Diabetes Prediction using Classification Methods

Abhishek Karmakar[1[0000-0002-4813-9568]] and Sharik Gazi[1[0000-0001-5368-9508]] and Varsha Singh[1[0000-0003-4777-1278]]

[1] Indian Institute of Information Technology Allahabad, Prayagraj, U.P.-211015, India.
varshagaur@gmail.com

Abstract. When the body organ pancreas are not capable enough of producing sufficient amount of insulin in the body, sugar level in the bloodstream increases. This results in serious health related issues. The deregulation of glucose in the body give rise to chronic disease denotes as Diabetes. Diabetes complications cause kidney problems, heart strokes, nerve issues and so on and so forth. Getting an aid or a precaution at an early level can reduce the chance of Diabetes. This work on diabetes prediction can help in the improvement of human life style. Our motive is to predict the outcome whether or not the patient is suffering from diabetes on the basis of basic human features such as glucose level, blood pressure, insulin, BMI (Body Mass Index), number of pregnancies, skin thickness, diabetes pedigree function and age. To achieve our goal, we have used the famous Pima Indian type-2 Diabetes Mellitus Classification Dataset. In this paper we did an analysis using twelve machine learning algorithms and compared the results with others.

Keywords: Diabetes complications, Diabetes Prediction, Pima Indian type-2 Diabetes Mellitus Classification Dataset, Machine Learning

1 Introduction

Our brain needs glucose as its prime source of energy and so are the tissues and muscles of our body. Potentially other than normal blood sugar level in our body can lead to a metabolic disorder. The uncontrolled condition of this disease leads to the damage of nerves, eyes, blood vessels and can cause neuropathy. The excess of the glucose into the blood stream causes a type-1 or type-2 diabetic chronic condition. The type-2 condition is also denoted as adult-onset as its non-insulin dependent whereas the type-1 condition is called as childhood-onset. As per the survey made by WHO (World Health Organization) 95% of people are suffering from type-2 diabetes. Also 9 million were suffering from type-1 condition in 2017 [1]. According to the survey [1] recorded from 2000 to 2019, the mortality rate by age has rose by 3% for diabetic people. It has been observed that people over 30 years of age are more likely to suffer from type-2 diabetic disorder. Over the past few years, in the list of global health problems the type-2 diabetes mellitus has been counted to be a major problem[2].

As the general style of living is changing gradually, diabetes has become a common disease. So an early detection can be really helpful. To seek out a solution, in this paper we had analyzed the prediction of type-2 diabetes mellitus using twelve different machine learning algorithms like Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, Artificial Neural Network, Adaboost. Logitboost, Xgboost and K-Means and evaluated them using accuracy, precision, recall, f1-score, support, ROC curve and AUC value.

The outline of this paper has been organized as follows: Section II presents the details of the PIMA Indian Diabetes Dataset; Section III explains the related works till date; Section IV describes the proposed architecture, Section V consists of the comparison of our work to others. Lastly, Section VI concludes the paper.

2 About the Dataset

In this experiment we have worked on the prediction of type-2 diabetes disorder using the PIMA Indian Diabetes Dataset. This dataset consists of the record of type-2 diabetic and non-diabetic people living in U.S. and Mexico [3]. The various health parameters examined were age, insulin, glucose, blood pressure level, number of pregnancies, skin thickness, Body Mass Index and diabetes pedigree function. The record consists of 268 diabetic and 500 non-diabetic people. The statistical description of the numerical data for each feature is shown in Fig 1.

	count	mean	std	min	10%	25%	50%	75%	90%	95%	99%	max
Pregnancies	768.0	3.845052	3.369578	0.000	0.000	1.00000	3.0000	6.00000	9.0000	10.00000	13.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	85.000	99.00000	117.0000	140.25000	167.0000	181.00000	196.00000	199.00
Blood_Pressure	768.0	69.105469	19.355807	0.000	54.000	62.00000	72.0000	80.00000	88.0000	90.00000	106.00000	122.00
Skin_Thickness	768.0	20.536458	15.952218	0.000	0.000	0.00000	23.0000	32.00000	40.0000	44.00000	51.33000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.000	0.00000	30.5000	127.25000	210.0000	293.00000	519.90000	846.00
BMI	768.0	31.992578	7.884160	0.000	23.600	27.30000	32.0000	36.60000	41.5000	44.39500	50.75900	67.10
Diabetes_Pedigree_Function	768.0	0.471876	0.331329	0.078	0.165	0.24375	0.3725	0.62625	0.8786	1.13285	1.69833	2.42
Age	768.0	33.240885	11.760232	21.000	22.000	24.00000	29.0000	41.00000	51.0000	58.00000	67.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.000	0.00000	0.0000	1.00000	1.0000	1.00000	1.00000	1.00

Fig. 1. Statistical Description of the data.

The PIMA Indian Diabetes Dataset is widely used in research, but like many datasets, it has potential biases and limitations that can influence the conclusions drawn from it. Here are some potential biases present in the dataset:

1. **Geographical Bias:** The dataset is collected from Arizona, USA. Therefore, predictions made using this dataset may not be generalizable to populations from other geographical regions.
2. **Ethnic Bias:** The dataset is primarily focused on the PIMA Indian population. As a result, the findings may not necessarily extend to other ethnic or racial groups.

3. **Potential Outcome Imbalance:** The dataset contains a specific number of diabetic and non-diabetic individuals. With 268 diabetic and 500 non-diabetic records, there's an imbalance in the dataset. The sampling method used to select these individuals could introduce bias. This could potentially affect the performance of machine learning algorithms, making them more sensitive to predicting the majority class (in this case, non-diabetic).
4. **Gender Bias:** The dataset is a collection of all female having at least 21 years of age having type-2 diabetes. This bias found in the dataset as it lacks information of male patients.

Privacy considerations in data handling with respect to PIMA Indian Dataset:

1. **Data Privacy and Anonymization:** Despite the public availability and frequent use of the PIMA dataset, the individual identities remain paramount. In our utilization of this dataset, we adhered rigorously to best practices in data privacy, even though the dataset has undergone de-identification processes. Such steps are crucial in upholding the ethical standards associated with data-driven research and maintaining the trust of the communities under study.
2. **Bias and Representation Concerns:** The specificity of the PIMA dataset to the Pima Indian population introduces inherent biases that must be acknowledged. While this dataset provides valuable insights into the health and demographic trends of this particular group, it is imperative to approach the generalization of these results with caution. Findings from this dataset, though rigorous, are principally reflective of the Pima Indian community. Consequently, extrapolating these outcomes to wider ethnic or demographic contexts may not be directly applicable and requires careful consideration.

3 Related Works

Classification algorithms are very much popular in the medical field. It helps in estimating the future trend. Likewise, for the prediction of diabetes, the proposal of machine learning approaches has been a part of several research works. Machine learning techniques reduce the time consuming process performed by any human for prediction of diabetes. As the diagnosis of diabetes is complex, the application of Machine Learning algorithms enhances the accurate judgment. In the field of Machine Learning researchers has performed various approaches for diabetes prediction using the PIMA Indian Dataset. Our works also revolves around the same.

A proposal of diabetes prediction using an Artificial Neural Network has been shown by researchers in [4]. The diabetes dataset chosen was from Naokhali Medical College, Bangladesh. Another work on hospital physical examination data in Luzhou, China has been proposed for diabetes prediction using PCA (Principal Component

Analysis) [5]. Authors in [6] and [7] had also discussed on approaches for prediction of diabetes.

The paragraph above has showed a few work done on Diabetes prediction other than PIMA dataset whereas this paragraph is only focused to demonstrate the research work performed on Pima Indian Diabetes Dataset. In a recent work on the same, researchers had achieved accuracies of 81.16% by implementing Logistic Regression and Support Vector Machine, 77.92% using KNN, 77.07% using Decision tree and 77.92% by implementing Random forest and neural network [8]. In another work the writers has achieved accuracies of 77.5% by implementing Logistic Regression, 77.9% through Support Vector Machine, 76% using KNN, 75.8% using Decision tree, 79.7% by implementing Random forest, 79.2% using gradient boosting and 78.4% by artificial neural network [9]. The work showed in [10] has a remark as achieved 77% in using Logistic Regression and KNN on the Pima Dataset. On the same dataset another work [11] has accomplished an accuracy of 78.8% through artificial neural network. Paper [12] presents the model KNN which helps to get 83.76%. Unlike these works, authors of [13] has achieved an accuracy level of 89% using Support Vector Machine, 88% using KNN and 86% using the artificial neural network. Paper [14] shows that Naïve Bayes and Decision Tree are also an efficient model that can be useful for the prediction of diabetes. They proclaim to have 79.56% and 76.95% of correct classification output. Through neural network in [15] the publishers has put 88.41% as accuracy. In the next work [16] the accuracies are 78.64% using Decision Tree and 66.40% using gradient boosting, 77.86% using adaboost and 77.49% using logitboost. In [17] J48 and CART shows an accuracy of 78.95% and 78.64%. In [18] Support Vector was useful to show the accuracy to be 78%. In the research work of [19] three algorithms were used. Naïve Bayes showed 76.30%, 65.10% for Support Vector and 73.82% for Decision Tree. In the work of [20] in total nine algorithms has been used. In [20] the extreme gradient boosting showed an accuracy of 80.52% whereas gradient boosting showed an accuracy of 76.62%. The multinomial Naïve Bayes had an accuracy of 68.83 and Gaussian naïve Bayes showed an 80.52%. Logistic Regression was useful to achieve 81.92% of accuracy level; Support Vector was 83.12%, 81.82% for KNN, 77.92% for Decision Tree, 75.32% for Random Forest, 84.42% by using neural network and 76.62% by adaboost classifier. Authors in [21] have demonstrated accuracies of 74.78%, 79.57% and 78.67% for decision tree, random forest and Naïve Bayes when all the features have been taken in consideration. Similarly, accuracies of 75.22% 75.22% and 79.13% for decision tree, random forest and Naïve Bayes when 3 features have been taken in consideration. Lastly, for 5 features, accuracies obtained are 75.65%, 73.91% and 77.83%. In the comparison table, only the highest accuracies have been mentioned for the models that have been trained on 3, 5 and all features.

Our works present the following accuracies – Naïve Bayes – 87.66%, Logistic Regression – 89.61%, SVM – 88.96%, KNN – 88.96%, Decision Tree – 88.31%, Random Forest – 90.90%, Gradient Boosting – 91.55%, Neural Network – 84.41%, Ad-

aboost – 89.61%, Logitboost – 90.25%, Xgboost – 90.90%, K-Means Clustering – 87.66%.

4 Proposed Work

To perform a correct classification, the most important step which is to be performed is data pre-processing. After handling the improper data, normalizing the data, splitting it into test and train, then the data is fed into the data modelling section. The overall structure is shown below in Fig 2.

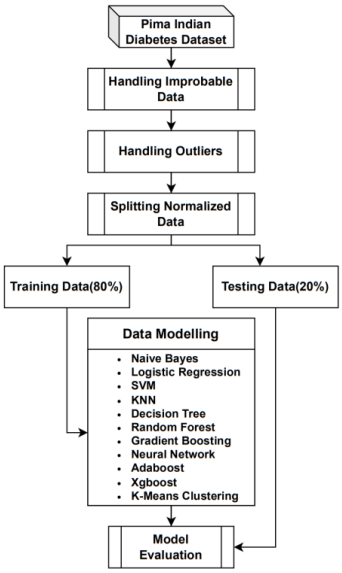

Fig. 2. Outline of the overall methodology

4.1 Handling Improbable Data

Data pre-processing is a technique which the data is prepared to enhance the model performance. In the given dataset there are attributes such as glucose level, blood pressure level, skin thickness and BMI which cannot be zero. So as in for the first step we had replaced all such insignificant values to Nan (or Null). Samples consisting of such insignificant data have been pictured in Fig. 3.

The dataset now consists of the Nan values, also it contains outliers. So the best way to replace all the Nan values is with median of the data. For each attribute the Nan values has been replaced with the median value pertaining to diabetic or non-diabetic patient. For example, the median glucose value for non-diabetic people is 107 whereas for diabetic people the median value is 140. Table 1 shows median value for the category of diabetic and non-diabetic people.

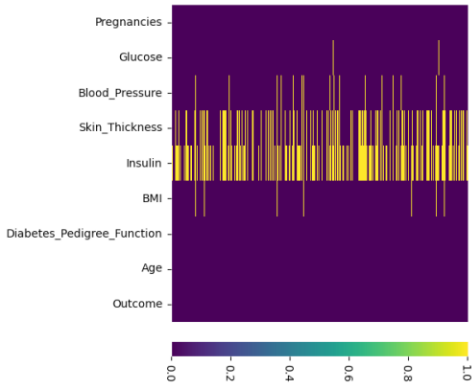

Fig. 3. Samples consisting of the insignificant value

Table 1. Table captions should be placed above the tables.

Health Parameters	**Outcome = 0**	**Outcome = 1**
Glucose	107.0	140.0
Insulin	102.5	169.5
Skin Thickness	27.0	32.0
Blood Pressure	70.0	74.5
BMI (Body Mass Index)	30.1	34.3

4.2 Handling Outliers

The spread of any kind of distribution of the data is tends to decrease at the extremes. The middle half of the data which lies within the second and third quartiles are called as interquartile range. This IQR (Interquartile Range) method is used for outlier detection as next step of the process. The outliers have been replaced with the median values again.

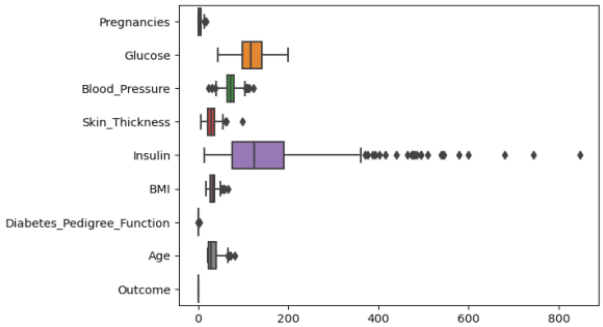

Fig. 4. Before removing the outliers

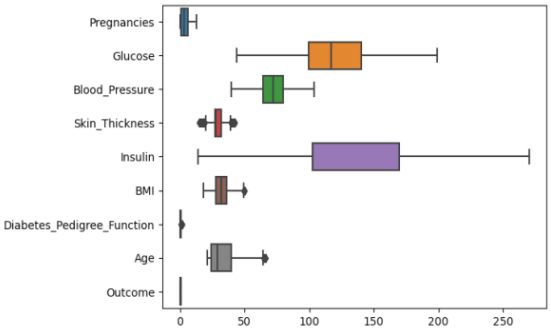

Fig. 5. After removing the outliers

4.3 Handling the Imbalance Data

In order to balance the dataset SMOTE (Synthetic Minority Over-sampling Technique) has been used to prevent the diabetic and non-diabetic biased learning of the model.

4.4 Splitting Normalized Data

The data has been normalized and then divided into training set (80%) and testing set (20%) in the ratio 4:1.

4.5 Data Modeling

The training data has been fed into the following models:-

i. **Naïve Bayes:** Gaussian Naïve Bayes, Multinomial Naïve Bayes and Complement Naïve Bayes models has been used to train the model. For 10 fold cross validation with var_smoothing parameter value to be 0.01 for GaussianNB() obtains accuracy of 87.66%. MutinomialNB() shows an accuracy of 82.46% and ComplementNB() showsn an accuracy of 80.51%. Gaussian Naïve Bayes shows us the higher accuracy with an AUC value of 80.6.

ii. **Logistic Regression:** A 10-fold cross validation has been performed on this model, it was seen that parameters 'C': 0.001, 'max_iter': 100, 'penalty': 'l2' gave us an optimal accuracy of 89.61% with AUC value of 87.8%

iii. **Support Vector Machine:** For a 5-fold cross validation and with most optimal parameters 'C': 1, 'gamma': 0.001, 'kernel': 'rbf', the model gave an accuracy of 88.96% with AUC value to be 87.3%.

iv. **K-Nearest Neighbors:** For the nearest neighbors to be 8, the model gave us an accuracy of 88.96% with AUC value of 87.9%.

v. **Decision Tree:** For the most optimal parameters criterion = 'entropy', max depth=20, max features=6, min samples leaf = 4, the model gave an accuracy of 88.31% with AUC value to be 86.2%.

vi. **Random Forest:** The model gave us an accuracy of 90.90% for parameter values – criterion = 'entropy', max depth = 60, max features = 8, min samples leaf = 2, n_estimators = 40 with AUC value 88.8%

vii. **Gradient Boosting:** For the parameters n_estimators = 55, learning rate = 0.25, max depth = 1, loss = 'log_loss', the model gave an accuracy of 91.55% with AUC value of 89.7%.

viii. **Neural Network:** The 3-Dense layers consists of 30 units,10 units and 1 unit. The first two layers consists of relu as activation function where as the last layer consist of sigmoid as activation function considering loss function to be binary cross entropy and the optimizer be adam version of the stochastic gradient descent. The model gave an accuracy of 84.41% with AUC value to be 82.8%.

ix. **Adaboost:** For n_estimators=30, the model shows us an accuracy of 89.61% with AUC value to be 88.3%.

x. **Logitboost:** For n_estimators=16, the accuracy obtained was 90.25% with AUC value of 90%.

xi. **Xgboost:** For the parameters 'min_child_weight': 9, 'max_depth': 19, 'learning_rate': 0.1105, 'gamma': 0.8, the model accuracy comes out to be 90.90% with AUC value of 89.9%.

xii. **K-Means Clustering:** The model has been trained for 2 clusters for which the accuracy obtained is 87.66% with AUC value of 87.8%.

5 Results

The results of the evaluation of the above methods explained have been demonstrated in Table 2.

Table 2. Evaluation Table

Models	**Accuracy**	**F1-score**		**AUC**
		outcome=0	outcome=1	
Naïve Bayes	87.66	0.91	0.81	0.806
Logistic Regression	89.61	0.93	0.83	0.878
Support Vector Machine	88.96	0.92	0.82	0.873
K-Nearest Neighbours	88.96	0.92	0.82	0.879
Decision Tree	88.31	0.92	0.81	0.862
Random Forest	90.90	0.94	0.85	0.888
Gradient Boosting	91.55	0.95	0.86	0.89.7
Neural Network	84.41	0.89	0.76	0.828
Adaboost	89.61	0.92	0.83	0.883
Logitboost	90.25	0.93	0.85	0.90
Xgboost	90.90	0.93	0.85	0.899
K-Means	87.66	0.91	0.80	0.878

A comparison with the present work has been shown to mark a boundary between our work with the work done till date in Table 3.

Table 3. Evaluation Table

ML Algorithms	Paper[8]	Paper[9]	Paper[10]	Paper[11]	Paper[12]	Paper[13]	Paper[14]	Paper[15]	Paper[16]	Paper[17]	Paper[18]	Paper[19]	Paper[20]	Paper[21]	Our work
Naïve Bayes	--	--	--	--	--	--	79.56	--	--	79.56	--	76.30	68.83 and 80.52	77.83	87.66
Logistic Regression	81.16	77.5	77	--	--	--	--	--	--	--	--	--	81.92	--	89.61
Support Vector Machine	81.16	77.9	--	--	--	89	--	--	--	--	78	65.10	83.12	--	88.96
K-Nearest Neighbours	77.92	76	77	--	83.76	88	--	--	--		--	--	81.82	--	88.96
Decision Tree	72.07	75.8	--	78.17	--	--	76.95	--	78.64	76.95 and 78.64	--	73.82	77.92	75.65	88.31
Random Forest	77.92	79.7	--	--	--	--	--	--	--	--	--	--	75.32	79.57	90.90
Gradient Boosting	--	79.2	--	--	--	--	--	--	66.40	--	--	--	76.62 and 80.52	--	91.55
Neural Network	77.92	78.4	--	--	--	86	--	88.41		--	--	--	84.42	--	84.41
Adaboost	--	--	--	--	--	--	--	--	77.86	--	--	--	76.62	--	89.61
Logitboost	--	--	--	--	--	--	--	--	77.49	--	--	--	--	--	90.25
Xgboost	--	--	--	--	--	--	--	--	--	--	--	--	--	--	90.90
K-Means	--	--	--	--	--	--	--	--	--	--	--	--	--	--	87.66

6 Conclusion

We have used various classification algorithms as they are very much popular in the medical field. It helps in prediction by solving complex time consuming problems. In this paper work we had tried to predict the type-2 diabetes using the human parameters through the implementation of twelve machine learning algorithms. A comparison with the present work has been shown to mark a boundary between our work with the work done till date. It can be seen that the bagging and boosting techniques yield better output compared to pure models. This work can be used to note down the pros and cons of using different models on the PIMA Indian Diabetes Dataset and become handy to be useful in practical purpose of early prediction of diabetes. The PIMA dataset encapsulates attributes of individuals from both the US and Mexico, offers a rich repository of salient features, enhancing its real-world applicability. Integrating additional data from diverse healthcare settings can foster the development of a more robust and generalizable predictive model. Such advancements can significantly support early risk assessments, enabling timely preventive measures for individuals predisposed to Type-2 diabetes.

References

1. https://www.who.int/news-room/fact-sheets/detail/diabetes, last accessed 12/12/22.
2. https://www.nature.com/articles/nrdp201519#Abs1, last accessed 12/12/22.

3. Kowsher, Md, et al. "Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers." 2019 22nd International Conference on Computer and Information Technology (ICCIT). IEEE, 2019.
4. Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.
5. Lai, Hang, et al. "Predictive models for diabetes mellitus using machine learning techniques." BMC endocrine disorders 19.1 (2019): 1-9.
6. Sarwar, Muhammad Azeem, et al. "Prediction of diabetes using machine learning algorithms in healthcare." 2018 24th international conference on automation and computing (ICAC). IEEE, 2018.
7. Mishra, Shubham, A. Vinod, and S. Kala. "Machine Learning Approaches for Type-2 Diabetes Software Predictor." 2022 International Conference on Innovative Trends in Information Technology (ICITIIT). IEEE, 2022.
8. Barhate, Rahul, and Pradnya Kulkarni. "Analysis of classifiers for prediction of type ii diabetes mellitus." 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, 2018.
9. Singh, Varsha, and A. K. Mishra, "A genetic algorithm for k-mean clustering," International Journal of Emerging Technologies in Computational and Applied Sciences, vol. 7, no. 4, pp. 359–364, 2013.
10. Singh, Varsha, Singh, Vijai, and U. S. Tiwary, "Clustering using genetic algorithm: A collaborative performance analysis," in 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2023, pp. 407–412.
11. Patra, Radhanath. "Analysis and prediction of Pima Indian Diabetes Dataset using SDKNN classifier technique." IOP Conference Series: Materials Science and Engineering. Vol. 1070. No. 1. IOP Publishing, 2021.
12. Kaur, Harleen, and Vinita Kumari. "Predictive modelling and analytics for diabetes using a machine learning approach." Applied computing and informatics (2020).
13. Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." arXiv preprint arXiv:1502.03774 (2015).
14. Ashiquzzaman, Akm, et al. "Reduction of overfitting in diabetes prediction using deep learning neural network." IT convergence and security 2017. Springer, Singapore, 2018. 35-43.
15. Sen, S.K. and Dash, S. (2014) Application of Meta Learning Algorithms for the Prediction of Diabetes Disease. International Journal of Advance Research in Computer Science and Management Studies, 2, 396-401
16. Iyer, A., Jeyalatha, S. and Sumbaly, R. (2015) Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5, 1-14. https://doi.org/10.5121/ijdkp.2015.5101
17. Kumari, V.A. and Chitra, R. (2013) Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications (IJERA), 3, 1797-1801.
18. Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.
19. Mahabub, Atik. "A robust voting approach for diabetes prediction using traditional machine learning techniques." SN Applied Sciences 1.12 (2019): 1-12.
20. Chang, Victor, et al. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms." Neural Computing and Applications (2022): 1-17.