Journal of Basrah Researches ((Sciences)) Vol. (47). No. 1 (2021)

Available online at: https://www.iasj.net/iasj/journal/260/issues

ISSN -1817 -2695

Received date: 16-7-2021 Accepted date: 25-9-2021 Available online date: 31-12-2021

A Comparative Analysis for Diabetic Prediction Based on Machine Learning Techniques

Nada Ali Noori[1], Ali A. Yassin[2]

Department of Computer science, Education College for Pure Sciences, University of Basrah, Basrah, 61004, Iraq.

Email: asnsn5@gmail.com[1], aliadel79yassin@gmail.com[2]

Abstract

In the last decades, living in modern large cities and unhealthy lifestyles have a negative impact on the health of people. According to the World Health Organization (WHO), Diabetic Mellitus (DM) is the most common diseases in the twenty-first century. DM is a chronic disease that has no final cure. With the rising mortality in the last years, the number of patients with diabetes worldwide will exceed 642 million in 2040. All the research on diabetes prediction in the previous years indicates that early diagnoses can decrease death rates and save human lives. In this regard, the machine learning techniques show promising results for analysis and predict diabetes mellitus at an early stage. However, the previous studies need to increase accuracy, more available medical information, balancing between performance and accuracy. This paper proposes a comparative Analysis study that can overcome the issues mentioned above and depend on classification and prediction techniques. The proposed work uses five common machine learning algorithms: K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) for obtaining the early predication, high accuracy, and performance compared with the related works. The experimental results indicate that SVM gets the highest accuracy (83 %) on the diabetic prediction based on the Pima Indian diabetes dataset (PIDD).

Keywords: Diabetic Prediction, Machine learning Algorithms, Classification, SVM, Comparative Analysis

1. Introduction

According to the reports of WHO, around 422 million people have diabetes in the past three decades, and nearly 1.6 million deaths are directly related to diabetes mellitus each year [1]. Figure (1) show diabetes prevalence worldwide in 2019 [2]. Diabetes is a common chronic disease that happens when the levels of blood sugar are abnormally high. Blood glucose is the primary energy source in the human body; it originates from the food we eat. Insulin is a blood hormone that travels from the bloodstream to the cells, instructing them to consume blood sugar and convert it to energy. Once the pancreas produces insufficient insulin, the cells cannot absorb sugar/glucose; consequently, the glucose remains in the bloodstream. As a result, blood glucose/blood sugar levels rise to unacceptably high levels [3]. High blood sugar causes various symptoms in humans, including excessive hunger, intense thirst, and frequent urine. Normal sugar levels in the human body are typically in the range of 75 to 100 mg per deciliter. Diabetes is diagnosed when the blood glucose level exceeds 126 mg/dl. If a person's blood glucose level is between 100 and 125 mg/dl, they have prediabetes [4].

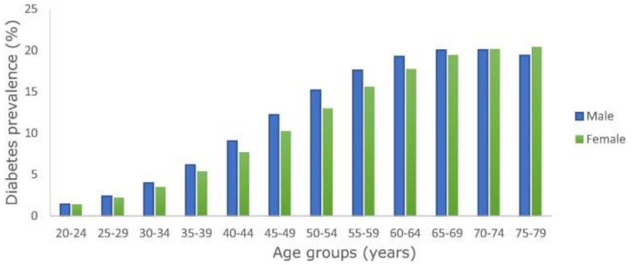


Figure (1). Explains the diabetes prevalence worldwide in 2019 [2]

If the blood sugar in the human body rises to a dangerous level, that will affect other organs and causes complication such as heart disease, nerve damage, kidney failure, and stroke [5]. Diabetes does not have a permanent cure [6]. Long-term diabetes produces macrovascular and microvascular complications, which are the most frequent health concerns. The macrovascular problem causes Damage to the big blood arteries of the heart, legs, and brain. The microvascular issue affects the small blood veins creating complications to the nerves, eyes, kidneys, and feet [7]. Diabetes can be effectively controlled if it is discovered at an early stage. A healthy workout routine and consuming a well-balanced diet, both habits help to avoid or prevent diabetes [8]. The National Diabetes Prevention Program, coordinated by the Centers for Disease Control and Prevention (CDCP), is a healthier lifestyle program that can help prediabetes patients to change and improve their lifestyle and avoid becoming Type 2 diabetes [9].

The healthcare industry gathers a massive amount of medical data, including patient's medical records, hospital information, and labs results. The disease's early diagnosis is examined through knowledge and doctor's experience; however, this can be imprecise and vulnerable. Hence the doctor's decisions can be inaccurate; that leads to preventing patients from the appropriate treatment. Therefore, more accurate identification with reasonable accuracy is required to predict diabetes[10-12]. We notice that the highest burden falls on the doctor in the diagnosis and treatment, which may be accompanied by a medical error that may lead to the patient's death.

In the era of information technology, machine learning and data mining have been developing, supporting a reliable tool in the healthcare domain, and taking a significant role in supporting doctors in making accurate diagnostic and preventing medical errors[27]. The data mining methods help preprocess the medical data and select the related features, while machine learning techniques are used to automate the diabetes prediction process [13].

This paper presents a practical study to analyze how a variety of classification algorithms (KNN, NB, LR, RF, SVM) behave when implemented to a PIDD data set. On the prediction side, the main benefit can protect the patients' lives in the future by presenting a tool that helps the doctor make accurate decisions and better diagnose for an early detection of diabetes mellitus. The experimental results were shown in the PID data set; the proposed work is appropriate for data analysis using Python programming language. Conversely, the essential contribution is to apply the results of the above referred ML algorithms for the early predicting diagnosis of diabetes. Furthermore, the results recommend the use of SVM early predicting the diagnosis of diabetes based on the accuracy (83%), precession (79%), recall (70%); finally, we obtain good results compared with previous work.

The organization of this paper, Section (2), is a literature review describing the previous work. Next, section (3) describes the methodology of the proposed system. Then, section (4) highlights the performance evaluation and the Experimental results gained after building the proposed system. Finally, section (5) is the Conclusion.

2. Related Work

Several papers used machine learning (ML) techniques to predict diabetes with the Pima Indian diabetes dataset (PIDD).

Alam et al. [14] use the Artificial Neural Network (ANN) technique on PIDD, the results showed the accuracy of (75%). Perveen et al. [19] used two datasets: Canadian Primary Care Sentinel Surveillance Network (CPCSSN) and PIDD. The information in CPCSSN related to gender, diastolic blood pressure (DBP), systolic blood pressure (SBP), triglycerides (TG), Body Mass Index (BMI), and fasting blood sugar (FBS). They used the decision tree model, Bootstrap Aggregating (BA), and Adaptive Boosting (AD). They found AD has a good result and can be implemented to predict other diseases like heart disease, diabetes, and hypertension. Although they obtain good accuracy, their work suffers from unbalancing between performance and accuracy. Sisodia et al. [15] use different machine learning methods NB, DT, and SVM, on PIDD; their results show that the NB algorithm has better accuracy (76%). Tigga et al. [16], using PIDD, implement logistic regression (LR) for diabetic prediction. They focus on some attributes such as glucose level, pregnancies, and BMI, that they consider these attributes most important than other attributes. To process the data and visualize the results, they use RStudio. Their Model is showing good prediction results with accuracy (75%).

Diwani et al. [17], using ten cross-validations to train and test all the patient's data with a decision tree (DT) and Naive Bayes (NB). Using WEKA tool to evaluate, compare and investigate the performance of their Model with many classification algorithms. The results display that the best predicted algorithm is NB with a value of accuracy (76%). Zou et al. [18] use DT, RF, and ANN for the classification model on PIDD. In addition, minimum Redundancy Maximum Relevance (MRMR) and Principal Component Analysis (PCA) methods are used for features reduction. Their results found that the RF algorithm with the MRMR has the best accuracy (77%).

We notice that the main problem in the machine learning methods is to select the best features and the appropriate classifier based on the type of system. This paper proposes a comparative analysis study for the early diagnoses of patients' diabetic based on two parts classification and prediction. Our work pays more attention to select and enhance features from PIDD that are applied on machine learning algorithms like Naive Bayes (NB), SVM, Linear Regression (LR), Random Forest (RF), K-Nearest Neighbor (KNN). As a result, the proposed study gets the best result based on accuracy and achieves a good balance between high performance and accuracy. We use two evaluation methods: K-Cross Validation and Train Test Split indicated the proposed study that obtains the best results compared with the previous works.

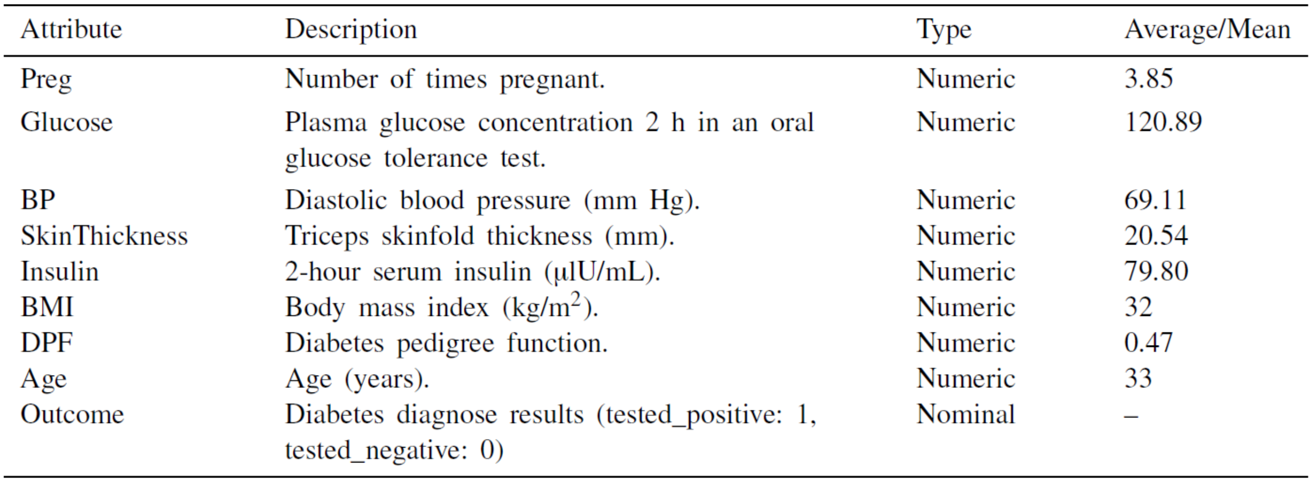
3. The proposed Model

In this section, we propose a classification model that consists of five parts as follows.

3.1. Data Set Part

In our research, the Pima Indian diabetes dataset (PIDD) is used. All the patients in the PID dataset are females. The dataset contains (768) instances and their associated nine attributes. Table 1 demonstrates the description of the dataset and the corresponding attributes. The nine attributes are Pregnancy, Age, BMI, Insulin level, Skin thickness, Blood pressure, Diabetes pedigree function, Glucose, and Outcome. The target or dependent variable is the outcome attribute, consisting of two binary values (1 means diabetes and 0 means non-diabetes). The remaining attributes are considering the independent features/ variables [20].

Table 1
The attributes of PIMA dataset.



Attribute	Description	Type	Average/Mean
Preg	Number of times pregnant.	Numeric	3.85
Glucose	Plasma glucose concentration 2 h in an oral glucose tolerance test.	Numeric	120.89
BP	Diastolic blood pressure (mm Hg).	Numeric	69.11
SkinThickness	Triceps skinfold thickness (mm).	Numeric	20.54
Insulin	2-hour serum insulin (μlU/mL).	Numeric	79.80
BMI	Body mass index (kg/m^2).	Numeric	32
DPF	Diabetes pedigree function.	Numeric	0.47
Age	Age (years).	Numeric	33
Outcome	Diabetes diagnose results (tested_positive: 1, tested_negative: 0)	Nominal	–

3.2. Data Preprocessing Part

The primary function of this part is used to enhance data that existed in the data set. The preprocessing output helps to build the best machine learning model that can provide better accuracy. The preprocessing performs various functions: delete the outlier values, filling the missing data, and do data normalization. For example, in the PID dataset, 500 instances were classified as non-diabetics, and 268 instances are diabetic.

3.2.1. Missing data identification

Table 2 explains the missing values of each attribute in the PID datasets. We contribute to overcoming missing values by replaced with the corresponding mean value. So, the importance of the current process will reflect on the results (accuracy and performance) of the following parts.

Table 2. The missing values in each attribute of the PID dataset.

Attributes	**Numbers of missing values**
Glucose	5
Pregnancy	0
BP	36
Insulin	374
Skinthikness	227
DPF	0
Age	0
BMI	11

3.2.2. Normalization

Normalizing the data in the range [0-1] help us to performed feature scaling, which boosted the time processing of ML algorithms [21]. We achieve the targets of this process by increasing the time processing of our proposed Model. After finishing preprocessing part, we have 698 instances/patients where 467 instances have no diabetes, and 232 instances have diabetes. Therefore, this part enhances the data set, which becomes more ready to use in the following parts.

3.3.Feature/Attribute Extraction Part

To enhance the quality of the data, feature extraction is essential in the classification model. Pearson's correlation approach is a widely used method for determining the most relevant features. Table (3) shows the relationship between input and output features after preprocessing; we can observe that Glucose and Outcome are highly correlated have a correlation coefficient of 0.49.

The value of the coefficient remains in the range of (1,-1). A significant correlation is shown by a value above 0.5, whereas no correlation is indicated by a value of zero. The outcomes are displayed in Table 3; for relevant attributes, we selected a cut-off of 0.2. As a result, factors such as blood pressure, DPF, and skin thickness are removed. Insulin, Glucose, BMI, Pregnancy, and Age are the five most important input variables. IN our proposed Model, this part reduces the size of the data set, leading to increased performance and reducing data dimension.

Table 3. The correlation values between the input and the output features

Features	**Correlation values**
Insulin	0.30
Glucose	0.49
Body Mass Index	0.31
Pregnancy	0.23
Age	0.23
Skin thickness	0.21
BP	0.21
DPF	0.20

3.4 Dataset Training Part

After preprocessing and data filtering, the dataset is ready to train and test. In our system, we used two methods for training and testing the dataset. The first one, known K-fold cross-validation used to split the data set, while the second one is called train-test splitting, implemented the same task of the first method, but it is different in the mechanism work. Figure (2) show each technique. This part plays a crucial role in providing data to the classification part.

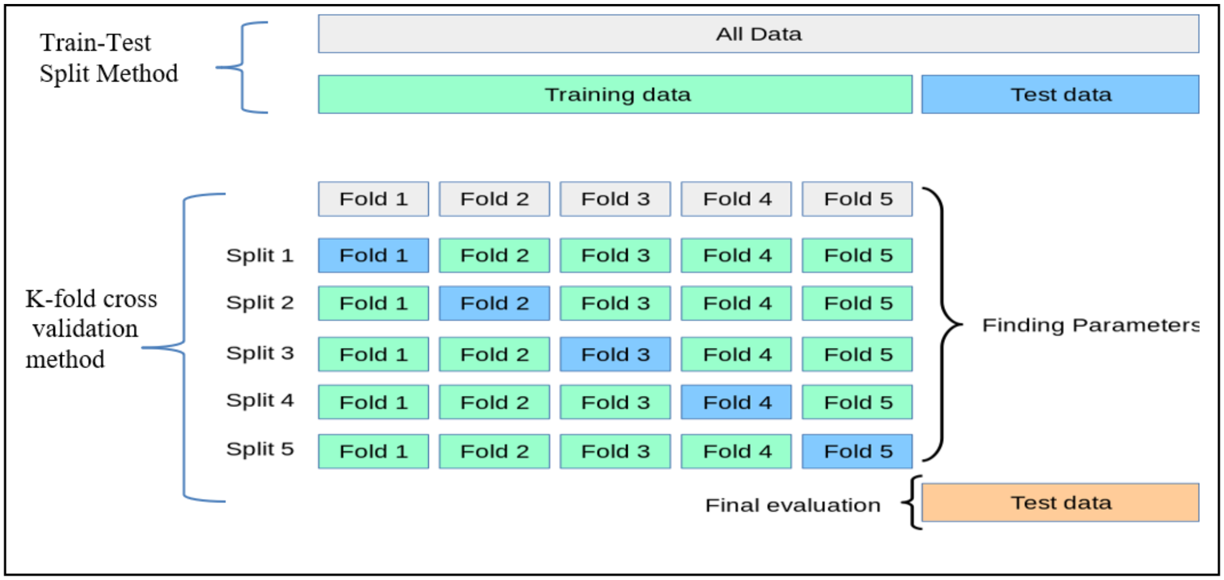


Figure 2.K-fold cross-validation method and train-test splitting method [22]

3.5 ML Algorithms and Classification Part

The classification model depends on the above parts to receive the training data/testing data. The training data will process using five machine learning algorithms: K-Nearest Neighbor, Naive Bayes, Logistic Regression, Random Forest, and SVM. The output refers that SVM can get the best result for classification data. On the other side, the testing data also needs to pass in the current part for performing classification based on the best parameters that enroll from the ML algorithms. The following figure (3) depicts the workflow of the proposed Model.

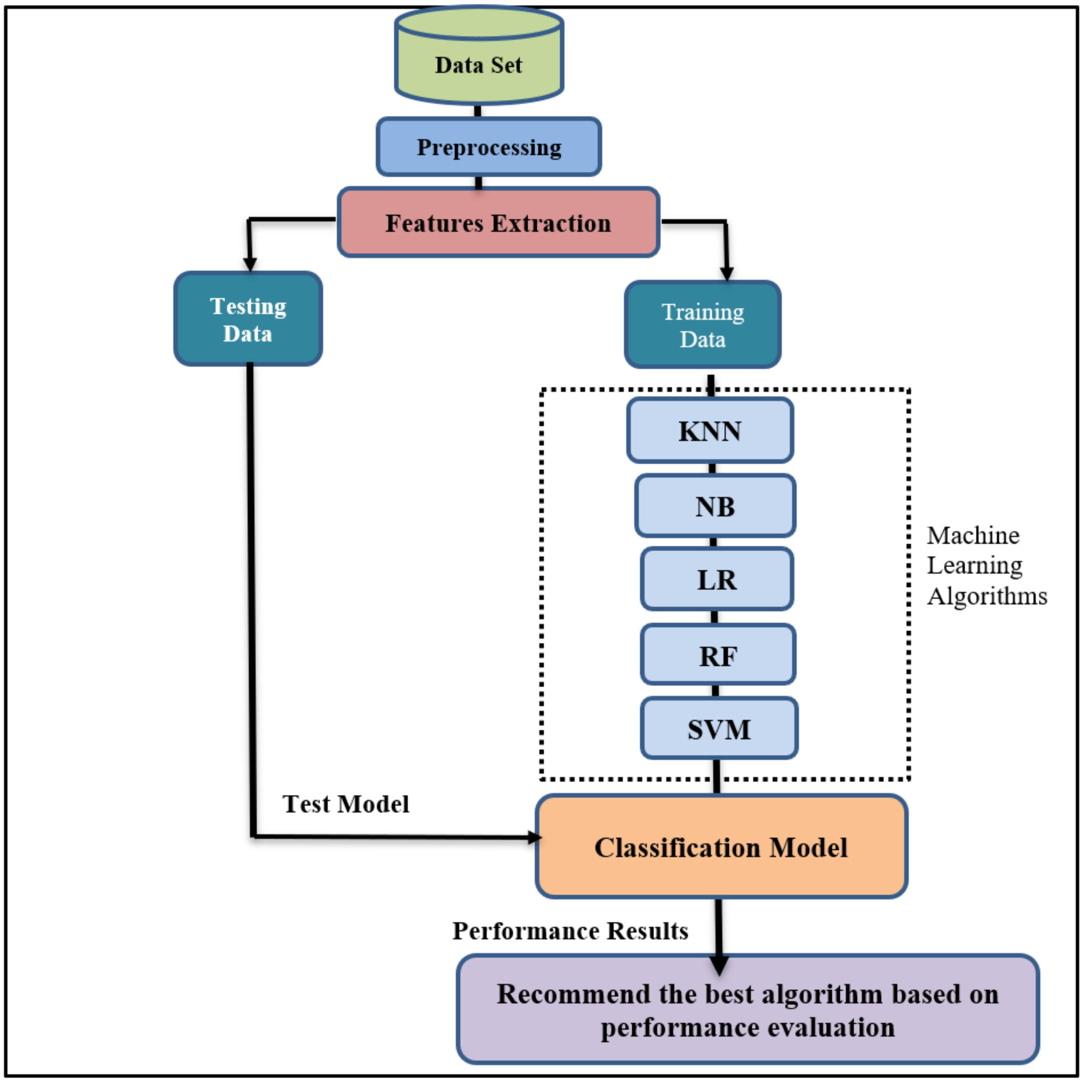


Figure 3. The Proposed Classification Model

The following steps describe the Proposed Classifier Model

Step 1: the input dataset is preprocessed, then select the best features from the dataset

Step 2: Using two methods to splitting the dataset: in the train-test splitting method, the percentage split of (70%) for the Training set and (30%) for the Testing set.
The second method is the cross-validation method uses One-fold for validation and the remaining are used for training the dataset.

Step 3: Implement the machine learning algorithm: K-Nearest Neighbor, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine algorithm.

Step 4: Based on the training set, building the classification model for the ML algorithms

Step 5: Based on the testing set, the classification model is testing for the ML algorithm

Step 6: For each algorithm, perform a comparative evaluation of the experimental performance results

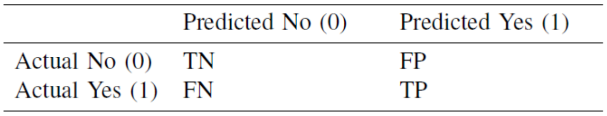
Step 7: Based on various measurements, the system is analyzed, and the best performing algorithm is detected.

The proposed classification model has been built using the Python programing language and based on the execution of each step; the experimental results are evaluated in the next part.

Result and Discussion

This section discusses our proposed model's results based on some measurements (Equations (1 - 4)) such as accuracy, precision, recall, and F1-Score. The confusion matrix considers the main factor for computing these measurements. Table 4 refers to the confusion matrix as below.

Table 4. Confusion matrix structure



	Predicted No (0)	Predicted Yes (1)
Actual No (0)	TN	FP
Actual Yes (1)	FN	TP

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{True Positive+True Negative+False negative+False Positive}} \quad \ldots \quad (1)$$

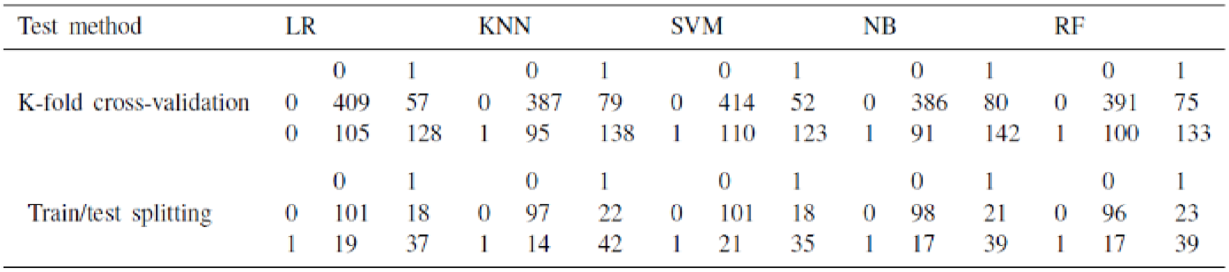
$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False negative}} \quad \ldots \quad (2)$$

$$\text{Precession} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad \ldots \quad (3)$$

$$\text{F1-SCOR} = \frac{2\times Precession \times Recall}{Precession \times Recall} \quad \ldots \quad (4)$$

In our work, table 5 denotes our study's confusion matrix based on cross-validation and train-test splitting. The performance measure values of the proposed model are shown in Table 6. The preprocessing and feature extraction parts are played the leading role in enhancing the results of our Model. For example, we can see that all classification ML algorithms used in our work are above 75%. Consequently, the SVM algorithm is showing better accuracy for both testing methods. In table 6, we see the results in the train/test splitting better than K-fold cross-validation because train/test splitting does not require overlapping between training data and testing data, and this method recommends medical data classification (see Figures (4,5)).

Table 5. Confusion matrices for KNN, RF, NB, LR, SVM classifier.



Test method	LR			KNN			SVM			NB			RF		
K-fold cross-validation		0	1		0	1		0	1		0	1		0	1
	0	409	57	0	387	79	0	414	52	0	386	80	0	391	75
	0	105	128	1	95	138	1	110	123	1	91	142	1	100	133
Train/test splitting		0	1		0	1		0	1		0	1		0	1
	0	101	18	0	97	22	0	101	18	0	98	21	0	96	23
	1	19	37	1	14	42	1	21	35	1	17	39	1	17	39

Table 6. The performance measure of all classification methods for K-fold cross-validation and Train/Test splitting method.

Classification	Precession	Recall	F-Score	Accuracy
KNN(K-Fold)	0.75	0.75	0.75	75%
KNN(Splitting)	0.76	0.64	0.69	79.1%
NB(K-Fold)	0.75	0.76	0.76	76.3%
NB (Splitting)	0.71	0.74	0.73	80.0%
LR (K-Fold)	0.76	0.77	0.75	77.2%
LR(Splitting)	0.75	0.61	0.67	81.0%
RF (K-Fold)	0.74	0.76	0.75	75.6%
RF(Splitting)	0.70	0.74	0.72	81.3%
SVM (K-Fold)	0.76	0.77	0.76	77.4%
SVM (Splitting)	0.79	0.63	0.70	83%

Table (7) compares the accuracy of the classification model in our work and previous works. Finally, the results showed that our Model considers the best classification technique.

Table (7) comparative analysis of Model's accuracy (%) with previous studies

Classifier	[23]	[24]	[25]	[26]	Proposed model
KNN	72.0	63.04	73.43	71.3	79.1
NB	67.0	73.48	75.52	-	81.1
LR	76.0	-	77.60	72.39	81.2
RF	72.0	75.39	74.30	74.4	81.4
SVM	68.0	77.73	65.63	73.43	83

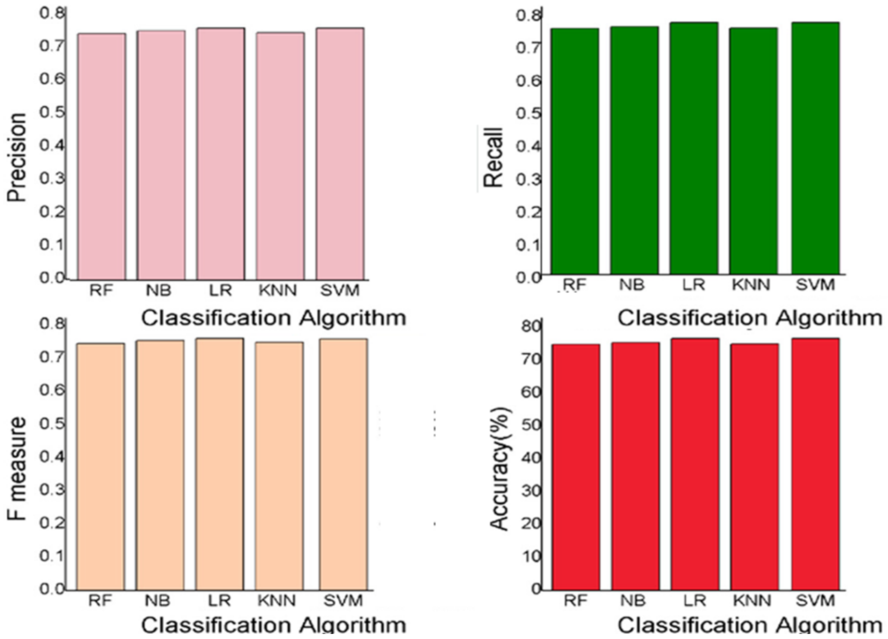


Fig. 4. The performance of all classifiers with a 10-fold cross-validation method.

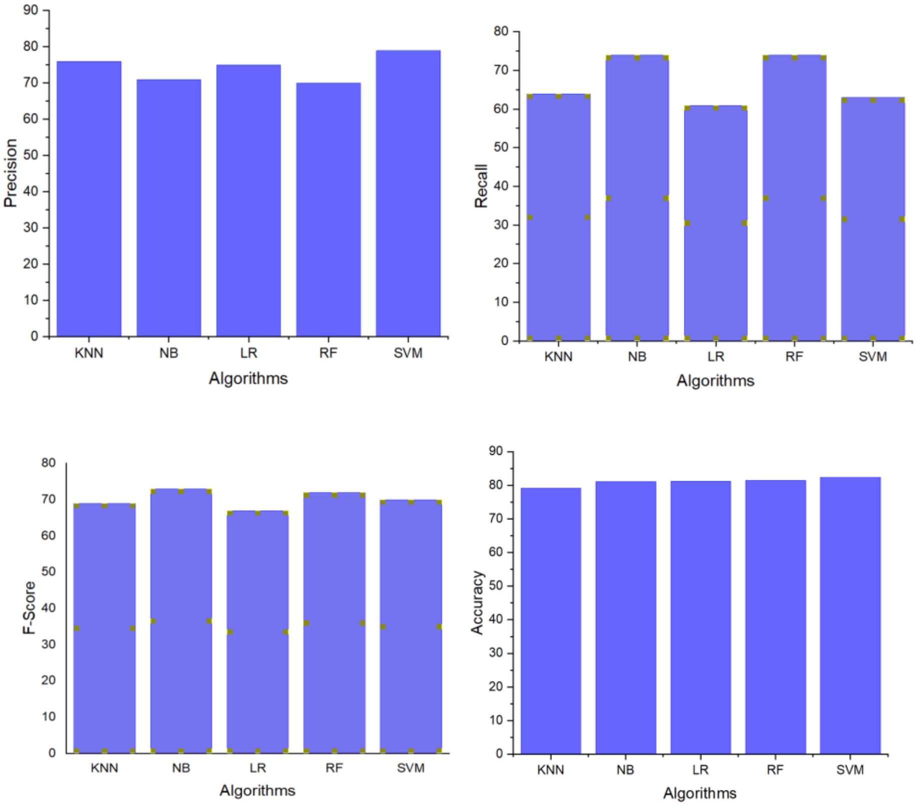


Fig.5.The performance of all classifiers with the train-test split method

Conclusion

People's health is always been a priority ov er the years, even before the existence of advanced technology. Therefore, the researchers always propose the best methods to provide better results, especially when dealing with medical data. Machine learning techniques and data mining tools show promising results in Health Care Information Technology, exclusively with the revolution of big data and the urgent need to analyze, understand, and interpret all this medical information. This study aims to build a classification model based on five machine learning algorithms Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Random Forest, and Support Vector Machine, and then select the best algorithm that has the best performance for early detection of diabetic Mellitus. Our model enhances the PID data set by overcoming the missing data, duplicate data, and outlier data. This enhancement reflects on the performance and accuracy of our Model compared with related works. On the healthcare side, our Model supports the medical systems by early detection of diabetes mellitus and impacts how a patient will live for the rest of his life. The experimental results show that our work has the best results comparing to other works.

REFERENCES

[1] https://www.who.int/health-topics/diabetes.

[2] https://www.diabetesresearchclinicalpractice.com

[3] D. Meetoo, “Chronic diseases: the silent global epidemic.,” *Br. J. Nurs.*, vol. 17, no. 21, pp. 1320–1325, 2008, doi:10.12968/bjon.2008.17.21.31731.

[4] M. Güemes, S. A. Rahman, and K. Hussain, “What is a normal blood glucose?,” Arch. Dis. Child., vol. 101, no. 6, pp. 569–574, 2016, doi: 10.1136/archdischild-2015-308336.

[5] American Diabetes Association, Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2010 Jan; 33(Suppl 1): S62–S69.doi: 10.2337/dc10-S062.

[6] https://www.healthgrades.com/right-care/diabetes/is-there-a-cure-for-diabetes.

[7] Viigimaa, Margus & Sachinidis, Alexandros & Toumpourleka, Maria & Koutsampasopoulos, Konstantinos & Alliksoo, Signe & Titma, Tiina. (2019). Macrovascular Complications of Type 2 Diabetes Mellitus. Current Vascular Pharmacology.17.10.2174/1570161117666190405165151.

[8] S.A. Kaveeshwar, J. Cornwall, The current state of diabetes mellitus in India, Australas. Med. J. 7 (1) (2014) 45.

[9] https://www.cdc.gov/diabetes/basics/prediabetes.html.

[10] C.L. Huang, M.C. Chen, C.J. Wang, Credit scoring with a data mining approach based on support vector machines, Expert Syst. Appl. 33 (4) (2007) 847–856, http://dx.doi.org/10.1016/j.eswa.2006.07.007.

[11] J. Chaki, S. Thillai Ganesh, S.K. Cidham, S. Ananda Theertan,Machine learning and artificial intelligence-based diabetes mellitus detection and self-management: A systematic review, J. King Saud Univ. - Comput. Inf. Sci. (2020).

[12] I. Contreras, J. Vehi, Artificial intelligence for diabetes management and decision support: Literature review, J. Med. Internet Res. 20 (5) (2018) e10775.

[13] G. Swapna, R. Vinayakumar, K.P. Soman, Soman KP diabetes detection using deep learning algorithms, ICT Express 4 (4) (2018) 243–246, http://dx.doi.org/10.1016/j.icte.2018.10.005, Elsevier B.V.

[14] T.M. Alam, et al., Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.

[15] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.

[16] N.P. Tigga, S. Garg, Predicting type 2 Diabetes using Logistic Regression accepted to publish in: Lecture Notes of Electrical Engineering,Springer.

[17] Salim Amour Diwani, Anael Sam, Diabetes forecasting using supervisedlearning techniques, Adv. Comput. Sci.: Int. J. [S.l.] (ISSN:2322-5157) (2014) 10–18.

[18] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9,Frontiers in genetics, 2018, p. 515.

[19] S. Perveen, M. Shahbaz, A. Guergachi, K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes,Procedia Comput. Sci. 82 (2016) 115–121.

[20] M. Lichman, Pima Indians diabetes database, ed. Center for machine learning and intelligent systems.: UCI Machine Learning repository.

[21] H. Benhar, A. Idri, J. Fernández-Alemán, Data preprocessing for decision making in medical informatics: potential and analysis, in:World Conference on Information Systems and Technologies, 2018,pp. 1208–1218.

[22] https://scikitlearn.org/stable/modules/cross_validation.

[23] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, DOI: 10.1016/j.procs.2020.01.047.

[24] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” *J. Big Data*, vol. 6, no. 1, 2019, DOI: 10.1186/s40537-019-0175-6.

[25] R. Deo and S. Panigrahi, “Performance Assessment of Machine Learning-Based Models for Diabetes Prediction,” *2019 IEEE Healthc. Innov. Point Care Technol. HI-POCT 2019*, no. 11, pp. 147–150, 2019, DOI: 10.1109/HI-POCT45284.2019.8962811.

[26] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” *Procedia Comput. Sci.*, vol. 167, no. 01, pp. 706–716, 2020, DOI: 10.1016/j.procs.2020.03.336.

[27] Shamis N. Abd, Mohammad Alsajri, and Hind Raad Ibraheem, “Rao-SVM Machine Learning Algorithm for Intrusion Detection System,” *Iraqi J. Comput. Sci. Math.*, pp. 23–27, 2020, doi : 10.52866/ijcsm.2019.01.01.004.