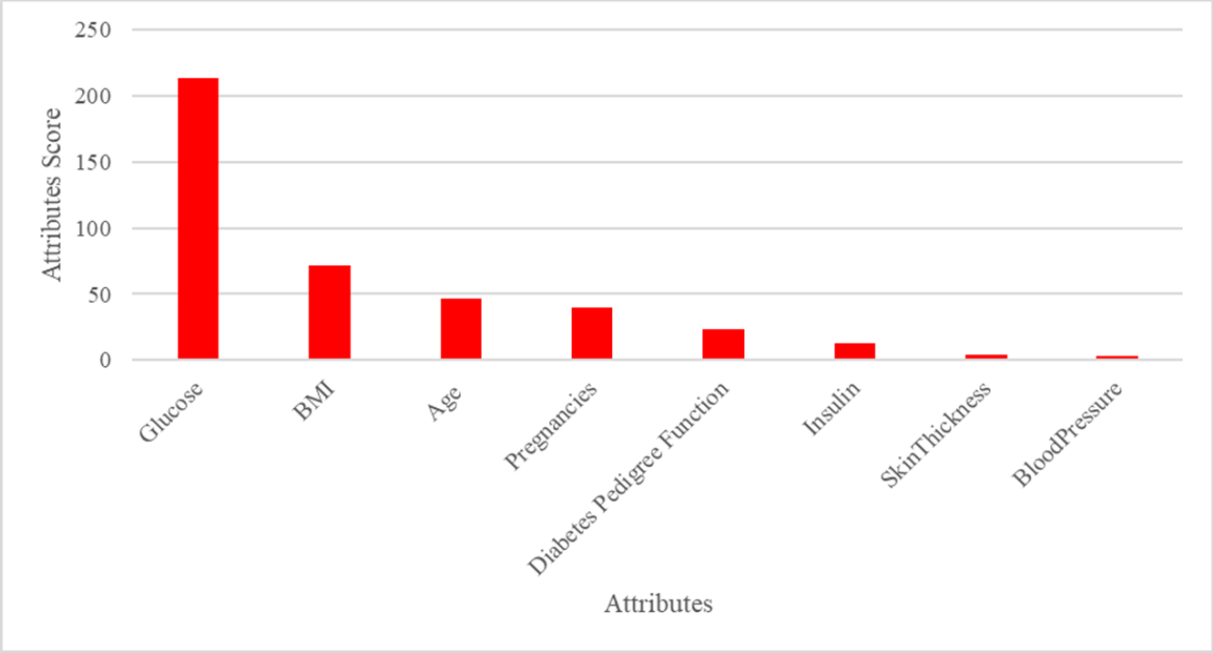

Figure 2: The importance degrees of attributes after applying ANOVA F-Test

The hyperparameters, search intervals and the best parameters obtained after applying the Grid Search method of the machine learning algorithms are presented in Table 3.

Table 3: The hyperparameters, search intervals and the best parameters of ML algorithms

Algorithm	Hyperparameters	Search Range	Best Parameter
Gradient Boosting	"n_estimators"	[5-500]	250
	"learning rate"	[0.1-1]	0.1
	"max_depth"	[1-9]	3
	"max_features"	[0.3-1]	1
	"min_samples_split"	[2-4]	2
Ada Boosting	"n_estimators"	[5-500]	76
	"learning rate"	[0.1-1]	0.7

The classification results obtained after applying the feature selection method and hyperparameter optimization are shown in Table 4.

Table 4: The 5-fold cross validation for all models after applying feature selection and hyperparameter optimization.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AdaBoosting Classifier	86.85	82.24	79.48	80.83
Gradient Boosting Classifier	89.19	85.71	82.84	84.25

5. Discussion

In this study, experiments designed to predict diabetes effectively were conducted in two stages. In the first step, the experiment is carried out with the default parameters of the models. In the second stage, the experiment is carried out by applying the feature method and hyperparameter optimization, respectively. The FS and HO effect on AdaBOOST and Gradient Boosting algorithms are shown in Figure 3. Accordingly, feature method-hyperparameter optimization significantly increased the classification performances of both algorithms.

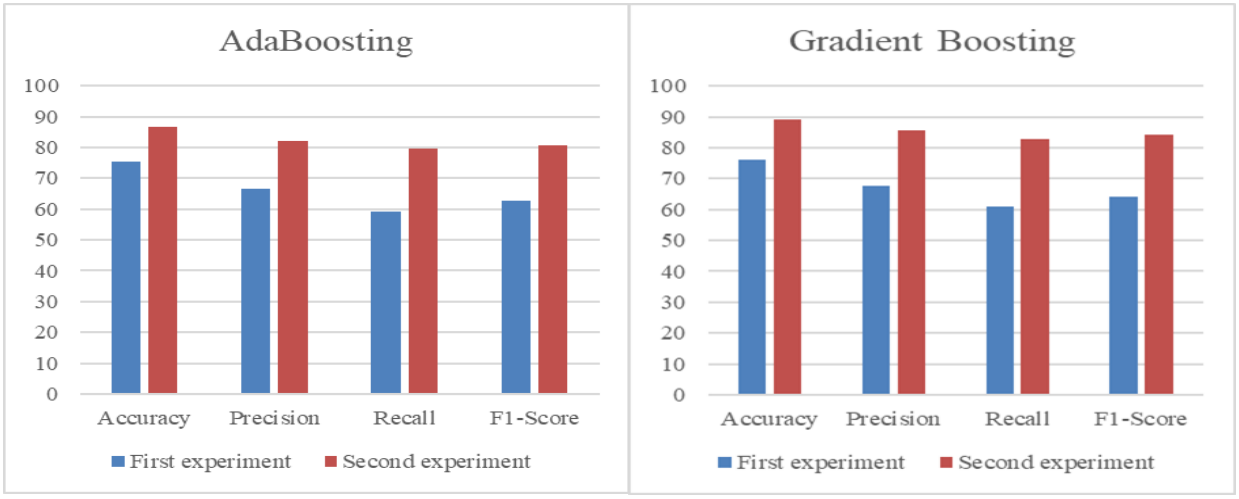

Figure 3: The classification results of AdaBoosting and Gradient Boosting algorithm

In Figure 4, the AUC-ROC metric is used to compare the classification performances of AdaBoosting and Gradient Boosting algorithms. As a result of the comparison, the Gradient Boosting algorithm shows better result than the AdaBoosting algorithm with a value of 89.19% of accuracy, 85.71% of precision, 82.24% of recall, 84.25% of F1-score and 0.88 AUC-ROC for the prediction of diabetes.

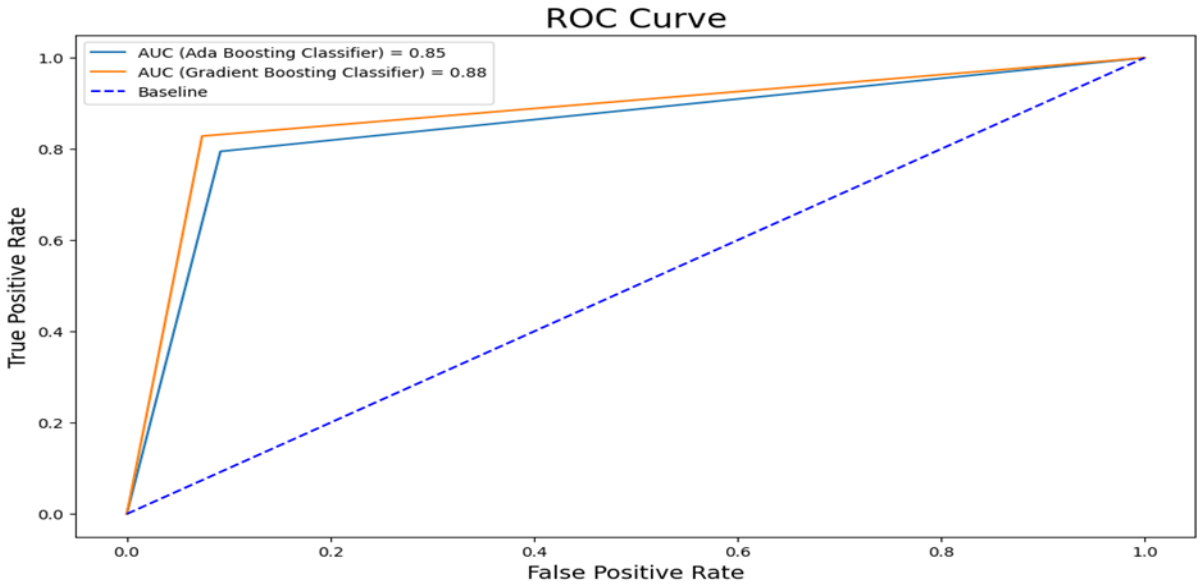

Figure 4: AUC-ROC curve of the classification results of AdaBoosting and Gradient Boosting algorithm

6. Conclusion

Early detection of diabetes and timely intervention are critical to prevent the disease from causing further problems. Recently, ML algorithm-based models have taken a place in the prediction of diabetes. In this study, the prediction is performed on a diabetes dataset using AdaBoost and Gradient Boosting algorithms. In order to

increase the classification rates of these algorithms, feature selection method and hyperparameter optimization are applied. With the use of feature selection method and hyperparameter optimization, the classification rates of machine learning algorithms have been increased significantly. When the classification performances of the Gradient Boost and AdaBoost algorithms are compared, the Gradient Boost algorithm achives a better result for the prediction of diabetes with 89.19% of accuracy, 85.71% of precision, 82.24% of recall, 84.25% of F1-score and 0.88 AUC-ROC.

References

Maniruzzaman, M., Rahman, M. J., Ahammed, B., Abedin, M. M. (2020). "Classification and prediction of diabetes disease using machine learning paradigm". *Health information science and systems*, *8*, 1-14. https://doi.org/10.1007/s13755-019-0095-z.

Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2022). "Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review". *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 3204-3225. https://doi.org/10.1016/j.jksuci.2020.06.013

Miao J., Niu. L., (2016). "A survey on feature selection". *Procedia Comput. Sci.*, 91, 919-926. https://doi.org/10.1016/j.procs.2016.07.111

Claesen M., De Moor B., (2015). "Hyperparameter search in machine learning". *arXiv preprint* arXiv:1502.02127. https://doi.org/10.48550/arXiv.1502.02127

Mohan, N., Jain, V. (2020). "Performance analysis of support vector machine in diabetes prediction". *In: International Conference on Electronics, Communication and Aerospace Technology*, pp. 1–3 (2020).

Sivaranjani, S., Ananya, S., Aravinth, J., & Karthika, R. (2021, March). "Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction". *In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 141-146). IEEE.

Khaleel, F. A., & Al-Bakry, A. M. (2023). "Diagnosis of diabetes using machine learning algorithms". *Materials Today: Proceedings*, 80, 3200-3203.

Mercaldo, F., Nardone, V., & Santone, A. (2017). "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques". *Procedia computer science*, 112, 2519-2528. https://doi.org/10.1016/j.procs.2017.08.193

Birjais R, Mourya AK, Chauhan R, Kaur H, (2019). "Prediction and diagnosis of future diabetes risk: A machine learning approach". *SN Appl Sci*., 1:1–8.

Pima Indians Diabetes Database, Kaggle, https://www.kaggle.com/datasets/uciml/pima-indiansdiabetes-database/ Accessed 09 May, 2023.

Joshi, AP, Patel BV, (2021). "Data preprocessing: The techniques for preparing clean and quality data for data analytics process". *Orient. J. Comput. Sci. Technol*, 13(0203), 78-81.

Ahsan MM, Mahmud MP, Saha PK, Gupta KD, (2021). Siddique Z, "Effect of data scaling methods on machine learning algorithms and model performance". *Technologies*, 9(3):52.

Sai, M. J., Chettri, P., Panigrahi, R., Garg, A., Bhoi, A. K., & Barsocchi, P. (2023). "An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes" *International Journal of Computational Intelligence Systems*, 16(1), 14.

Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). "Classification and prediction of diabetes disease using machine learning paradigm." *Health Inf. Sci. Syst.*, 8, 1-14.

Belete DM, Huchaiah MD, (2022). "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results". *Int. J. Comput. Appl*, 44(9): 875-886.