Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 273–284
DOI : 10.47974/JSMS-1253

Diabetes prediction using feature engineering and machine learning algorithms with security

Jyoti Arora *
Department of Information and Technology
Maharaja Surajmal Institute of Technology
Janakpuri
Delhi
India

Sonia Rathee †
Department of Computer Science and Engineering
Maharaja Surajmal Institute of Technology
Janakpuri
Delhi
India

Mamta Gahlan
Department of Information and Technology
Maharaja Surajmal Institute of Technology
Janakpuri
Delhi
India

Amita Yadav
Shalu
Maharaja Surajmal Institute of Technology
Janakpuri
Delhi
India

* *E-mail:* joy.arora@gmail.com (Corresponding Author)
† *E-mail:* sonia.rathee@msit.in

Abstract

The prevalence of diabetes has been steadily increasing, necessitating accurate prediction models to assist in early diagnosis and proactive management. In this paper, a hybrid machine learning-based diabetes prediction model has been proposed. To evaluate the model, the dataset was subsequently divided into training and testing subsets. We used the Random Forest Classifier, Light Gradient Boosting Mechanism Classifier, Gradient Boosting Classifier, Logistic Regression, K-Nearest Neighbours (KNN) Classifier, Naive Bayes Gaussian, Decision Tree Classifier, XGBoost Classifier, and Support Vector Classifier as nine different classifiers. Several metrics were used to evaluate the models, including testing accuracy, recall score, F1 score, and precision score. We have evaluated our model on the "Pima Indian Diabetes Database"[1], which served as the main dataset, for diabetes prediction. The proposed model serves as a practical framework for researchers and practitioners interested in leveraging machine learning techniques for diabetes prediction.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Diabetes prediction, Light gradient boosting machine (LGBM), Naive Bayes (NB), Random forest, Logistic regression (LR), Gradient boosting (GB), K-nearest neighbour (k-NN), XGBoost, Decision tree (DT), Support vector machines (SVM).*

1. Introduction

Diabetes mellitus is a chronic metabolic disorder with elevated blood glucose levels that affects millions of people worldwide. Diabetes is estimated to affect 422 million people worldwide, primarily in low and middle-income countries, according to the World Health Organization (WHO). Furthermore, by the year 2030, this figure could reach 490 billion. However, a few countries, including Canada, China, and India, have higher diabetes prevalence rates. India now has over 100 million people, and 40 million people among these are diabetics [2]. Statistical methods (such as median and mean) that deal with missing value data typically generate an excessive number of duplicate samples and certainly reduce the relationships between characteristics in the sample data. When classification predictions are made with uneven data, the decision boundary is skewed in favor of the dominant class, which confuses the results. Most of the physicians are still unable to properly forecast results or are inefficient due to a lack of understanding of data analysis methodologies. This leads many people to go untreated, resulting in life-threatening diabetic complications [3]. Early and accurate prediction of diabetes is crucial for timely intervention and effective management of the disease [4]. With the help of data-driven models, machine learning algorithms have demonstrated great promise in the prediction of diabetes

by revealing hidden patterns and connections among a variety of patient features. One such technique, known as FOCUS (Feature Optimization with One-hot Encoding and Scaling) [5], has emerged as a novel approach specifically designed for diabetes prediction.

Further, paper is structured as follows: In Section 2, presents a comprehensive review of relevant literature on diabetes prediction, feature engineering techniques, and their impact on model performance. Section 3 outlines the proposed methodology, including the dataset used, pre-processing steps, and the implementation details of FOCUS. In Section 4, we present the experimental analysis and performance evaluation of FOCUS in comparison to other feature engineering approaches and discusses the findings, highlights the strengths and limitations of FOCUS, and provides insights into the practical implications of our research. At last, in Section 5, concludes by summarizing the key contributions.

2. Literature Survey

Various classifiers are employed in machine learning by R. Moungmai [7] to predict and detect diabetes. like logistic regression, extreme gradient boosting, and naive Bayes statistical modelling. Testing is done on Pima Indian diabetes databases. The Extreme Gradient Boosting algorithm is more accurate than the other two algorithms, with an accuracy of 81%.

A. Chaudhary et al. [8], has compared different machine learning algorithms and different classifier used for predicting the Pima Indian diabetes dataset to predict diabetes illness with the lowest cost and best performance. The decision tree (DT), naive based (NB), and support vector machine learning (SVM) algorithms were utilised by Sisodia et al. in [9] to predict diabetes. The Naive Bayes classifier's accuracy was found to be 76.30%. For classification, Zou et al. [10] used Random Forest, J48, ANN after reducing the number of features using unsupervised techniques such as Principle Component Analysis (PCA) and Minimal Redundancy Maximum Relevance (mRMR) approaches. It is discovered that mRMR accuracy is superior to PCA with all characteristics. Nongyao et al. investigated decision trees,

ANN, logistic regression, and naive Bayes as four classification methods. Additional bagging, boosting, and random forest analysis were all applied to each. Maximum accuracy of 84% to 86% was achieved by all groups. When Wu et al. in [11] used data mining techniques (such as enhanced kNN and logistic regression), they were able to predict a person's likelihood of getting type 2 diabetes up to 95.42% accuracy.

Somnath et al. used the "Probabilistic Neural Network (PNN) to forecast diabetic illness" [12]. The "PIMA Indian dataset" was used to implement the approach. The author decides not to use the pre-processing method. However, "the dataset is divided into 90% for the training dataset and 10% for the typical setting. When tested, the suggested approach had an accuracy of 81.49%, and when taxing the data, it had an accuracy of 89.56%. The LGBM =Algorithm idea has been examined [13]. They used two techniques: gradient-based one-sided sampling and theoretical analysis of the same.

3. Methodology

The proposed model uses machine learning algorithms to forecast diabetes in a systematic manner to predict diabetes using machine learning algorithms and evaluate the effectiveness of FOCUS (Feature Optimization with One-hot Encoding and Scaling) as a feature engineering technique. The Pima Indians Diabetes Database, obtained from Kaggle, serves as the primary data source for this study. The methodology can be summarized into steps.

a) *Data Collection*

The dataset, which is available on Kaggle, was compiled from the Pima Indians Diabetes Database [1]. This dataset was originally stored by the National Institute of Diabetes and Digestive and Kidney Diseases. Based on precise diagnostic parameters that are present in the data, the

Table I
Details of the Dataset

S. No.	Name of Attributes	Description
1	Pregnancies	Number of pregnancies
2	Glucose	Plasma Glucose Amount
3	Blood Pressure	Diastolic blood pressure
4	Skin Thickness	The thickness of Triceps skin fold
5	Insulin	2-h serum insulin value
6	BMI	Body Mass Index value
7	Diabetes Pedigree Function	Diabetes Pedigree function value
8	Outcome	Class Variable (0 or 1)
9	Age	Age of patient

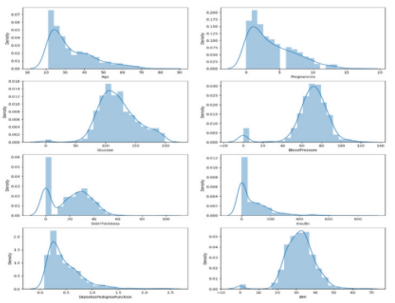

Figure 1
Skew of Data.

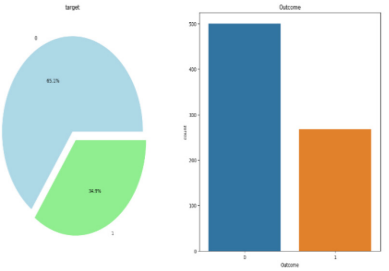

Figure 2
Bar Plot for Outcome Class

dataset seeks to predict whether a patient has diabetes. Many patients at this clinic are Pima Indian women over the age of 21. Table I. gives the details of the dataset.

b) *Data Analysis and Visualization*

This stage involves analyzing and visualizing data. In this stage, the dataset was initially imported from a CSV file format, and an overview of the dataset was obtained by retrieving the first five data units along with their corresponding sizes.

By conducting this data analysis and visualization, a comprehensive understanding of the dataset's characteristics and distributions was obtained as shown the Figure 1.

We further explored the dataset by visualizing the distribution of the outcome class using a bar plot as represented in Figure 2 below.

c) *Data Preprocessing*

Extensive data pretreatment methods were used, utilizing the FOCUS (Feature Optimization with One-hot Encoding and Scaling) [5] framework, to guarantee the best performance of the machine learning models. The preprocessing phase involved several crucial steps, including missing observation analysis, outlier observation analysis, feature engineering, one-hot encoding, and scaling.

Missing Observation Analysis: Initially, the dataset was examined for null values, and although none were found, it was observed that certain variables, such as glucose levels and blood pressure, had missing values

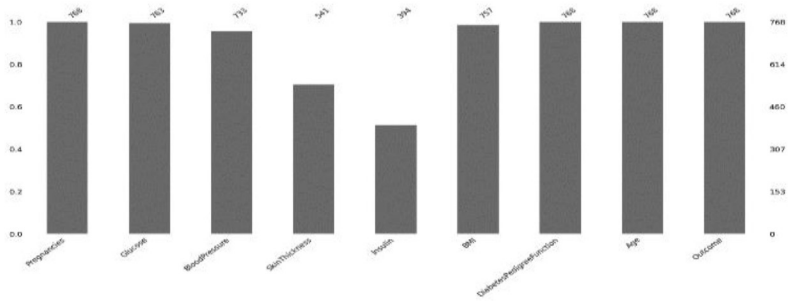

Figure 3

Visualization of the Missing Values using the Missing Library

replaced with 0[14]. Recognizing that these replacements were invalid, the missing values were replaced with NaN values as shown in Figure 3.

Outlier Observation Analysis: Outliers can significantly affect how well machine learning algorithm's function. To identify and address outliers, an analysis was conducted by comparing the attribute values to the 25th and 75th percentiles. A specific focus was given to the insulin variable, as shown in Figure 4.

4. Experiment Result and Analysis

By looking into various algorithms of machine learning, we sought to maximize the potential of the FOCUS framework in accurately predicting diabetes. The following sections of this paper will delve into the details of the experimental results, performance analysis, and comparison among the employed algorithms to determine the most effective approach for diabetes prediction using the FOCUS methodology.

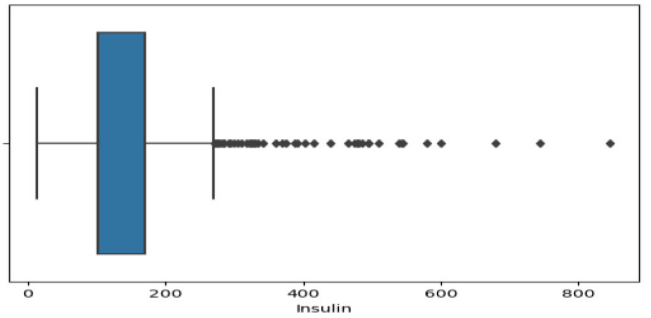

Figure 4

Outlier observation chart for Insulin.

Random Forest Classifier: A categorization method used in forecast modelling and behavioral feature analysis is random forest [12]. The decision trees that make up most of the random forest method each represent a unique instance.

Light Gradient Boosting Mechanism Classifier: Tree-based learning strategies are used by the gradient boosting system LightGBM [19]. Exclusive feature bundling (EFB) and gradient-based on-side sampling (GOSS) are the two methods it uses. GOSS will therefore ignore most of the data component with minor gradients and only use the remaining data to estimate the overall information gain. A dataset's accuracy can be determined using a variety of machine-learning algorithms.

Gradient Boosting Classifier: As the name implies, gradient boosting classifiers are a particular class of techniques used for classification tasks [20]. The machine learning algorithm's characteristics are the inputs that are supplied to it. These inputs are used to calculate the output value. The characteristics of the dataset serve as the variables in the equation in a mathematical sense.

Logistic Regression: A statistical technique called logistic regression is used to create machine learning models using binary, or dichotomous, dependent variables and independent variables. The types of independent variables include nominal, ordinal, and interval ones.

K-Nearest Neighbors Classifier: A fundamental machine learning algorithm called K-NN is based on the Supervised Learning approach. K-NN is frequently used for both classification and regression. The K-NN method considers how closely the new case and data resemble the old case and data.

Naïve Bayes Gaussian: The Bayes Theorem underpins the Naive Bayes classification model. This model considers the presumption of predictor independence. An issue instance is categorized using Naive Bayes using a conditional probability model. For each of the K possible outcomes or classes Ck, the probability that an entity has x=(x1,x2,x3,... xn) number of features (independent variables) is calculated as P(Ck | x1, x2,...xn). This is how the conditional probability is represented:

$$p(\mathrm{C}_k \mid x) = \frac{p(\mathrm{C}_k) \times p(x \mid C_k)}{p(k)} \tag{1}$$

In this instance, p(Ck) denotes the prior probability of class C, p(k) is the prior probability of the predictor, and p(x | Ck) denotes the likelihood, or probability of the predictor given class.

When considering any category C_k, j i and assuming that each feature x_i is conditionally independent of every other feature x_j, the following model can be represented:

$$p(C_k \mid x_1, x_2, x_3, \ldots x_n) = p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{2}$$

Decision Tree Classifier: Decision Tree is one of the robust and popular algorithms for classification and prediction. The categorization of instances based on features is represented by the decision tree model's tree structure. The training samples are represented by the first node of the decision tree, which has a tree-like structure. If every sample is a member of the same class, the node becomes the leaf and is identified by the class. If not, the algorithm chooses the discriminatory property as the current node of the decision tree.

XGBoost Classifier: The gradient boosting framework is used by the decision tree-based ensemble machine learning method XGBoost [20]. In forecasting issues involving complex data (images, text, etc.), artificial neural networks usually exceed all other methodologies or architectures. On the other hand, decision tree-based algorithms are currently recognized as the industry standard for managing small to medium-sized structured/ tabular data.

Support Vector Classifier: SVMs are employed to address the issues of classification and regression [21]. The decision boundary that the SVM provides is defined by the following equation:

Various classification algorithms, including Random Forest, Light Gradient Boosting Mechanism, Gradient Boosting, Logistic Regression, K-Nearest Neighbors, Decision Tree, Naïve Bayes Gaussian, Support Vector Classifier, and XGBoost were evaluated within the FOCUS framework for diabetes prediction [22].

Performance Analysis and Comparison

The analysis of the results of the proposed model using a variety of performance metrics, including Accuracy, F1 Score, Precision, and Recall Score is done. The optimal algorithm for developing web applications is the one with the highest prediction accuracy. The accuracy of the idea was compared to various other recent attempts at diabetes prediction in terms of performance comparison. The performance findings show that the concept can perform better than earlier experiments that are equivalent.

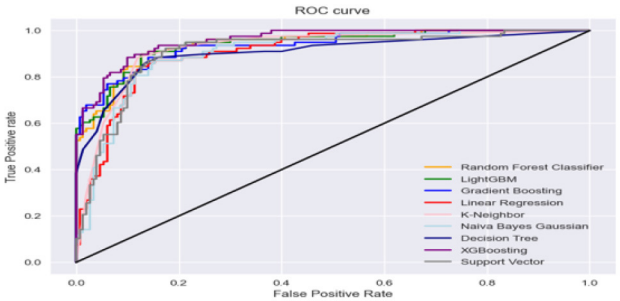

Figure 5

ROC Curve Analysis

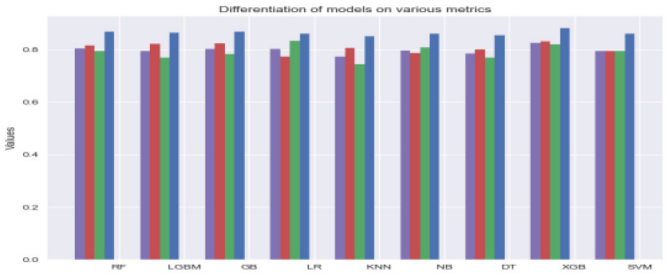

Figure 6

Performance Analysis of Applied Algorithms

Table II

Metrics result table for every algorithm

S. No.	Name of the Model	Accuracy	Recall Score	Precision	F1 Score
1	Random Forest Classifier	86.84	79.49	81.58	80.52
2	Light GBM Classifier	84.21	70.51	80.88	75.34
3	Gradient Boosting Classifier	86.84	78.20	82.43	80.26
4	Logistic Regression	85.96	83.33	77.38	80.25
5	K Nearest Neighbors	85.09	74.36	80.56	77.33
6	Naïve Bayes	85.96	80.77	78.75	79.74
7	Decision Tree	85.53	76.92	80.00	78.43
8	XG-Boost	88.15	82.05	83.11	82.58
9	Support Vector Machine	85.96	79.48	79.48	79.48

In Figure 5, we demonstrate the ROC curve analysis, which shows the differences between the models graphically. This analysis allows us to assess the models performance across different classification criteria and compare their overall effectiveness in distinguishing between positive and negative instances.

Furthermore, in Figure 6, we provide a comprehensive performance analysis of the applied algorithms. This graph demonstrates how the models differ based on a number of parameters, including Accuracy, Precision, Recall Score, and F1 Score.

Table II presents the metrics result table for each algorithm evaluated in the study. Among these algorithms, the XGBoost model, after hyperparameter tuning, achieved the highest accuracy of 88.15%.

5. Conclusions And Future Scope

The research objective was to develop a model that might identify diabetics who are most likely to be hospitalized. Predicting the likelihood of hospital admission is an extremely difficult task. The process and the outcome are affected by many variables. The understanding of the factors that affect predicting the likelihood of hospital admission by healthcare facilities must be improved immediately. This research evaluates the patient's medical history and using several machine-learning models to analyze several crucial markers, such as the person's body mass index and blood glucose level. Based on the relevant medical information obtained via a secure web application, the proposed model forecasts when diabetes will manifest itself in a person. A fair and trustworthy XGBoost model produced the forecast, and its accuracy rate was 88.15%.

References

[1] Kaggle Dataset, "Pima Indian Diabetes Database". https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (2017). Accessed: 2022-11-25.

[2] International Diabetes Federation, (2021). "Diabetes Facts and Figures". https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html#:~:text=Diabetes%20facts%20%26%20figures,-Last%20update%3A%2009&text=In%202021%2C,low%2D%20and%20middle%2Dincome%20countries/ . Accessed: 2022-11-20.

[3] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus—A machine learning approach," in 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 122–127, Trivandrum, India (2015).

[4] Md Shahin Ali et.al, "A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights", *BioMed Research International*, vol. 2023, Article ID 8583210, 15 pages (2023). https://doi.org/10.1155/2023/8583210.

[5] S. Perveen, M. Shahbaz, "Performance analysis of data mining classification techniques to predict diabetes," Procedia Computer Science, vol. 82, pp. 115–121 (2016).

[6] N. Nai-arun. and R. Moungmai "Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science., 69 , pp. 132-142 (2015).

[7] A. Choudhury and D. Gupta, "A survey on medical diagnosis of diabetes using machine learning," in Recent Developments in Machine Learning and Data Analytics: IC3 2018, Springer Singapore (2019).

[8] Anil Kumar & Sandeep Kumar Sharma. Information cryptography using cellular automata and digital image processing, *Journal of Discrete Mathematical Sciences and Cryptography*, 25:4, 1105-1111 (2022), DOI: 10.1080/09720529.2022.2072437.

[9] Q. Zou, K. Qu, Y. Luo, D. Yin, and Y. Ju and H. Tang, "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9: 515 (2018).

[10] H. Wu, S.Yang, Z. Huang, J. He, and X. Wang, X "Type 2 diabetes mellitus prediction model based on data mining." *Informatics in Medicine Unlocked* 10: 100-107 (2018).

[11] M. Somnath et.al, "Prediction of Diabetes Type-II Using a Two-Class Neural Network" Proceedings of the (2017) International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25, pp. 65–72 (2017).

[12] Sandeep Kumar Sharma, Anil Kumar, and Uday Pratap Singh. Enhanced Edges Detection from Different Color Space. In Proceedings of the 4th International Conference on Information Management & Machine Intelligence (ICIMMI '22). Association for Computing Machinery, New York, NY, USA, Article 16, 1–6 (2023).

[13] J. A. Hussain, I. R. White, M. J. Johnson, et al. "Development of guidelines to reduce, handle and report missing data in palliative care trials: A multi-stakeholder modified nominal group technique", *Palliative Medicine.*, 36(1), pp. 59-70 (2022).

[14] Jasuja Arush and Sonia Rathee. "Emotion Recognition Using Facial Expressions." *IJIRR* vol. 11, no. 3 : pp. 1-17 (2021). http://doi.org/10.4018/IJIRR.2021070101.

[15] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," in *IEEE Access*, vol. 9, pp. 114381-114391 (2021), doi: 10.1109/ACCESS.2021.3104357.

[16] S. N. Dhage and C. K. Raina "A review on Machine Learning Techniques" *International Journal on Recent and Innovation Trends in Computing and Communication* ISSN: 2321-8169, vol. 4 (3), pp. 395-399 (2016).

[17] Sharma, Sandeep Kumar, Kumar, Anil, Ashtagi, Rashmi & Jain, Rekha. OCA: An intelligent model for improving security breach of biometric based authentication systems, *Journal of Discrete Mathematical Sciences and Cryptography,* 26:5, 1415–1425 (2023), DOI: 10.47974/JDMSC-1765.

[18] Sharma, S.K., Kumar, A., Digital Image Transformation Using Outer Totality Cellular Automata. Machine Intelligence Techniques for Data Analysis and Signal Processing. Lecture Notes in Electrical Engineering, vol 997. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-0085-5_69.

[19] D. Mishra,et.al, "Light gradient boosting machine with optimized hyperparameters for identification of malicious access in IoT network" *Digital Communications and Networks*, vol. 9(1), pp. 125-137 (2023).

[20] A-A.Tanvir, I. A. Khandokar, "A gradient boosting classifier for purchase intention prediction of online shoppers" *Heliyon*, vol. 9(4) (2023).

[21] C. Cortes and V. Vapnik, "Support-vector networks", *Mach. Learn.,* 20 (3) , pp. 273-297 (1995).