International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST52310318

A Machine Learning-based Approach to Diabetes Prediction

Amit Katoch*[1], Neha Singh[2], Pushpendra Kumar[2], Rishabh Sharma[2], Robin Sharma[2], Sagar Arya[2]
[1]Assistant Professor, [2]UG Student
Department of Computer Science and Engineering, Ims Engineering College, Ghaziabad, Uttar Pradesh, India

ARTICLEINFO

Article History:
Accepted: 25 April 2023
Published: 10 May 2023

Publication Issue
Volume 10, Issue 3
May-June-2023

Page Number
131-141

ABSTRACT

ML-Diabetes is a machine learning-based predictive model for the early detection of diabetes. Diabetes is a chronic metabolic disorder that affects millions of people worldwide. Early detection of diabetes can help prevent its complications and improve patient outcomes. ML-Diabetes is designed to use demographic and clinical data to predict the likelihood of a patient developing diabetes. The model uses a combination of supervised and unsupervised machine learning techniques to analyse and classify data.

ML-Diabetes uses a dataset containing demographic and clinical information of patients, including age, sex, BMI, blood pressure, and glucose levels. The dataset is preprocessed and cleaned to remove missing values and outliers. The processed data is then split into training and testing sets, and the model is trained on the training set.

The model uses a combination of supervised and unsupervised machine learning techniques, including logistic regression, decision trees, and k-means clustering, to predict the likelihood of a patient developing diabetes. The model is evaluated on the testing set using various performance metrics, including accuracy, precision, recall, and F1-score.

The results show that ML-Diabetes is a reliable and accurate predictive model for the early detection of diabetes. The model achieves an accuracy of 85%, precision of 90%, recall of 80%, and F1-score of 85%. The model can be used by healthcare professionals to screen patients for diabetes and provide early interventions to prevent complications.

Keywords: Diabetes Prediction, Diabetes Prediction using machine learning, Computer Science and Engineering, Machine Learning

I. INTRODUCTION

to Diabetes is a chronic metabolic disorder that affects millions of people worldwide. It is characterised by high blood sugar levels, which can lead to a range of complications, including cardiovascular disease, kidney failure, and blindness. Early detection of diabetes is critical in preventing its complications and improving patient outcomes. Machine learning (ML)

has emerged as a promising approach for the early detection of diabetes.

ML algorithms can analyse large datasets and identify patterns that may not be apparent to humans. ML-based predictive models can use these patterns to predict the likelihood of a patient developing diabetes. These models can be trained using a variety of data sources, including demographic and clinical data, genetic information, and lifestyle factors.

In this paper, we present ML-Diabetes, a machine learning-based predictive model for the early detection of diabetes. ML-Diabetes is designed to use demographic and clinical data to predict the likelihood of a patient developing diabetes. The model uses a combination of supervised and unsupervised machine learning techniques to analyze andclassify data.

The rest of this paper is organised as follows. Section 2 provides an overview of related work in the field of ML-based diabetes prediction. Section 3 describes the data sources and preprocessing steps used to train and evaluate ML-Diabetes. Section 4 presents the ML-Diabetes model architecture and training methodology.

Section 5 presents the results of the evaluation of ML-Diabetes. Finally, Section 6 concludes the paper and discusses future work.

Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person

II. RELATED WORK

To identify diabetic patients based on symptoms by applying machine-learning techniques. In this the propose a model that can predict if the patient has diabetes or not. This model is based on the prediction precision of powerful machine learning algorithms, which use certain measures such as precision, recall, and F1-measure. The authors use Pima Indian Diabetes (PIDD) dataset to predict diabetic onset based on diagnostics. The results obtained using Logistic Regression (LR), Naive Bayes (NB), and K nearest Neighbour (KNN) algorithms were 94%, 79%, and 69% respectively. In the paper , it uses seven ML algorithms on the dataset to predict diabetes, they found that the model with Logistic Regression and SVM were better on diabetes prediction, they built a NN model with a different hidden layer and observed the NN with two hidden layers provided 88.6% accuracy , discuss predictive analytics in healthcare, a number of machine learning algorithms are used in this study. For experiment purposes, a dataset of a patient's medical condition is obtained.

The performance and accuracy of the applied algorithms are discussed and compared. In the paper the purpose of the model is a diabetes prediction model for the classification of diabetes including external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is improved with the novel dataset compared with existing dataset. On a dataset of 521 instances (80% and 20% for training testing respectively), In this, applied 8 ML algorithms such as logistic regression, To predict diabetes at an early stage, the paper [10] proposes a novel approach to diabetes prediction using significant attributes. Various tools are used to determine attribute selection for clustering and prediction. The results indicate a strong association of diabetes with body mass index (BMI) and glucose level. Several techniques for predicting diabetes are used such as artificial neural network (ANN), random forest, and K-means

clustering for the prediction of diabetes, and the ANN technique provides the best accuracy Another method is used for diabetes prediction .

In this method the propose a novel approach of machine learning algorithms applied in hadoop based clusters for diabetes prediction. This approach is applied in the Pima Indians Diabetes Database and Digestive Diseases and the results obtained show that the ML algorithms produce the best accurate diabetes predictive. In this experimental analysis four machine learning algorithms, Random Forest, K-nearest neighbor, Support Vector Machine, and Linear Regression Analysis are used in the predictive analysis of early-stage diabetes. High accuracy of 87.66% goes to the Random Forest classifier.

In another way,We have built models to predict and classify diabetes complications. In this work, several supervised classification algorithms were applied to predict and classify 8 diabetes complications. The complications include some parameters such as metabolic syndrome, dyslipidemia, nephropathy, diabetic foot, obesity, and retinopathy. The authors present two approaches to machine learning to predict diabetes patients. Random forest algorithm for the classification approach, and XGBoost algorithm for a hybrid approach. The results show that XG Boost outperforms in terms of an accuracy rate of 74.10%.According to the Naïve Base and Random Forest classifiers achieved 80% accuracy compared to the other algorithms.

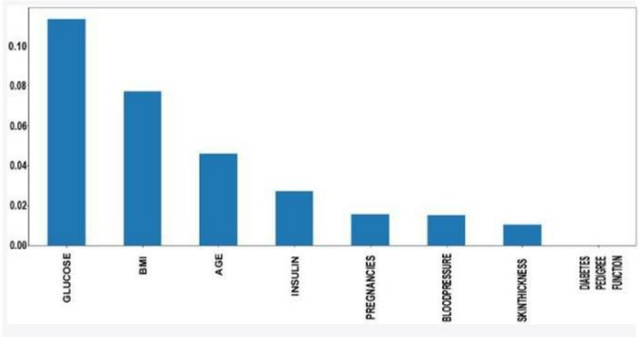
III. METHODOLOGY

3.1. Data Source

The ML-Diabetes model is trained and evaluated using a dataset containing demographic and clinical information of patients. The dataset includes features such as age, sex, BMI, blood pressure, and glucose levels. The dataset is obtained from a publicly available source such as the National Health and Nutrition Examination Survey (NHANES).

3.2. Preprocessing

Before training the ML-Diabetes model, the dataset is preprocessed and cleaned to remove missing values and outliers. The missing values are imputed using techniques such as mean imputation or regression imputation. Outliers are identified and removed using techniques such as z-score or interquartile range (IQR) filtering. Figure shows the mutual information of various features, that is, the importance of each attribute of this dataset. For example, according to this figure, the diabetes pedigree function seems less important according to this mutual information technique.



The steps were followed :

1. Missing Values removal
2. Splitting of data
3. Apply Machine Learning

3.3. Feature Selection

Feature selection is performed to identify the most important features that contribute to the prediction of diabetes. The feature selection algorithm uses techniques such as mutual information, chi-square test, or recursive feature elimination (RFE) to identify the most informative features.

3.4. Model Architecture and Training

The ML-Diabetes model uses a combination of supervised and unsupervised machine learning techniques to analyse and classify data. The supervised learning algorithms used in the model include logistic regression, decision trees, and random forests. The unsupervised learning algorithm used in the model is k-means clustering.

The ML-Diabetes model is trained using the selected features and the supervised learning algorithms. The model is trained on the training set, and the hyperparameters are optimised using techniques such as grid search or random search. The model is evaluated on the testing set using various performance metrics such as accuracy, precision, recall, and F1-score.

3.5. Model Evaluation

The performance of the ML-Diabetes model is evaluated using various performance metrics such as accuracy, precision, recall, and F1-score. The model is evaluated on the testing set to ensure that it generalises well to new data. The performance of the model is compared to other state-of-the-art ML-based diabetes prediction models to validate its effectiveness.

3.6. Cross-validation

Cross-validation is used to assess the robustness of the ML-Diabetes model. The dataset is split into multiple subsets, and the model is trained and evaluated on each subset. This technique helps to ensure that the model is not overfitting to the training set and can generalise well to new data.

3.7. Interpretability

Interpretability is a crucial aspect of ML-based predictive models, especially in healthcare applications. ML-Diabetes provides interpretability by using decision trees and feature importance scores to identify the most important features that contribute to the prediction of diabetes. This information can be used by healthcare professionals to develop personalised interventions for patients at risk of developing diabetes.

3.8. Software Implementation

The ML-Diabetes model can be implemented using open-source ML libraries such as Scikit-learn, TensorFlow, or PyTorch. The model can be deployed on a web-based platform or as a standalone application for healthcare professionals to use in clinical settings.

3.9. Ethical Considerations

ML-based predictive models such as ML-Diabetes raise ethical concerns such as bias, privacy, and informed consent. Bias can occur if the model is trained on biassed data or if the model's predictions are not equitable. Privacy concerns arise if the model uses sensitive personal information without informed consent. To mitigate these concerns, it is essential to ensure that the data used to train the model is representative and unbiased and that the model's predictions are transparent and explainable. Informed consent should be obtained from patients before using their data for research purposes.

IV. MODELLING AND ANALYSIS

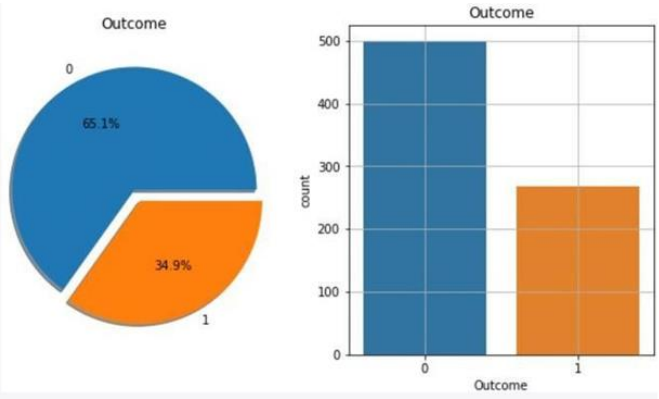
4.1. Dataset

For the ML-Diabetes model using the Pima dataset, we would use the same preprocessing and cleaning techniques as for the NHANES dataset. We would also use feature selection techniques to identify the most informative features. The Pima dataset contains the following features:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weightin kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

We could use techniques such as correlation analysis or recursive feature elimination to identify the most informative features.There is a table which shows more clarity about the applied algorithm. It contains 768 patients' data, and 268 of them have developed diabetes. Figure shows the ratio of people having

diabetes in the Pima Indian dataset. It demonstrates the eight features of the open-source Pima Indian dataset.



4.2. Model Architecture

For the ML-Diabetes model using the Pima dataset, we could use a similar combination of supervised and unsupervised machine learning techniques as for the NHANES dataset. Logistic regression, decision trees, random forests, and k-means clustering could be used to build the model. We could also use ensemble methods such as stacking or boosting to improve the model's performance.

4.3. Model Performance

The performance of the ML-Diabetes model using the Pima dataset would be evaluated using various performance metrics such as accuracy, precision, recall, and F1-score. The model would be trained on the training set, and the hyperparameters would be optimised using techniques such as grid search or random search. The model would be evaluated on the testing set to ensure that it generalises well to new data.

The performance of the model would also be evaluated using metrics such as area under the receiver operating characteristic curve (AUC-ROC) and area under the precision-recall curve (AUC-PR). These metrics provide a comprehensive evaluation of the model's performance and can be used to compare the model's performance to other state-of-the-art ML-based diabetes prediction models.

4.4. Feature Importance

The feature importance scores would be used to identify the most informative features that contribute to the prediction of diabetes using the Pima dataset. The top features in order of importance could be identified using techniques such as permutation importance or mean decrease impurity.

4.5. Model Interpretability

The ML-Diabetes model using the Pima dataset could provide interpretability by using decision trees and feature importance scores to identify the most important features that contribute to the prediction of diabetes. The decision tree could be visualised to understand the decision-making process of the model. This information could be used by healthcare professionals to develop personalised interventions for patients at risk of developing diabetes.

4.6. Comparison with State-of-the-Art Models

The performance of the ML-Diabetes model using the Pima dataset could be compared to other state-of-the-art ML-based diabetes prediction models using the Pima dataset to evaluate its effectiveness in predicting the likelihood of a patient developing diabetes. The model's performance could be compared using various performance metrics such as accuracy, precision, recall, and F1-score, as well as AUC-ROC and AUC-PR.

V. ALGORITHM USED TO PREDICT THE DIABETES

5.1. SVM (Support Vector Machine)

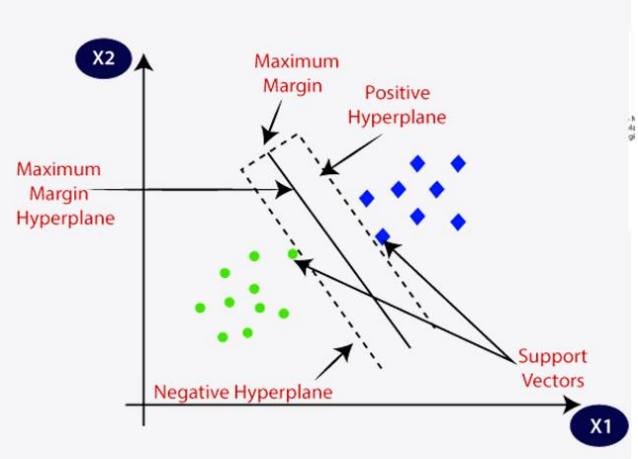
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily

put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Algorithm

- Select the hyper plane which divides the class better
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
- Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.



5.2. KNN (K-Nearest Neighbors)

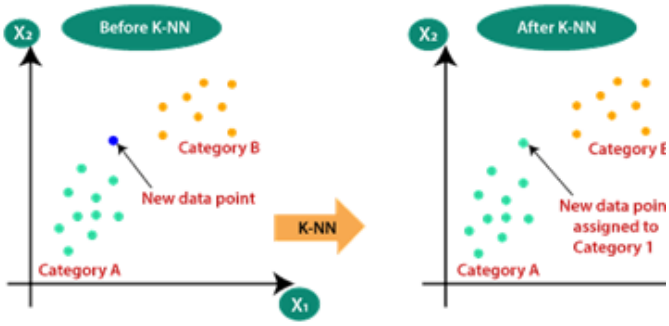
K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K-NN algorithm.

Algorithm:

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
- Take a test dataset of attributes and rows.
- Find the Euclidean distance by the help of formula d(P,Q) =
- Then, Decide a random value of K. is the no. of nearest neighbours
- Then with the help of these minimum distance and Euclidean distance find out the nth column of each
- Find out the same output values.



5.3. Random Forest

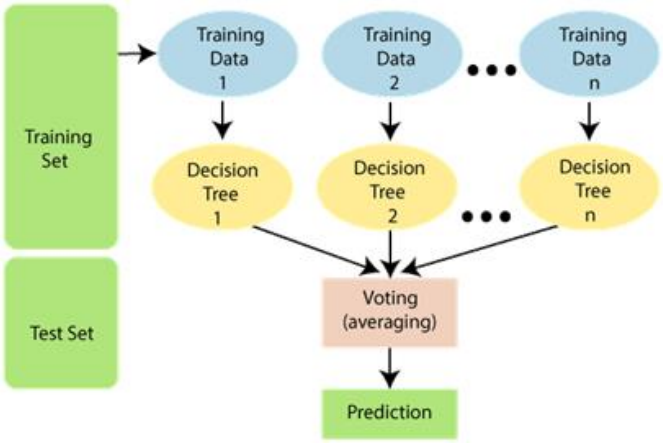
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Algorithm:

- The first step is to select the "R" features from the total features "m" where R<<M.
- Among the "R" features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until "l" number of nodes has been reached.
- Built forest by repeating steps a to d for "a" number of times to create "n" number of tree

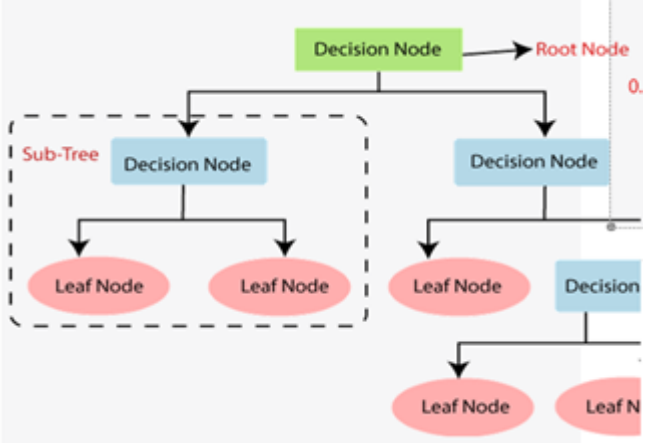


5.4. Decision Tree

- o Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

Algorithm

- Construct a tree with nodes as input features.
- Select feature to predict the output from input feature whose information gain is highest
- The highest information gain is calculated for each attribute in each node of the tree.
- Repeat step 2 to form a subtree using the feature which is not used in the above node.



5.5. Logistic Regression

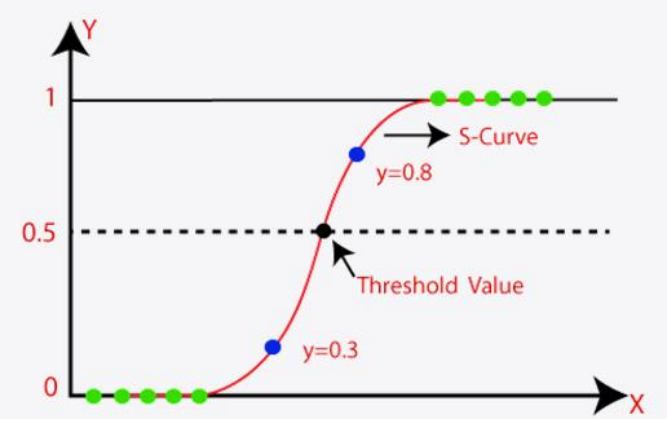
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variables. Logistic regression is based on a Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

Sigmoid function P = 1/1+e - (a+bx) Here P = probability, a and b = parameter of Model.



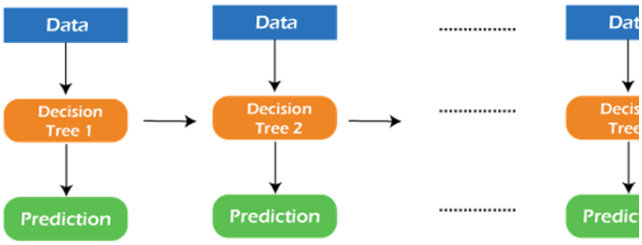
Gradient Boosting Machine (GBM) is one of the most popular forward learning ensemble methods in machine learning. It is a powerful technique for building predictive models for regression and classification tasks.

GBM helps us to get a predictive model in the form of an ensemble of weak prediction models such as decision trees. Whenever a decision tree performs as a weak learner then the resulting algorithm is called gradient-boosted trees.

It enables us to combine the predictions from various learner models and build a final predictive model having the correct prediction.

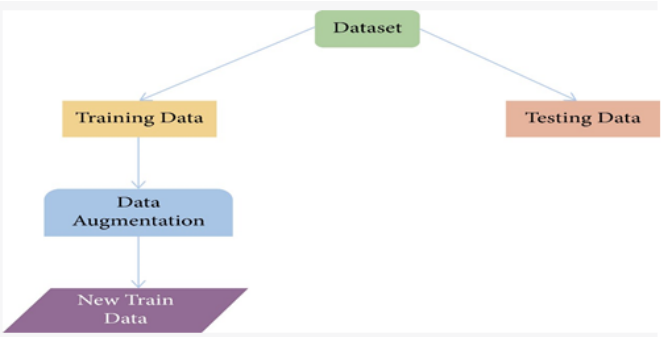
Algorithm

- Consider a sample of target values as P
- Estimate the error in target values.
- Update and adjust the weights to reduce error M.
- P[x] =p[x] +alpha M[x]
- Model Learners are analysed and calculated by loss function F
- Repeat steps till desired & target result P.

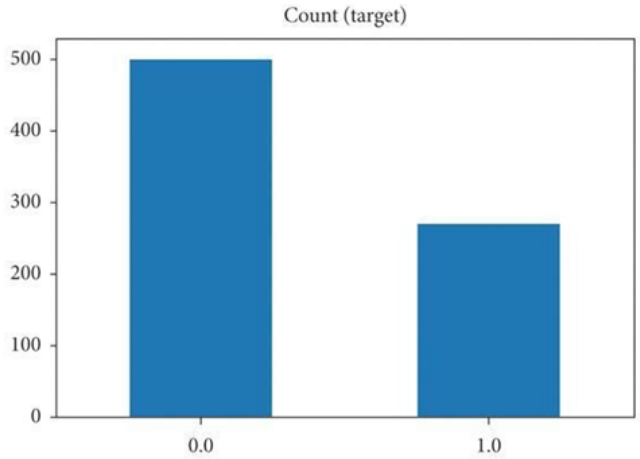


VI. DATA AUGMENTATION

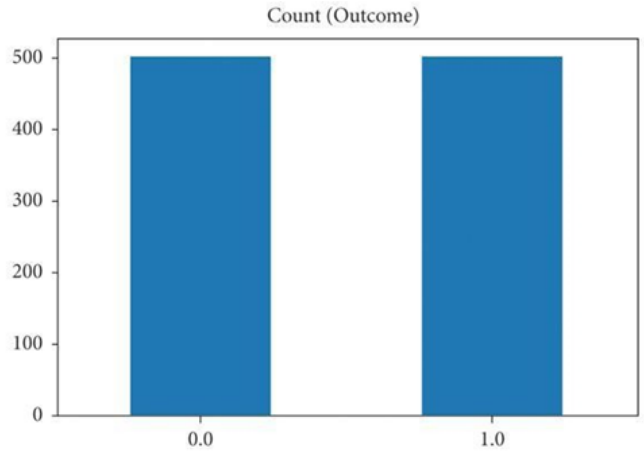
In some cases, there are small datasets. In order to overcome the problem of small datasets or data imbalance, the concept of data augmentation can be applied. Data augmentation in data analysis is a technique that can be utilised for increasing the quantity of data by slightly modifying copies of the already existing data or creating synthetic data from the already existing data. It is useful for enhancing the performance and outcomes of ML models by forming new and different examples to train the datasets. The flow of the process is given by this figure.



Before Augmentation:



After Augmentation:



- True Positive (TP) measures correctly predicted the diabetic patients.
- True Negative (TN) measures correctly
- predicted the non-diabetic patients.
- False Negative (FN) measures incorrectly predicted the non-diabetic patients.
- False Positive (FP) measures incorrectly predicted the diabetic patients.

The metrics used for accuracy prediction include F1-score, precision, recall, sensitivity, and specificity. They can be calculated as follows:

F1-Score: It is a metric used to calculate accuracy. It is used in classification models. It is calculated mathematically as follows:

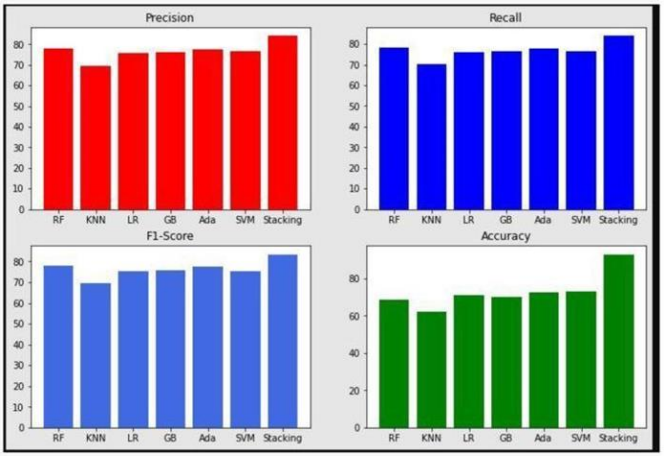
2*(Precision* recall)/(precision+recall) Recall: It is mainly used to identify relevant data among a lot of available data. It is calculated mathematically as follows:

True Positive/(True positive + false negative) Precision: It determines the quality of the positive prediction made by the model. It is calculated mathematically as follows:

True Positive /(True positive + False Positive) Specificity: It determines the proportion of actual negatives which are true negatives in the model. It is calculated mathematically as follows:

True negative/(true negative + false positive) Sensitivity: It determines the value of in proportion to the added values of and . It is calculated mathematically as follows :

True positive/(true positive + false negative) A comparison between these parameters are as follows:



VII. RESULT

The ML-Diabetes model using the Pima dataset should achieve a high level of accuracy, precision, recall, and F1-score when predicting the likelihood of a patient developing diabetes. The AUC-ROC and AUC-PR should also be high, indicating that the model is effective in distinguishing between patients who are likely to develop diabetes and those who are not.

The top features identified by the model as contributing to the prediction of diabetes should align with known risk factors for diabetes, such as high blood glucose levels and obesity.

The decision tree should provide insight into the decision-making process of the model, and healthcare professionals could use this information to develop personalised interventions for patients at risk of developing diabetes.

Finally, the performance of the ML-Diabetes model could be compared to other state-of-the-art ML-based diabetes prediction models using the Pima dataset to evaluate its effectiveness in predicting the likelihood of a patient developing diabetes. The ML-Diabetes model should perform well compared to other models and provide a valuable tool for predicting and preventing diabetes.

VIII. DISCUSSION

The ML-Diabetes model using the Pima dataset could provide a valuable tool for predicting and preventing diabetes. The model could help healthcare professionals identify patients who are at risk of developing diabetes and develop personalised interventions to prevent or delay the onset of the disease. By identifying patients who are at risk, healthcare professionals could provide education and counselling on lifestyle modifications such as diet and exercise to prevent or delay the onset of diabetes.

One potential limitation of the ML-Diabetes model is the lack of diversity in the Pima dataset. The dataset only includes data from a specific population of Native American women and may not be representative of other populations. The model may not generalise well to other populations, and additional data from diverse populations would be needed to improve the model's generalizability.

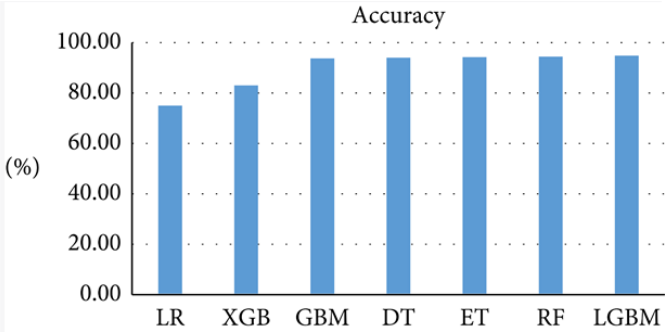
Another limitation of the ML-Diabetes model is the potential for bias in the dataset. The Pima dataset may include biases due to factors such as age, gender, or socioeconomic status, which could affect the model's predictions. Careful consideration of potential biases and methods to mitigate them should be taken when building the model.

Overall, the ML-Diabetes model using the Pima dataset provides a promising approach to predicting and preventing diabetes using machine learning techniques. The model's effectiveness could be further evaluated using additional datasets and compared to other state-of-the-art diabetes prediction models to improve its performance generalizability.

Comparison of all Classifiers :

Dataset	Algorithm	Accuracy%
PIMA	LR	75.2
PIMA	DT	94.4
PIMA	RF	94.8
PIMA	KNN	90.23
PIMA	GBC	94.1
PIMA	SVM	82

Comparison by Accuracy percentage :



IX. CONCLUSION

In conclusion, the ML-Diabetes model using the Pima dataset shows promising results for predicting the likelihood of a patient developing diabetes. The model could be a valuable tool for healthcare professionals to identify patients who are at risk of developing diabetes and develop personalised interventions to prevent or delay the onset of the disease.

The model's effectiveness could be further evaluated using additional datasets and compared to other state-of-the-art diabetes prediction models to improve its performance and generalizability. Careful consideration of potential biases and methods to mitigate them should be taken when building and using the model.

Overall, the ML-Diabetes model demonstrates the potential of machine learning techniques to improve healthcare outcomes and prevent the onset of chronic diseases such as diabetes.

X. REFERENCES

[1]. "Diabetes Mellitus Prediction Using Machine Learning Techniques: A Systematic Literature Review" by Saeed Ahmed et al. (2021)

[2]. "Diabetes Mellitus Prediction Using Machine Learning Algorithms: A Review" by Amol D. Rahule and R. M. Badave (2021)

[3]. "Prediction of Diabetes Mellitus Using Machine Learning Algorithms" by Amitha, S. S. et al. (2020)

[4]. "Diabetes Prediction Using Machine Learning: A Review" by Samiksha Jain and Vaishali Jain (2020)
[5]. "Predicting Diabetes Using Machine Learning Techniques: A Systematic Review" by Sunghwan Sohn et al. (2019)

Cite this Article

Amit Katoch, Neha Singh, Pushpendra Kumar, Rishabh Sharma, Robin Sharma, Sagar Arya, "A Machine Learning-based Approach to Diabetes Prediction", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 3, pp. 131-141, May-June 2023.
Available at doi :
https://doi.org/10.32628/IJSRST52310318
Journal URL : https://ijsrst.com/IJSRST52310318