EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Modelling of Diabetic Cases for Effective Prevalence Classification

Shrey Shah[1], Monika Mangla[1*], Nonita Sharma[2], Tanupriya Choudhury[3] and Maganti Syamala[4]

[1]Dwarkadas J Sanghvi College of Engineering, Mumbai, India
[2]Indira Gandhi Delhi Technical University for Women, New Delhi, India
[3]Department of Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, 248002, Uttarakhand, India
[4]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh - 522302, India

Abstract

INTRODUCTION: This study compares and contrasts various machine learning algorithms for predicting diabetes. The study of current research work is to analyse the effectiveness of various machine learning algorithms for diabetes prediction.
OBJECTIVES: To compare the efficacy of various machine learning algorithms for diabetic prediction.
METHODS: For the same, a diabetic dataset was subjected to the application of various well-known machine learning algorithms. Unbalanced data was handled by pre-processing the dataset. The models were subsequently trained and assessed using different performance metrics namely F1-score, accuracy, sensitivity, and specificity.
RESULTS: The experimental results show that the Decision Tree and ensemble model outperforms all other comparative models in terms of accuracy and other evaluation metrics.
CONCLUSION: This study can help healthcare practitioners and researchers to choose the best machine learning model for diabetes prediction based on their specific needs and available data.

Keywords: Machine Learning, Ensemble Model, Diabetes Prediction, Healthcare, Accuracy, sensitivity, Specificity

Received on 15 December 2023, accepted on 16 March 2022, published on 22 March 2024

doi: 10.4108/eetpht.10.5514

*Corresponding author. Email: manglamona@gmail.com

1. Introduction

As per a report published by International Diabetes Federation (IDF), around 537 million adults in age range of 20 years to 79 years have been living with diabetes in year 2021. These statistics of diabetic patients may reach around 643 million and 783 million by 2030 and 2045 respectively [1] These statistics are alarming especially in middle-income and low-income countries and hence necessitates preventive and corrective measures as diabetes may lead to number of problems namely renal failure, blindness, and heart disease etc. In order to curb the spread of disease, it is imperative to have efficient methods to perform early detection of the disease.

The preventive models that can assist in identification of diabetes are the unparalleled need of the entire world. The evolution in the technology have provided a helping hand in this direction and have been assisting the health experts. Among various evolving technologies, the research in the field of artificial intelligence and machine learning (ML) has completely revolutionized the facet of diabetes detection. As a result of motivating results by implementing ML in healthcare, a huge number of researchers have been working in this direction and resultantly number of efficient ML models have been suggested by numerous researchers.

Authors in this paper also employ the different ML models to perform diagnosis of diabetes. For the same, authors initially employ statistical approaches to find patterns and associations that can be used to generate predictions. This statistical analysis is followed by implementation of various ML models. Authors in this paper have employed different ML models namely K nearest neighbour, Decision Tree, Random Forest, Logistic Regression, Gausian Naıve Bayes, Support Vector

Machine (SVM), XGBoost. Authors also employ an ensemble model that uses random forest. Finally, the performance of different ML models is compares in terms of different performance metrics so as to determine the optimal ML model for diabetes detection.

The paper is organized in various sections. Here, section I establishes the need of research and educate the reader regarding its need. Related work by different researchers have been discussed in section 2. Proposed methodology is given in section 3 and results are discussed in section 4. Finally, the conclusion and future work is presented in section 5.

2. Related Work

Although several researchers have undertaken the problem of diabetes prediction and suggested implementation of various technologies for the same. The authors in this section aim to discuss the significant findings and results by different researchers in this domain. Although there are numerous technologies that have been employed for diabetes prediction, the employment of ML has outperformed several other technologies and resultantly disease prediction is kind of dominated by the evolution in the field of artificial intelligence and ML. Therefore, authors in this paper have primarily focused on the application of ML in disease detection.

The 2023 study by Mohanty, Ghosh, Rahat [2] and Reddy, "Advanced Deep Learning Models for Corn Leaf Disease Classification", focuses on the application of deep learning in classifying diseases in corn leaves based on a field study in order to perform the experimental evaluation, authors have considered the Pima Indians Diabetes dataset [3] that contains a variety of demographic and medical characteristics of female patients.

Input: Dataset

1. *Import Dataset*
2. *Preprocessing the Data*
   a. *Handling missing values*
   b. *Removing outliers*
   c. *Normalization*
3. *Principal Feature Extraction*
4. *Split the dataset into Training and validation set.*
5. *Performance Evaluation*

Authors have performed the preprocessing in order to standardize the data which is followed by implementing different ML algorithms namely Decision Tree, Naive Bayes, Gradient Boosting, SVM, Logistic Regression and K nearest neighbor. Performance of these is compared on the basis of different performance metrics namely accuracy, sensitivity, F1-score, and specificity to assess the effectiveness of these algorithms. Authors in [4] also worked on the same dataset and during preprocessing data is normalized and irrelevant features are removed. During the experimental evaluation, it is noticed that Random Forest outperforms other comparative models in terms of accuracy by achieving an accuracy of 77.92%. Additionally, authors also performed Recursive Feature Elimination (RFE) to determine the most important features which are known to be Glucose, BMI, Age, and Insulin. Overall, the work demonstrates the effectiveness ML algorithms for diabetes predictions and significance of feature selection to improve accuracy of predictions.

Similar kind of research is also carried out by authors in [5] who used WEKA software for data preprocessing. However, WEKA does not suggest any ML algorithms it is to be decided by the end user as different data mining algorithms have different requirements. Authors used different ML methods like Artificial neural network (ANN), Bayes Classifier, Decision Tree, Regression and SVM. During the comparative evaluation it is evident that C4.5 based Decision Tree yields highest accuracy of 79%. Carrying further the research, authors in [6] achieved an accuracy of 82.3% using Naïve Bayes Classifier. This significant rise in the accuracy is attributed to the efficient feature selection techniques.

3. Proposed Methodology

Authors in this paper propose to use different ML algorithms [6]. Ahead of applying different ML models, data is preprocessed so as to remove any anomalies and outliers in the dataset that could influence the results [7]. After preprocessing, different ML models are implemented and comparative analysis is performed to determine the optimal ML algorithm for disease detection [8]. The stepwise description of the proposed methodology is illustrated in Fig. 1.

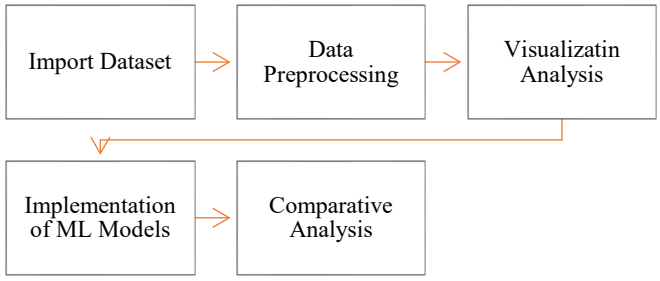


Figure 1. Illustration of Proposed methodology

Authors have used following ML algorithms in the study:

K Nearest-neighbor: KNN is an instancebased learning approach that can be used for classification of the data based on its k nearest neigbors[9]. The distance is generally Euclidean distance and is measured as follows:

$$d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$

Logistic Regression: It is a supervised ML algorithm that determines the probability of a binary response based on one or more predictors. Here, attributes can be both continuous and discrete and it can be used to categorize the

data points into two groups [10]. Here these two groups refer to diabetic and non-diabetic patients. Logistic regression is based on linear regression and the prime objective of the logistic regression is to determine an optimal among the predictor and target variables. Logistic regression uses sigmoid activation function to forecast the likelihood of the positive and negative classes.

Naive Bayes: It is a Bayes' theorem-based algorithm primarily used for classification which describes the probability of a hypothesis being true given certain evidence [11]. The algorithm calculates the probability of an instance belonging to a certain class based on the features of the instance. It assumes that the features are independent of each other and hence is considered "naive".

Support Vector Machine: SVM, a powerful ML algorithm can be used for regression and classification. SVM is efficient for complex and nonlinear datasets where traditional linear models may not yield optimal performance.

Decision Tree: Decision tree is a supervised learning method which is used when target variable is categorical. Random Forest:

Random forest creates multiple decision trees and combines their predictions for classification, regression, and other tasks. It randomly selects the input data and features to create decision tree so as improve the accuracy of the model.

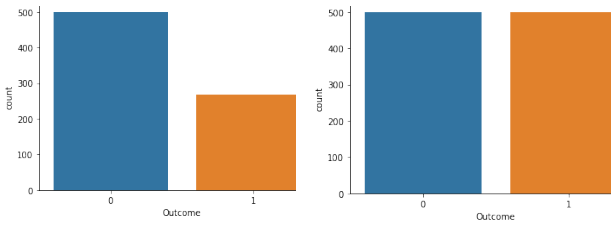
Ensemble Learning: Ensemble learning integrates the results of numerous independent base ML models so as to enhance the efficiency of entire model. The basic idea behind ensemble learning is that it aggregates the predictions by various models so as a capture a robust representation of the underlying data and thus provides better predictions than individual models [12].

XGBoost: Extreme Gradient Boosting also employs ensemble learning where multiple weak learners are integrated to generate a stronger model. It can be used for classification, regression, and ranking problems. Decision trees are used as weak models by the XGBoost classifier, which iteratively enhances their performance by minimising a loss function and adding new trees to remedy the mistakes produced by older ones. This procedure continues until a user-specified stopping criterion is satisfied, such as a minimum increase in performance or a maximum number of trees [12].

4. Results and Discussion

The authors have taken the diabetes dataset from Kaggle [3]. This dataset contains 768 entries among 268 are diabetic and 500 are non-diabetic as shown in Fig. 1. The features/attributes representing are:

- Pregnancies
- BloodPressure
- SkinThickness
- Insulin and Glucose
- BMI
- DiabetesPedigreeFunction
- Age
- Diabetic/Non-diabetic (Target variable)



Ahead of preprocessing *Post oversampling of minority class*

Figure 2. Illustration of Dataset

In Fig. 1, diabetic and non-diabetic patients are represented by 1s and 0s respectively. As evident from Fig. 1, the non-diabetic patients are nearly twice to the diabetic patients leading to imbalance in the dataset. Imbalanced dataset may lead to prediction bias and hence data must be balanced by performing data augmentation. Authors perform oversampling of the minority class (diabetic in current scenario) to eliminate the large difference of volume between diabetic and non-diabetic patients. For the same, few rows of the minority class are randomly selected and duplicated so as to increase the total number of the instances for minority class. The dataset ahead of oversampling and after oversampling is illustrated in Fig. 2(a) and 2(b) respectively.

Further, the various ML models are employed on the dataset and comparative analysis is carried out using various performance metrics namely accuracy, sensitivity, and specificity. As the dataset was initially imbalanced, f1-score is also used for comparative analysis. The mathematical formulations for various performance metrics is given below.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1\ Score = \frac{2\ X\ Recall\ X\ Precision}{Recall + Precision} = \frac{2\ TP}{2TP + FP + FN}$$

Here, T/F and P/N implies True or false and Positive or Negative. Thus, FP refers to False positive in the confusion matrix. The comparative analysis of various ML models is given in Table 1.

Table 1. Comparison of Results

Model	Accuracy	F1-Score	Sensitivity	Specificity
KNN (K=3)	0.785	0.804	0.88	0.69
DT	**0.875**	**0.88262**	**0.94**	**0.81**
RF (d = 10)	0.860	0.87156	0.95	0.77
LR	0.695	0.690	0.68	0.71
NB	0.695	0.687	0.67	0.72
SVM	0.77	0.785	0.84	0.7
XGBoost (d = 6)	0.82	0.829	0.87	0.77
Ensemble (RF + LR)	0.720	0.7021	0.66	0.78

In Table 1, the performance metrics for different ML algorithms have been given and the optimal value for each performance metric is highlighted. Here, authors also implemented an ensemble model that integrates random forest and logistic regression, and it is observed that the proposed ensemble model outperforms all traditional ML models. The graphical comparison of all models for accuracy and f1-score is illustrated in Fig. 3 and Fig. 4 respectively.

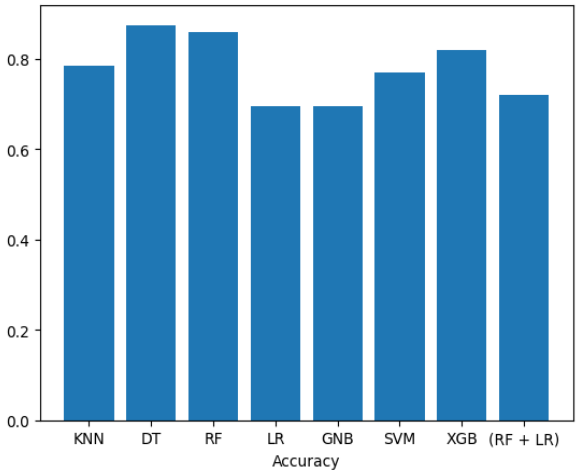


Figure 3. Comparative Analysis of Accuracy

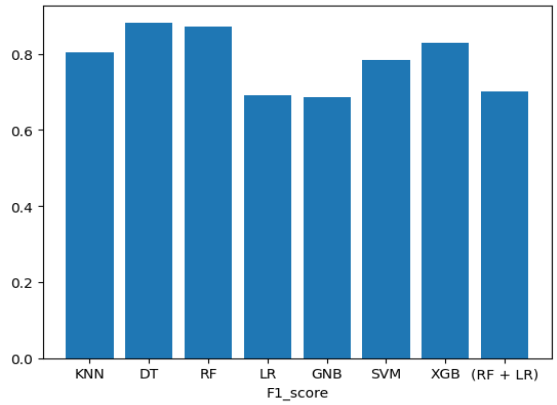


Figure 4. Comparative Analysis of F1_Score

5. Conclusion

This paper discusses the use of various classification algorithms to predict Diabetes using a set of attributes. These models include K- Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, XGBoost, and an ensemble of random forest and decision tree. After reviewing the performance of all models, we infer that an ensemble model of Decision Tree and Random Forest would provide us the maximum accuracy of 87.5%. The early prediction of diabetes can help in its treatment and possibly get it under control quickly.

References

[1] https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
[2] Mohanty, S.N.; Ghosh, H.; Rahat, I.S.; Reddy, C.V.R. Advanced Deep Learning Models for Corn Leaf Disease Classification: A Field Study in Bangladesh. Eng. Proc. 2023, 59, 69. https://doi.org/10.3390/engproc2023059069
[3] https://www.kaggle.com/datasets/mathchi/diabetes-data-set
[4] Kishan Patel, Manu Nair and Shubham Phansekar. Diabetes Prediction using Machine Learning International Journal of Scientific & Engineering Research Volume 12, Issue 3, March-2021
[5] P. Tomar, N. Sharma and S. D. Kamble, "Aspect Analysis Based on Statistical Description and Visualization of Data," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 38-42, doi: 10.1109/IC3I56241.2022.10073201.
[6] Llaha, O. and Rista, A., 2021, May. Prediction and Detection of Diabetes using Machine Learning. In RTA-CSIT (pp. 94-102).
[7] Khadidos A, Khadidos AO, Kannan S, Natarajan Y, Mohanty SN and Tsaramirsis G (2020) Analysis of COVID-19 Infections on a CT Image Using DeepSense Model. Front. Public Health 8:599550. doi: 10.3389/fpubh.2020.599550.
[8] N. Sharma, M. Mangla, M. Ishaque and S. N. Mohanty, "Inferential Statistics and Visualization Techniques for Aspect Analysis," 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah,

Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICAISC56366.2023.10085093.

[9] D. Patel, S. Chopra, N. Sharma and M. Mangla, "Analysis and Visualization of Heart Failure Prediction Dataset," 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 2022, pp. 1-5, doi: 10.1109/AIST55798.2022.10064888.

[10] S. Gupta, R. Sharma, N. Sharma and M. Mangla, "Aspect Analysis of Dementia Patients," 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2022, pp. 1-5, doi: 10.1109/ICICT55121.2022.10064519.

[11] N. Sharma, J. Dev, M. Mangla, V. Wadhwa,S.N. Mohanty, S. N., & Kakkar, D. (2021). A heterogeneous ensemble forecasting model for disease prediction. New Generation Computing, 1-15.

[12] M. Mangla, S. K. Shinde, V. Mehta, N. Sharma, & S. N. Mohanty, (Eds.). (2022). Handbook of Research on Machine Learning: Foundations and Applications. CRC Press.