Indian Journal of Modern Research and Reviews

This Journal is a member of the *'Committee on Publication Ethics'*

Online ISSN: 2584-184X

Research Article

Diabetes Prediction using Machine Learning Techniques

Ashna Merin Philip

Abstract:

Diabetes, characterized by elevated glucose levels, poses significant health risks, including heart problems, kidney issues, hypertension, and potential damage to various organs. Early detection and management are crucial to prevent severe complications. This paper aims to enhance the accuracy of diabetes prediction through the application of diverse machine learning techniques. By leveraging datasets collected from patients, machine learning classification and ensemble methods, such as K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF), will be employed. These models play a vital role in predicting diabetes, with varying accuracies. The project emphasizes the importance of identifying diabetes at an early stage for effective control. Notably, the results reveal that Random Forest outperforms other machine learning techniques, displaying its capability for highly accurate diabetes prediction. The findings underscore the significance of leveraging advanced machine learning methods to enhance predictive accuracy and, consequently, the effectiveness of diabetes management.

Keywords: *Diabetes, Machine, Learning, Prediction, Dataset, Ensemble*

1. Introduction

Diabetes stands as a pervasive global menace, attributed to factors such as obesity and elevated blood glucose levels. The condition disrupts insulin, impacting carbohydrate metabolism and raising blood sugar levels. Insufficient insulin production in the body is a key factor leading to diabetes. The World Health Organization (WHO) reports that approximately 422 million individuals, particularly in low-income countries, suffer from diabetes, with projections estimating a surge to 490 million by 2030. The prevalence is evident in various countries, including Canada, China, and India, where the diabetic population in India alone exceeds 40 million among its 100 million-plus residents. Diabetes stands as a significant contributor to global mortality rates. Recognizing the gravity of the issue, early prediction becomes paramount for effective intervention and life-saving measures. This project focuses on forecasting diabetes by utilizing diverse attributes associated with the disease, employing the Pima Indian Diabetes Dataset. Employing various Machine Learning classification and ensemble techniques, the project aims to harness the power of these methods in predicting diabetes. Machine Learning, a method designed to train computers, becomes instrumental in building efficient models using collected datasets, such as the Pima Indian Diabetes Dataset, for predicting diabetes. The challenge lies in selecting the most effective technique among various Machine Learning methods. To address this, popular classification and ensemble methods are applied to the dataset to enhance the accuracy and reliability of diabetes prediction.

Article History

- **ISSN**: 2584-184X
- **Received:** 03 Feb 2024
- **Accepted**: 25 Feb 2024
- **Published:** 05 Mar 2024
- **MRR:**2(3) Mar**.** 2024: 01-06
- **©2024, All Rights Reserved**
- **Peer Review Process:** Yes
- **Plagiarism Checked**: Yes

Authors Details

Ashna Merin Philip
Assistant Professor, Saintgits College of Applied Sciences, Pathamuttom, Kottayam, Kerala, India

Corresponding Author

Ashna Merin Philip
Assistant Professor, Saintgits College of Applied Sciences, Pathamuttom, Kottayam, Kerala, India

2. Related Works

K. Vijiya Kumar *et al.*, [2] introduced the Random Forest algorithm as a pivotal tool for diabetes prediction, aiming to achieve early detection with heightened accuracy in machine learning. Their proposed model demonstrated superior results in predicting diabetes, displaying its effectiveness, efficiency, and promptness. Nonso Nnamoko *et al.* [5] presented an ensemble supervised learning approach for predicting diabetes onset. Employing five widely used classifiers and a meta-classifier for output aggregation, their method exhibited enhanced accuracy compared to similar studies using the same dataset. Tejas N. Joshi *et al.*, [4] contributed to diabetes prediction using machine-learning techniques, utilizing three different supervised methods, namely SVM, Logistic Regression, and ANN. Their project focused on developing an effective technique for early diabetes detection. Deeraj Shetty *et al.*, [6] proposed a Diabetes Disease Prediction System using data mining assembly, incorporating algorithms like Bayesian and KNN on a diabetes patient's database. This system provided a comprehensive analysis of diabetes, leveraging various attributes for prediction. Muhammad Azeem Sarwar *et al.*, delved into a study on diabetes prediction using multiple machine learning algorithms in healthcare. Their work involved applying six different algorithms, discussing and comparing their performance and accuracy. The research aimed to determine the most suitable algorithm for diabetes prediction. The increasing interest in Diabetes Prediction as a research area emphasizes the need for developing systems that effectively address identified issues from previous studies, recognizing the importance of accurate classification in this crucial domain of computer science.

3. Methodology

The primary objective of this paper is to explore models for predicting diabetes with improved accuracy. The research involves experimentation with various classification and ensemble algorithms aimed at enhancing the precision of diabetes prediction. The subsequent sections will provide a concise overview of the phases involved in the study. Dataset Description - The data utilized in this study is sourced from the UCI repository and is identified as the Pima Indian Diabetes Dataset. This dataset comprises multiple attributes collected from 768 patients.

A. Dataset Description

The data utilized in this study is sourced from the UCI repository and is identified as the Pima Indian Diabetes Dataset. This dataset comprises multiple attributes collected from 768 patients.

Table 1: Dataset Description

S. No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

The 9th attribute in the dataset serves as the class variable for each data point, representing the outcome as either 0 or 1, indicating the absence (negative) or presence (positive) of diabetes, respectively. Regarding the distribution of diabetic patients, the model for predicting diabetes was constructed. However, it is noted that the dataset exhibited a slight imbalance, with approximately 500 instances labeled as 0, denoting the absence of diabetes (negative), and 268 instances labeled as 1, signifying the presence of diabetes (positive).

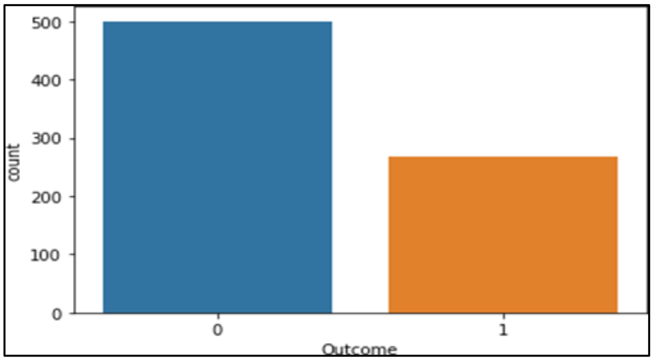

Figure 1: Ratio of Diabetic and Non Diabetic Patient

B. Data Pre-processing

Data pre-processing is a crucial step, especially in healthcare-related datasets, where missing values and other impurities can affect data effectiveness. This process is essential to enhance the quality and effectiveness of the data obtained after the mining process. It plays a vital role in preparing the dataset for the application of Machine Learning Techniques, ensuring accurate results and successful predictions. In the case of the Pima Indian Diabetes Dataset, data pre-processing is carried out in two steps to address any inconsistencies or imperfections.

1) Missing Values Removal

The first step in data pre-processing involves the removal of instances with zero (0) as their values. Instances with zero values are considered irrelevant, as zero is not a plausible measurement in this context. By eliminating these instances, the process of feature subset selection is initiated, resulting in a more streamlined dataset. This subset selection reduces the dimensionality of the data, contributing to faster and more efficient processing.

2) Splitting of Data

Following the cleaning process, the data is normalized and then split into training and testing sets for model evaluation. During this split, the algorithm is trained on the training dataset, while the test dataset is reserved for assessing the model's performance. The training process involves the creation of a model based on the logic, algorithms, and feature values within the training data. Normalization is a key aspect of this phase, aiming to standardize all attributes to the same scale for uniformity and effective model training.

C. Apply Machine Learning

Once the data is prepared, the next step involves the application of Machine Learning Techniques. Various classification and ensemble techniques are employed to predict diabetes using the Pima Indian Diabetes Dataset. The primary goal is to analyze the performance of these techniques, determining their accuracy, and identifying crucial features that significantly contribute to the prediction process. The application of Machine Learning Techniques is pivotal in understanding the effectiveness of different methods and discerning the key features that play a significant role in the accurate prediction of diabetes.

The techniques employed in this study include:

1. Support Vector Machine (SVM)

SVM is a leading supervised machine learning algorithm, widely recognized for classification tasks. It constructs a hyperplane that effectively separates two classes in high-dimensional space, adaptable for both classification and regression. SVM excels in differentiating instances into specific classes, even classifying entities not explicitly supported by the training data. The separation is achieved through a hyperplane, strategically positioned to maximize the margin to the closest training points of any class. SVM stands out for its versatility and efficiency in pattern recognition.

Algorithm involves the following steps:
1. Select the hyperplane that effectively separates the classes.
2. Determine the better hyperplane by calculating the distance between the hyperplane and the data, known as the Margin.
3. If the distance between classes is minimal, the likelihood of misclassification is high, and vice versa. Therefore, prioritize selecting the hyperplane with a higher margin.
4. The margin is calculated as the sum of the distance to the positive point and the distance to the negative point. Mathematically, Margin = distance to positive point + Distance to negative point.

2. K-Nearest Neighbor (KNN)

KNN is a supervised machine learning algorithm suitable for both classification and regression tasks. Operating as a lazy prediction technique, KNN assumes that similar items are proximate to each other. The algorithm groups new data points based on their similarity measures. It records and classifies data points by assessing their similarity. KNN calculates distances between points using a tree-like structure. In making predictions for a new data point, the algorithm identifies the closest data points in the training dataset, known as its nearest neighbors. The parameter 'K' represents the number of nearby neighbors, always a positive integer, and is selected from the set of classes. Closeness is primarily defined in terms of Euclidean distance**.** The Euclidean dis- tance between two points P and Q i.e. P (p1,p2, …. Pn) and Q (q1, q2,..qn) is defined by the following equation:-

$$d(P,Q) = \sum_{i=1}^{n} (P_{i-}Q_i)^2$$

Algorithm for K-Nearest Neighbor (KNN):
1. Utilize a sample dataset named Pima Indian Diabetes dataset, containing columns and rows.
2. Take a test dataset with attributes and rows for evaluation.
3. Calculate the Euclidean distance using the formula.

$$EculideanDistance = \sqrt{\sum_{i=1}^{y} \sum_{j=1}^{m} \sum_{l=1}^{n-1} (R_{(j,l)} - P_{(i,l)})^2}$$

4. Choose a random value for K, representing the number of nearest neighbors.
5. Determine the nth column of each data point using the minimum distance and Euclidean distance.
6. Identify the output values corresponding to the selected neighbors.

If the values are same, then the patient is diabetic, other- wise not.

3. Decision Tree

A fundamental classification method, the decision tree is a supervised learning algorithm specifically applied when the response variable is categorical. The decision tree adopts a tree-like structure that models the classification process based on input features. These input variables can encompass various types, such as graphs, text, discrete values, and continuous variables. The decision tree method is commonly employed for its ability to effectively handle categorical outcomes and provide interpretable results through its structured representation.

Steps for the Decision Tree Algorithm:
1. Construct a tree with nodes representing input features.

2. Select the feature for predicting the output from input features, considering the feature with the highest information gain.
3. Calculate the information gain for each attribute in each node of the tree.
4. Repeat step 2 to form a subtree using the feature that was not utilized in the previous node.

4. Logistic Regression

Logistic Regression is a supervised learning classification algorithm employed to estimate the probability of a binary response based on one or more predictors, which can be continuous or discrete. This algorithm is utilized when classifying or distinguishing data items into categories is required, such as determining whether a patient is positive or negative for diabetes (binary classification). Logistic regression fits a model that describes the relationship between the target and predictor variables. It is based on the linear regression model and utilizes the sigmoid function to predict the probability of positive and negative classes (P = 1/(1 + e^-(a + bx)). Here, P represents probability, and a and b are parameters of the model.

5. Ensembling

Ensembling is a machine learning technique that involves combining multiple learning algorithms for improved performance in a task. It offers superior predictions compared to individual models, addressing issues like noise, bias, and variance. Popular ensemble methods include Bagging, Boosting, Ada-Boosting, Gradient Boosting, Voting, and Averaging. In this work, Bagging (Random Forest) and Gradient Boosting ensemble methods are utilized to predict diabetes, contributing to enhanced accuracy and robustness.

6. Random Forest

Random Forest is an ensemble learning method utilized for both classification and regression tasks. It excels in providing higher accuracy compared to other models and efficiently handles large datasets. Developed by Leo Breiman, Random Forest is a popular ensemble learning method that enhances the performance of decision trees by reducing variance. The algorithm constructs numerous decision trees during training and outputs the mode of classes for classification or the mean prediction for regression, offering robust and reliable results.

Random Forest Algorithm:
1. Select "R" features from the total features "M," where R << M.
2. Identify the best split point for the node among the selected "R" features.
3. Split the node into sub-nodes using the determined best split.
4. Repeat steps a to c until the desired number of nodes "l" has been reached.
5. Build the forest by repeating steps a to d for "a" number of times, creating "n" trees in the ensemble.

The Random Forest algorithm determines the best split using the Gini Index Cost Function, expressed as:

$$Gini = \sum_{k=1}^{n} p_k * (1 - p_k) \; Where \; k = Each \; class \; and \; p = proption \; of \; training \; instances$$

In the Random Forest algorithm, the initial step involves considering the choices and utilizing the rules of each randomly created decision tree to predict outcomes. These predicted outcomes are stored within the target place. Subsequently, the algorithm calculates the votes for each predicted target. Finally, the algorithm relies on the highest voted predicted target as the ultimate prediction from the Random Forest formula. Random Forest is known for providing correct predictions across various applications due to its ensemble nature and the aggregation of predictions from multiple decision trees, leading to robust and accurate outcomes.

7. Gradient Boosting

Gradient Boosting is a powerful ensemble technique primarily employed for prediction, particularly in classification tasks. It combines weak learners to create strong learner models for enhanced prediction accuracy. Gradient Boosting utilizes Decision Tree models, excelling in classifying complex datasets. Known for its effectiveness and popularity, the performance of a gradient boosting model improves over iterations, making it a robust choice for achieving high predictive accuracy.

Gradient Boosting Algorithm:
1. Consider a sample of target values as (P).
2. Estimate the error in target values.
3. Update and adjust the weights to reduce the error (M).
4. Update the target values using the formula (P[x] = p[x] + alpha M[x])
5. Analyse and calculate model learners using the loss function (F).

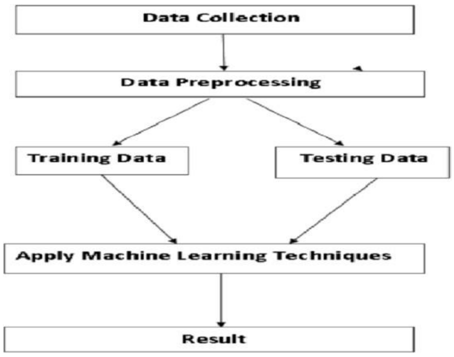

Figure 2: Overview of the Process

6. Repeat steps 2 to 5 until the desired and target result (P) is achieved.

Model Building

The model-building phase is crucial for predicting diabetes, involving the implementation of various machine-learning algorithms discussed earlier.

Procedure of Proposed Methodology

Step 1: Import required libraries, Import diabetes dataset.
Step 2: Pre-process data to remove missing data.
Step 3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.
Step 4: Select the machine learning algorithm i.e. K- Nearest Neighbor**,** Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.
Step 5: Build the classifier model for the mentioned ma- chine-learning algorithm based on training set.
Step 6: Test the Classifier model for the mentioned ma- chine-learning algorithm based on test set.
Step 7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
Step 8: After analysing based on various measures conclude the best performing algorithm.

4. Evaluation and Result Analysis

In this system, various steps were undertaken to implement the proposed approach using Python. The methodology involved the utilization of different classification and ensemble methods, which are standard Machine Learning techniques designed to achieve optimal accuracy from the dataset. Notably, the Random Forest classifier outperformed other methods in this study. The overall strategy was to leverage the best Machine Learning techniques to predict diabetes and achieve high-performance accuracy. The results of these Machine Learning methods are depicted in the figure, highlighting their effectiveness in the prediction task.

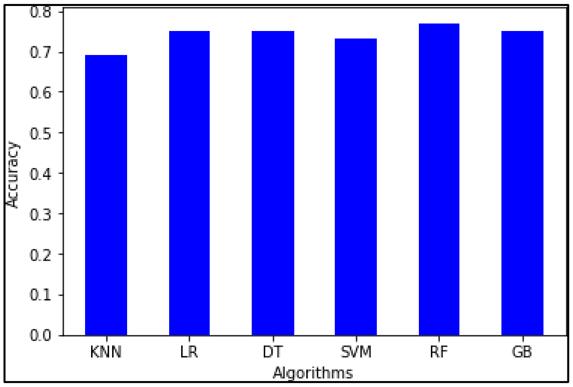

Figure 3: Accuracy Result of Machine learning methods

The importance of features in the prediction, particularly emphasized in the Random Forest algorithm, is presented in the following plot. The chart illustrates the cumulative importance of each feature, displaying their respective roles in predicting diabetes. The X-axis represents the importance of each feature, while the Y-axis displays the names of the features. This visualization aids in understanding the significance of individual features and their collective impact on the accuracy of the diabetes prediction model.

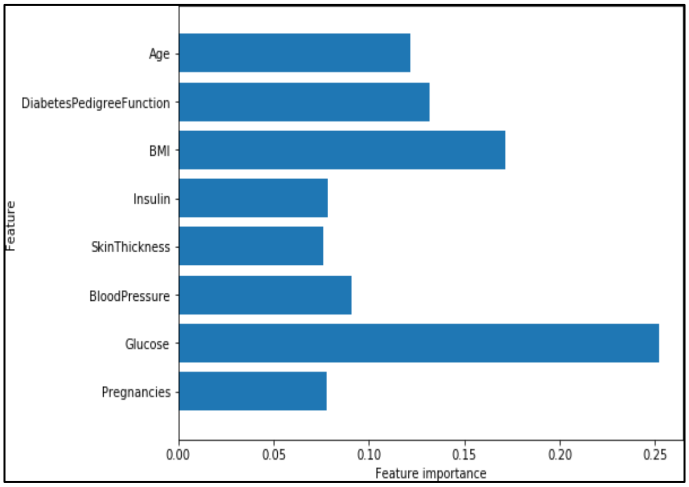

Figure 4: Feature Importance Plot for Random Forest

5. Conclusion

The primary objective of this project, which was to design and implement Diabetes Prediction using various Machine Learning methods and analyze their performance, has been successfully achieved. The proposed approach incorporates a range of classification and ensemble learning methods, including SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting classifiers. The project attains a classification accuracy of 77%. The experimental results provide valuable insights for healthcare, enabling early prediction and informed decision-making to address diabetes and potentially save lives.

References

1. Tasin I, Islam S. Diabetes prediction using machine learning and explainable AI techniques. NLM. 2022.
2. VijiyaKumar K, Lavanya B, Nirmala I, Caroline SS. Random Forest Algorithm for the Prediction of Diabetes. Proceedings of International Conference on Systems Computation Automation and Networking. 2019.
3. Faruque MF, Asaduzzaman, Sarker IH. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. International Conference on Electrical, Computer and Communication Engineering (ECCE). 7-9 February 2019.

4. Joshi TN, Chawan PM. Diabetes Prediction Using Machine Learning Techniques. Int J Eng Res Appl. 2018;8(1):09-13.
5. Nnamoko N, Hussain A, England D. Predicting Diabetes Onset: an Ensemble Supervised Learning Approach. IEEE Congress on Evolutionary Computation (CEC). 2018.
6. Shetty D, Rit K, Shaikh S, Patil N. Diabetes Disease Prediction Using Data Mining. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). 2017.
7. Nahla B, Andrew et al. Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine, IEEE Transactions. 2010;14:1114-20.
8. Dewangan AK, Agrawal P. Classification of Diabetes Mellitus Using Machine Learning Techniques. Int J Eng Appl Sci. 2015;2.
9. Dutta D, Paul D, Ghosh P. Analyzing Feature Importance's for Diabetes Prediction using Machine Learning. IEEE. 2018;pp 942-928.

.