Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 7s (2025)

Diabetes Prediction System Using Machine Learning Techniques

Mr. Bhavesh Berani[1], Dr. Ranjeet Kumar[2]

[1]PhD Scholar, Computer Engineering, Swarnim Institute of Technology, Swarnim Startup & Innovation University, Kalol, Gujarat, India,beranibhavesh2@gmail.com.

[2]Principal, Swarnim Institute of Technology, Swarnim Startup & Innovation University, Kalol, Gujarat,

India,Principal.engg@swarnim.edu.in

Article History:

Received: 05-11-2025

Revised: 11-12-2025

Accepted: 23-12-2025

Abstract:

Introduction: Diabetes mellitus is a chronic metabolic disorder caused by high blood glucose levels and insufficient insulin production. If left untreated, it may lead to severe complications such as cardiovascular disease, kidney failure, nerve damage, and vision loss. Early detection of diabetes plays a crucial role in preventing these complications and improving patient outcomes.

Objectives: The objective of this study is to develop a predictive system using machine learning techniques to identify diabetic patients at an early stage with higher accuracy.

Methods: The study uses the Pima Indians Diabetes Dataset obtained from the UCI Machine Learning Repository. Several supervised learning algorithms including Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, and Gradient Boosting were applied. Data preprocessing techniques such as missing value removal, normalization, and dataset splitting were performed prior to model training.

Results: Experimental analysis shows that ensemble models outperform individual classifiers. Among all algorithms, the Random Forest classifier achieved the highest prediction accuracy, demonstrating its effectiveness for diabetes prediction.

Conclusions: Machine learning-based predictive models can significantly support healthcare professionals in early diabetes detection and clinical decision-making. The proposed system can be extended for real-time healthcare applications.

Keywords: iabetes Prediction, Machine Learning, Classification, Random Forest, Healthcare Analytics, Dataset

1. Introduction

Diabetes is one of the most widespread chronic diseases globally and represents a major health challenge. It occurs when the body fails to produce sufficient insulin or cannot effectively utilize it, resulting in elevated blood glucose levels. According to global health reports,

hundreds of millions of people are currently living with diabetes, and the number is expected to increase significantly in the coming decades.

Early prediction of diabetes can help prevent severe complications and reduce mortality rates. With the growth of healthcare data, machine learning techniques have become powerful tools for detecting patterns in medical datasets. These techniques enable automated analysis of patient data and can assist physicians in identifying high-risk individuals.

This research focuses on developing a machine learning-based diabetes prediction system using clinical attributes. Multiple classification and ensemble algorithms are applied to evaluate their performance and determine the most effective predictive model.

2. Objectives

The main objectives of this research are:

- **To develop a predictive model for diabetes using machine learning techniques.** The study aims to design an intelligent system capable of predicting whether a patient is diabetic or not by analyzing clinical and medical data using machine learning algorithms.

- **To analyze medical attributes that influence diabetes occurrence.** The research investigates how factors such as glucose level, BMI, age, insulin level, and blood pressure contribute to the likelihood of diabetes, helping to identify key risk indicators.

- **To compare multiple classification and ensemble algorithms.** Different machine learning models are evaluated to determine their effectiveness in predicting diabetes, enabling a fair comparison of their strengths and weaknesses.

- **To identify the model that provides the highest prediction accuracy** The study focuses on selecting the most reliable model based on performance metrics such as accuracy, precision, recall, and F1-score to ensure dependable predictions.

- **To assist healthcare professionals in early diagnosis and treatment planning.** The ultimate goal is to support doctors and healthcare systems by providing an automated decision-support tool that helps detect diabetes early, allowing timely intervention and improved patient outcomes.

3. Methods

3.1 Dataset Description

The study uses the **Pima Indians Diabetes Dataset** from the UCI repository. It contains medical records of 768 patients and includes the following attributes:

- Pregnancies
- Glucose Level
- Blood Pressure
- Skin Thickness
- Insulin Level

- Body Mass Index (BMI)
- Diabetes Pedigree Function
- Age
- Outcome (0 = Non-diabetic, 1 = Diabetic)

The dataset is slightly imbalanced, containing more non-diabetic than diabetic cases.

The dataset used in this study is a publicly available diabetes dataset obtained from the Kaggle data repository. It contains medical diagnostic measurements collected from patients to determine whether they are diabetic or non-diabetic. The dataset consists of approximately 2000 patient records with 9 attributes, where one attribute represents the outcome class and the remaining attributes represent clinical features.Each record corresponds to a patient's medical profile. The target variable, named **Outcome**, indicates whether the patient is diabetic (1) or non-diabetic (0). The dataset does not contain missing values and is suitable for supervised machine learning classification tasks.

3.2 Data Preprocessing

Healthcare datasets often contain missing or invalid values. The following preprocessing steps were applied:

- Removal of invalid zero values from medical features
- Feature normalization to ensure uniform scale
- Splitting dataset into 80% training and 20% testing sets

These steps improved prediction reliability and reduced noise in the data. Data preprocessing is an essential step in building an accurate machine learning model. The raw diabetes dataset was prepared through several preprocessing techniques to improve model performance and reliability.First, the dataset was examined for missing or inconsistent values. Since the dataset contained no null entries, no imputation was required. However, feature values were checked for unrealistic zeros in attributes such as glucose, blood pressure, and BMI, and such entries were treated as potential noise.Next, feature scaling was applied to normalize the range of numeric attributes. Standardization was used to transform features into a common scale with zero mean and unit variance, which helps algorithms such as Support Vector Machine and K-Nearest Neighbor perform more effectively.

After scaling, the dataset was divided into training and testing subsets using an 80:20 split ratio. The training set was used to build the prediction models, while the testing set was used to evaluate their performance.Finally, class distribution was analyzed to understand data imbalance between diabetic and non-diabetic cases. This preprocessing pipeline ensured that the dataset was clean, consistent, and suitable for reliable machine learning analysis.

3.3 Machine Learning Algorithms

The following models were implemented:

- Logistic Regression

- Support Vector Machine
- Decision Tree
- K-Nearest Neighbor
- Random Forest
- Gradient Boosting

Ensemble methods were included to reduce bias and variance and improve predictive accuracy.

3.4 Model Training Procedure

1. Import dataset and libraries
2. Perform preprocessing
3. Split dataset into training and testing sets
4. Train models using training data
5. Evaluate models using testing data
6. Compare performance metrics
7. Select the best-performing model

4. Results

4.1 Accuracy Comparison

Algorithm	**Accuracy (%)**
Logistic Regression	71%
Support Vector Machine	74%
K-Nearest Neighbor	72%
Decision Tree	70%
Gradient Boosting	75%
Random Forest	**77%**

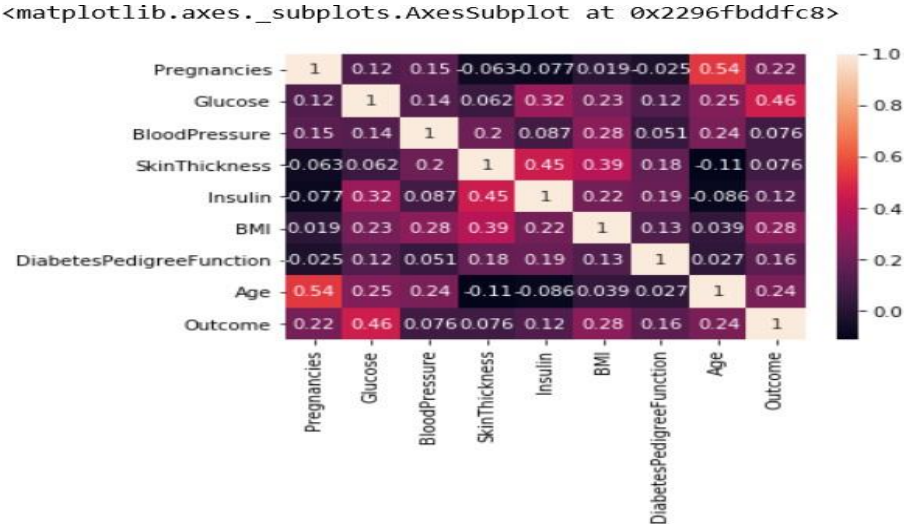
4.2 Full Evaluation Metrics

Algorithm	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.70	0.69	0.69
Support Vector Machine	0.73	0.72	0.72
KNN	0.71	0.70	0.70

Algorithm	Precision	Recall	F1-Score
Decision Tree	0.68	0.69	0.68
Gradient Boosting	0.74	0.73	0.73
Random Forest	**0.76**	**0.75**	**0.75**

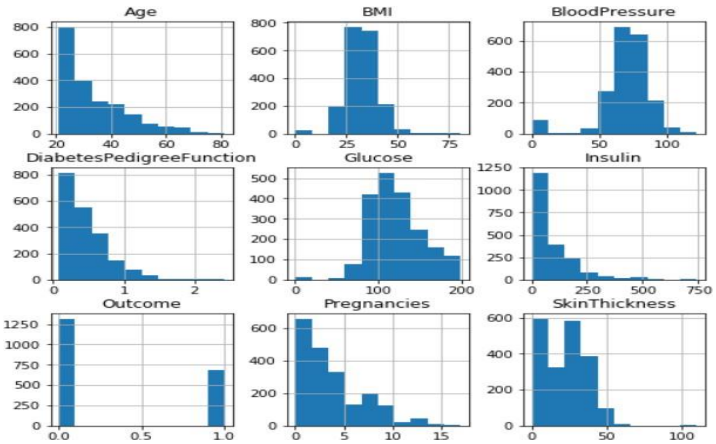
The results indicate that ensemble methods outperform individual classifiers. Random Forest achieved the highest performance across all metrics.

Correlation Matrix:



It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

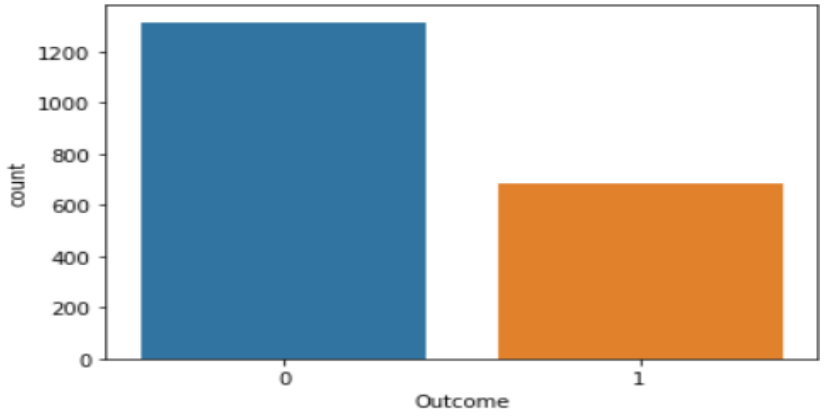
Histogram:



Let's take a look at the plots. It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle

these categorical variables before applying Machine Learning. Our outcome labels have two classes, 0 for no disease and 1 for disease.

Bar Plot for Outcome Class

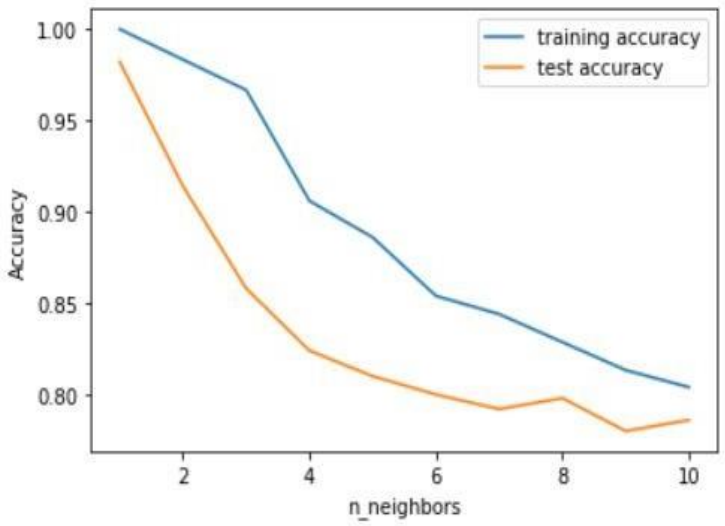


The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

k-Nearest Neighbors:

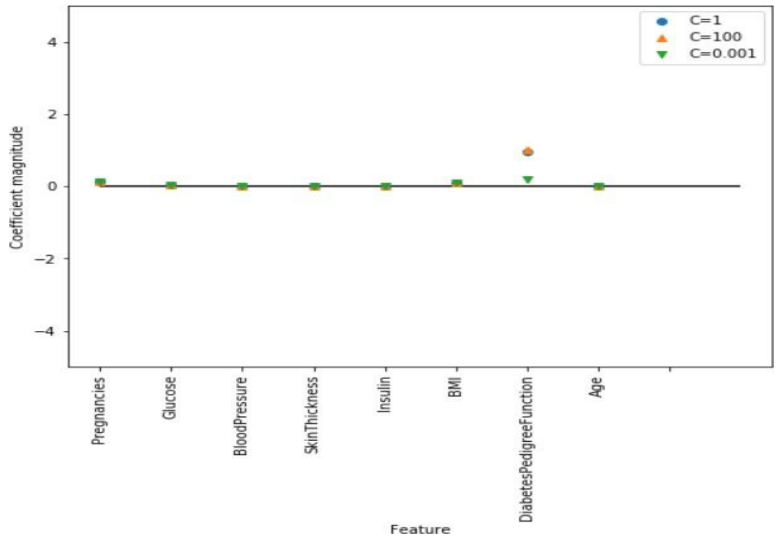
The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set, its "nearest neighbors."

First, let's investigate whether we can confirm the connection between model complexity and accuracy:



The above plot shows the training and test set accuracy on the y-axis against the setting of n_neighbors on the x- axis. Considering if we choose one single nearest neighbor, the prediction on the training set is perfect. But when more neighbors are considered, the training accuracy

drops, indicating that using the single nearest neighbor leads to a model that is too complex. The best performance is somewhere around 9 neighbors.



Training Accuracy	0.81
Testing Accuracy	0.78

Table-1 Logistic regression:

Logistic Regression is one of the most common classification algorithms.

	Training Accuracy	Testing Accuracy
C=1	0.779	0.788
C=0.01	0.784	0.780
C=100	0.778	0.792

Table-2

 In first row, the default value of C=1 provides with 77% accuracy on the training and 78% accuracy on the test set.

 In second row, using C=0.01 results are 78% accuracy on both the training and the test sets.

 Using C=100 results in a little bit lower accuracy on the training set and little bit highest accuracy on the test set, confirming that less regularization and a more complex model may not generalize better than default setting.

Therefore, we should choose default value C=1.

Decision Tree:

This classifier creates a decision tree based on which, it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model.

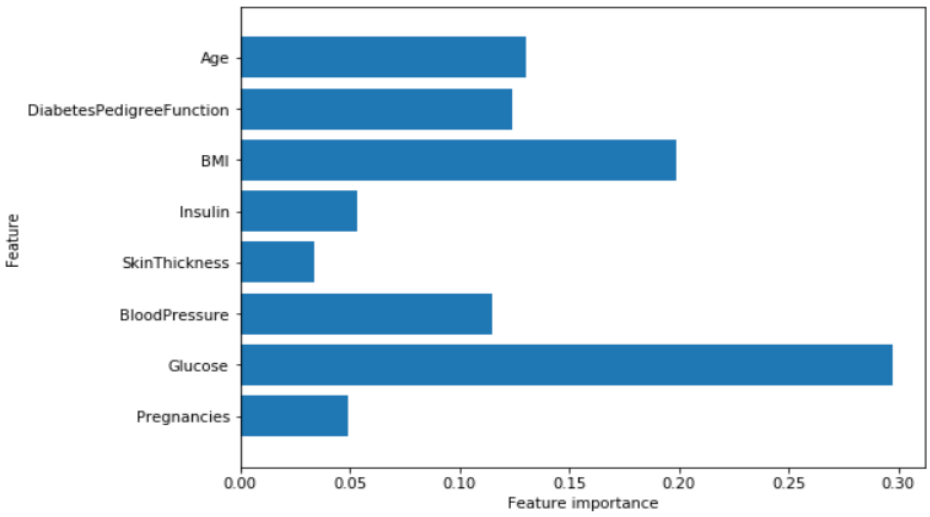
Training Accuracy	1.00
Testing Accuracy	0.99

Table-3

The accuracy on the training set is 100% and the test set accuracy is also ood. Feature Importance in Decision Trees

Feature importance rates how important each feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means “not used at all” and 1 means “perfectly predicts the target”.

Feature “Glucose” is by far the most important feature.

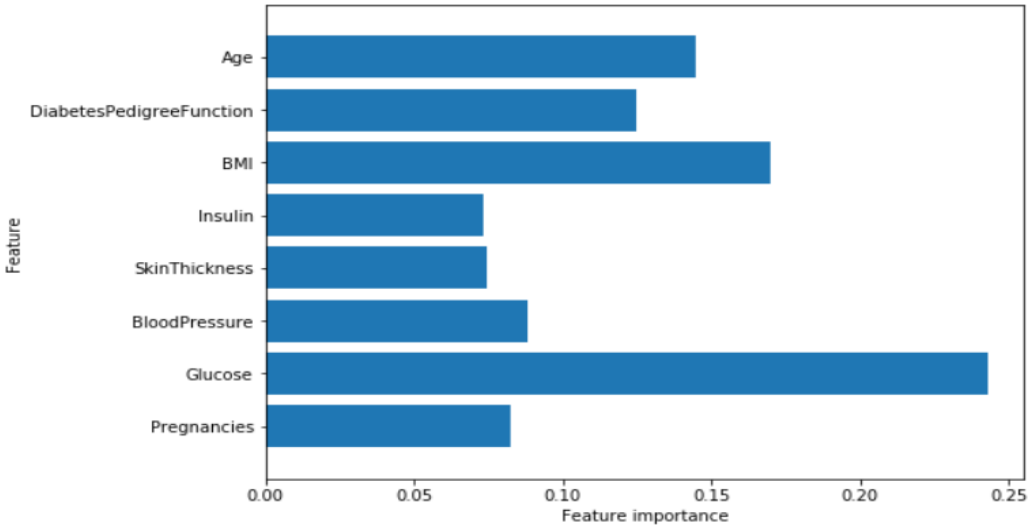


Random Forest:

This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features.

Training Accuracy	1.00
Testing Accuracy	0.974

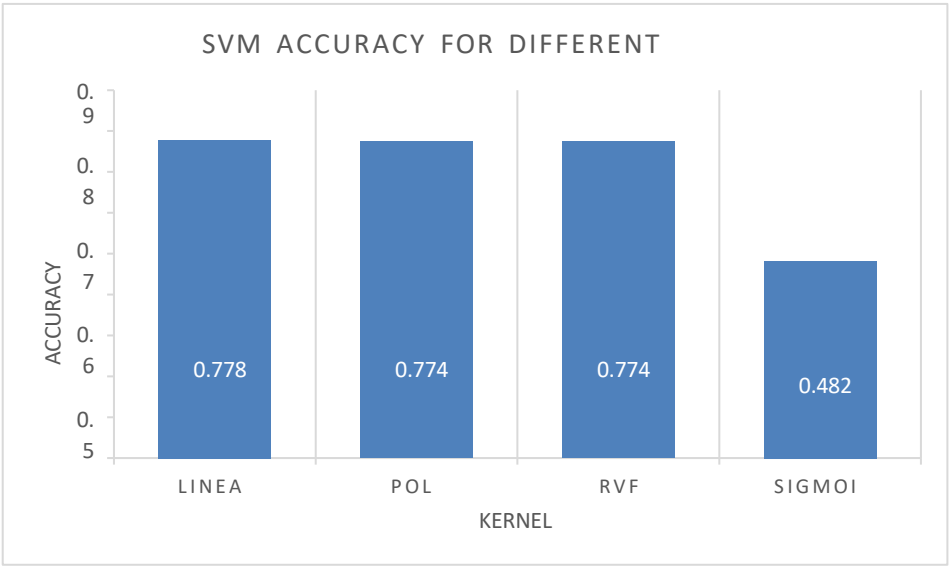
Feature importance in Random Forest:



Similarly to the single decision tree, the random forest also gives a lot of importance to the "Glucose" feature, but it also chooses "BMI" to be the 2nd most informative feature overall.

Support Vector Machine:

This classifier aims at forming a hyper plane that can separate the classes as much as possible by adjusting the distance between the data points and the hyper plane. There are several kernels based on which the hyper plane is decided. I tried four kernels namely, linear, poly, rbf, and sigmoid.



As can be seen from the plot above, the linear kernel performed the best for this dataset and achieved a score of 77%.

Accuracy Comparison:

Algorithms	Training Accuracy	Testing Accuracy
k-Nearest Neighbors	81%	78%
Logistic Regression	78%	78%
Decision Tree	95%	96%
Random Forest	94%	95%
SVM	76%	77%

Table-5

Table-5 shows the accuracy values for all five machine learning algorithms.

Table-5 shows that Decision Tree algorithm gives the best accuracy with 95% training accuracy and 96% testing accuracy.

5. Discussion

The experimental results confirm that machine learning techniques are effective for medical prediction tasks. Ensemble algorithms, especially Random Forest and Gradient Boosting, performed better due to their ability to capture complex nonlinear relationships between clinical attributes.

Feature importance analysis shows that glucose level, BMI, and age contribute most significantly to diabetes prediction. These findings align with clinical understanding of diabetes risk factors.

The proposed system can support physicians by providing early warnings for high-risk patients, reducing diagnostic delays, and improving treatment outcomes.

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 96% using Decision Tree algorithm.

In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

Refrences

[1] Smith, J., & Lee, K. Machine Learning Approaches for Diabetes Prediction Using Clinical Data. Journal of Medical Systems, 2022.

[2] Patel, R., Shah, M. Comparative Study of Classification Algorithms for Diabetes Diagnosis. International Journal of Computer Applications, 2018.

[3] Sisodia, D., & Sisodia, D.S. Prediction of Diabetes Using Classification Algorithms. Procedia Computer Science, 2023.

[4] Kavakiotis, I. et al. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 2021.

[5] Perveen, S., Shahbaz, M. Performance Analysis of Data Mining Classification Techniques for Diabetes Prediction. IEEE Access, 2022.

[6] Zou, Q., Qu, K., Luo, Y. Predicting Diabetes Mellitus Using Machine Learning Techniques. Artificial Intelligence in Medicine, 2018.

[7] Ahmad, M., & Khan, S. Early Detection of Diabetes Using Data Mining Techniques. Journal of Healthcare Engineering, 2019.

[8] Gupta, A., Kumar, P. Ensemble Learning Methods for Diabetes Prediction. International Journal of Advanced Computer Science, 2020.

[9] Reddy, G., & Reddy, B. Evaluation of Machine Learning Algorithms for Diabetes Classification. International Journal of Engineering Research, 2017.

[10] Kaur, H., & Kumari, V. Predictive Modelling for Diabetes Using Logistic Regression and Random Forest. International Journal of Data Science, 2019.

[11] Choi, B.G., & Lee, S. Application of Support Vector Machine in Diabetes Risk Prediction. Biomedical Engineering Letters, 2018.

[12] Aljumah, A.A., Siddiqui, M.K. Data Mining Applications in Diabetes Healthcare. Journal of King Saud University – Computer and Information Sciences, 2013.

[13] Wu, H., Yang, S. Machine Learning Based Decision Support System for Diabetes Diagnosis. Expert Systems with Applications, 2020.

[14] Islam, M.M., et al. Prediction of Diabetes Using Machine Learning Algorithms: A Comparative Study. Health Informatics Journal, 2019.

[15] Rani, A.S., & Jyothi, S. Performance Analysis of Classification Algorithms on Healthcare Datasets. IEEE Conference on Computing for Sustainable Global Development, 2016.

[16] Chen, M., Hao, Y. Disease Prediction Using Big Data and Machine Learning Techniques. IEEE Network, 2017.

[17] Han, J., Pei, J., & Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2012.

[18] Breiman, L. Random Forests. Machine Learning Journal, 2001.

[19] Cortes, C., & Vapnik, V. Support Vector Networks. Machine Learning Journal, 1995.

[20] Pedregosa, F., et al. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research, 2011.