*Asian Journal of Research in Computer Science*

*12(3): 33-43, 2021; Article no.AJRCOS.76525*
*ISSN: 2581-8260*

# Effectiveness of Classification Methods on the Diabetes System

**Ahmed T. Shawky [a*] and Ismail M. Hagag [a]**

[a] *El Madina Higher Institute of Administration and Technology, Egypt.*

***Authors' contribution***

*This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.*

***Article Information***

DOI: 10.9734/AJRCOS/2021/v12i330287
Editor(s):
(1) Dr. Xiao-Guang Lyu, Huaihai Institute of Technology, China.
Reviewers:
(1) Prasath S., VET Institute of Arts and Science Co-education College, India.
(2) Piotr Artiemjew, University of Warmia and Mazury in Olsztyn, Poland.
(3) Myelinda A. Baldelovar, DIT, Surigao del Sur State University, Philippines.
Complete Peer review History: https://www.sdiarticle4.com/review-history/76525

***Case Study***

***Received 07 September 2021***
***Accepted 15 November 2021***
***Published 19 November 2021***

## ABSTRACT

In today's world using data mining and classification is considered to be one of the most important techniques, as today's world is full of data that is generated by various sources. However, extracting useful knowledge out of this data is the real challenge, and this paper conquers this challenge by using machine learning algorithms to use data for classifiers to draw meaningful results. The aim of this research paper is to design a model to detect diabetes in patients with high accuracy. Therefore, this research paper using five different algorithms for different machine learning classification includes, Decision Tree, Support Vector Machine (SVM), Random Forest, Naive Bayes, and K- Nearest Neighbor (K-NN), the purpose of this approach is to predict diabetes at an early stage. Finally, we have compared the performance of these algorithms, concluding that K-NN algorithm is a better accuracy (81.16%), followed by the Naive Bayes algorithm (76.06%).

*Keywords: Data mining; machine learning algorithms; decision tree; naïve bayes; support vector machine; random forest; and k- nearest neighbor; health care.*

## 1. INTRODUCTION

Using classification methods in the medical field is considered to be very widely used in today's world trying to approach better detection or even vaccines for certain diseases under certain constraints. Diabetes is considered to be a chronic disease in the world and it's about the

*Corresponding author: E-mail: ah_taisser@hotmail.com;

level of sugar in the blood, where it becomes too high [1,2]. Diabetes Mellitus (DM) is an illness that causes the human body to lower the production of the insulin hormone, which in return makes the level of glucose raises, and the metabolism and especially of carbohydrates abnormal. The patient then faces some symptoms such as intensify thirst, and hunger, and frequent micturition square. However, the main problems happen further when diabetes stays in the human body as it causes more complications such as diabetic ketoacidosis and nonketotic hyperosmolar coma. And the problem doesn't stop there as diabetes is also spreading all over the world, as stated by the world health organization, by 2030 there will be approximately over 700 million patients with diabetes. Diabetes is a disease that is spread all over the world. However, it's more common in developed countries [3,4].

The most common symptoms for patients of diabetes are weight loss, slow-healing in wounds, thirst, frequent urination, increased hunger, and more.[5].

Type I: Insulin-dependent diabetes as it is being called, happens before the age of 30 and that is usually due to the lack (or) deficiency of insulin, and usually this type affects children. In people with Type I diabetes, the beta cells of the pancreas, as it is the body part that is responsible for producing insulin, are destroyed and that's because of the autoimmune system.

Type II: and it is called non-insulin-dependent diabetes. And that type usually occurs to people over 40 years old. The main cause of this type is overweight, obesity, lack of physical activities, poor diet, and family history.

Using extracting information from the data mining process, from a large database, and medical-related diabetes, as it is a multidisciplinary field that involves computational processes in the computing area. Using statistical techniques, machine learning, clustering, classification, and discovering hidden patterns in the diabetes disease. Data mining techniques have recently been employed very widely to predict the data patterns such as time-series[6]. Using different algorithms in the machine learning process to predict the disease with high accuracy, we can help save human life and reduce the effect of the treatment [7]. As the discovery of the disease in earlier stages the better ways that we are able to control it. Data mining techniques may help people make a better judgment regarding diabetes using their daily physical examination of the data. This paper presents the study of which we predict diabetes using different algorithms of the data mining techniques. The main focus of the paper is implementing different techniques from the data mining on diabetes dataset and building prediction models to predict the disease using supervised learning. So, first, we will focus on preparing the classifier models depending on several data mining techniques and then moving forward and implementing these models to predict diabetes on the available dataset and then we compare the performance of each classifier model in that we have produced in the first step to analysis the accuracy of the different models. We used different data mining classification algorithms in this study for creating the classifier models and these algorithms are, Naive Bayes, Support Vector Machines (SVM), Decision tree, Random forest, and K-NN [8, 9]. The importance of this study comes from helping to decide which one of these algorithm classifications are supposed to have the best accuracy for this database of diseases. As we perform evaluation depends on many standard parameters, but we valued accuracy and training time as the most important parameters. Which eventually leads to the most suitable algorithm for high accuracy and less time to build the model.

The remaining part of this paper will be divided into 5 sections. First, we introduce the main idea. Secondly, we will present the literature review and related work. In the third section, we will discuss the methodology and the models that are proposed. And the result of the models will be presented in the fourth section. And lastly in the fifth section, the conclusion, and recommended future work.

### 1.1 Related Works

The researchers in the area of data mining usually start with the discussion of how much importance of the data mining approaches and the different classification algorithms in classifying the data, and ow is it important for the decision-making process, and especially for the medical practitioners, and how such an approach could be used to prevent the disease at an early stage which usually leads to reduce the effect of the diseases over a specific group of people. A few exceptional works have been illustrated in this section.

Orabi et al. In this paper [10] worked on deciding a system to predict diabetes diseases, to find out at a specific age if the candidate is suffering from diabetes or not. This planned system is supposed to be supported by the conception of the machine learning techniques, and the researcher has done this using the decision tree algorithm. The results of this system were satisfactory and that's because the system is designed well in predicting diabetes incidents at a specific age, and the accuracy using the decision tree algorithm was high.

Pradhan et al. In this paper [11] focused on ANN logistic regression and J48 to create a classifier model that diagnoses diabetes. And the researchers have stated that the J48 technique of machine learning provides efficiency and scores better accuracy among other techniques, for ANN to predict better predictions of diabetes, the author suggested that using the multilayer Perceptron and Genetic algorithm for feature selection. The author didn’t evaluate the classification algorithm using the cross-validation evaluation method.

Kumar et al. In this paper [12] tried the predictions with other algorithms such as support vector machines (SVM), Naive Bayes, and, logistic regression, with the 10-fold cross-validation. The goal of their work was to find out the best model from different machine learning algorithms that can be used to predict diabetes by using these techniques on a dataset of patients, the researchers have compared the different algorithms based on the accuracy parameter, and they found that the support vector machine (SVM) was the best accuracy score among other algorithms.

Alkargole et al.[13] the authors in this paper tried different data mining techniques as a hybrid framework using the Pima Indian dataset. And they applied classification algorithms like decision tree, Naive Bayes, and support vector machines (SVM) to evaluate the Apache servers. And using the proposed method they achieved an accuracy of 94%.

Madhusmita Rout, Amandeep Kaur et al. in this paper [14] tried different data mining techniques to achieve a proposed predictive model to compare the accuracy between different machine learning algorithms. And they found that the logistic regression accuracy score was 82.35% and that was the highest score compared to other classifiers i.e., K-nearest neighbor (K-NN), Naive Bayes, Decision tree, and support vector machine (SVM).

Our proposal fills in an interesting gap for diabetes disease. In this research, we focus on patient analyzes through which we can determine whether the patient has the disease or not. Most of the research is not focused precisely on diabetics to identify the disease early, so we have submitted this paper to benefit patients, doctors, hospitals, and others to know the disease early. So, we noticed that there are some comparisons between different machine learning algorithms. However, there is no fair judgment to know which machine learning algorithm is a better fit for the dataset mining of diabetes disease.

## 2. METHODOLOGY

The main objectives of this paper research are to enhance the quality of the prediction of the diabetes disease using the diabetes dataset and to make the prediction of the disease using a high accuracy algorithm. In this study, we will try to achieve a fair judgment and know the best machine learning algorithm that can be used for the database of diabetes. This research has moved through a few phases that are described in the framework of the experimentation and that is shown in Fig.1. This framework has many stages: the collecting of the data, preprocessing and the selection method, transformation stage, selection of the mining tool for the dataset, programming language selection, and the stage of selecting the data mining tool algorithms, and that depends on the dataset binary classification or multi-classification.

This paper research uses different data mining techniques to analyze and evaluate the classification algorithms for the impact factors of the dataset of diabetes. Though we used R studio environment data mining techniques, we generated different predictive models for the classification process of the diabetes of the dataset, and then we evaluated the performance and accuracy through different techniques. These predictive modeling in health care could help patients and that's through the discovery of the diseases in early stages, another benefit that could be considered is the consistency of their long-term goals (cheap, better health, better choices). The progressive experience (and other auto insurers) is instructive, progressive (through the use of predictive modeling) finding patients who were a high-health risk classified (assigned to a high-risk pool) and could help in taking the necessary measures in advance in order to lessen the expected diabetes.

The experimental design explains the reasons why such a phenomenon occurs by making experiments where independent variables are manipulated, extraneous variables are controlled and therefore the conclusions are being drawn and that then leads to some other actions that the decision-maker should practice. Then the optimization as a technique suggests balancing the level of a certain variable that is related to other variables, thus identifying the ideal level of it, is the recommendation for the decision-maker.

### 2.1 The Proposed Model

During the four phases in the processing part in the conceptual model, this paper's research starts with gathering the necessary data, preprocessing of data, and filtering to clean, integrate, and then transform the data. And the required fields then are selected by programming language. The data then was transformed into a certain file format, that format is acceptable by the data mining tools. The data mining tools were the different data mining algorithms that are based on the datasets binary class or multi-class which are tested by Python and R programming language.

### 2.2 Data Gathering and Selection Phas

In this experimental work, this dataset has been collected from the National Institute of Diabetes and Digestive and Kidney diseases, which we obtain from Kaggle. The main goal of the dataset is to predict whether the patient has diabetes or not, and that through certain measurements of diagnostic in the dataset. The dataset contains various medical predictor variables and one target variable, outcome. A predictor variable includes some of the features such as BMI of the patient, insulin level, number of pregnancies the patient has had, age, and so on. However, there are several constraints were placed on the selection of these features from a larger database, in specific, all patients here are females at least 21 years old of Pima Indian heritage.

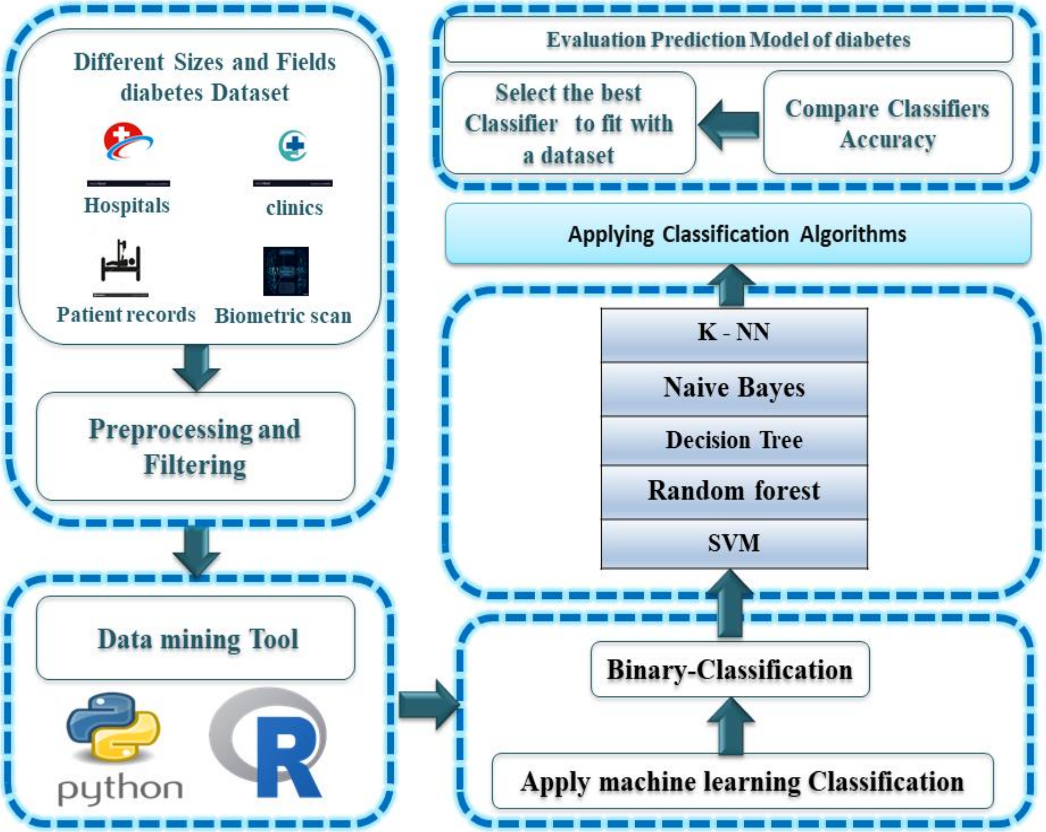

Different Sizes and Fields diabetes Dataset
Hospitals
clinics
Patient records
Biometric scan
Preprocessing and Filtering
Data mining Tool
python
R
Evaluation Prediction Model of diabetes
Select the best Classifier to fit with a dataset
Compare Classifiers Accuracy
Applying Classification Algorithms
K - NN
Naive Bayes
Decision Tree
Random forest
SVM
Binary-Classification
Apply machine learning Classification

**Fig. 1. The conceptual framework of the comparative study of diabetes dataset with different type classifiers**

The dataset consists of 768 instances and 8 attributes plus class. For Each Attribute: (all numeric-valued);

1. Number of times pregnant
2. Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/ (height in m) ^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

- Missing Attribute Values: Yes
- Class Distribution: (class value 1 is interpreted as "tested positive fordiabetes")
- Class Value Number of instances0: 5001: 268

## 2.3 Data Preprocessing and Transformation Phase

In this step, the dataset will be changed through different stages, such as data cleaning, data filtering, data integration, and data transformation. The gathered data was saved as a text document or excel spreadsheet. The cleaning process is needed to eliminate the missing values of the data and can run the analysis of the data based on the selected algorithms, or to correct the inconsistent data, identifying outliers, removing the duplicating data. The data was exemplified by numbers and stored in txt or CSV form files, so it can be used by the data mining tool. The dataset splits into two parts: the first part is a training set, and it was 80% of the dataset, and the second part was the remaining 20% to be used as a testing set for testing the trained model.

## 2.4 Selection of Data Mining Tool

There is no specific software to be used for this analysis. There are different tools that are used by different enterprises use different tools for data analysis. However, the decision of which tool to be used is depending on the type of data that is needed to be analyzed. The choice of the tools is also affected by the quality of the data which may have a significant impact on your analysis of the dataset[15].

R is a software open-source language that uses data modeling, statistics, prediction, time analysis, handling, and data visualization. The R language uses the computer RAM and depending on the RAM the R language can handle the data, the more the RAM the more the data R can work with. For the requirement, we have more than 4000 different packages created by various scholars. The R version that is used is the latest as R.3.0.0. R is not the best language to be used for large data analysis and that's because of the memory limit problems. However, some libraries such as R, Rodbc, rmr2, and Rhdfs are available to handle large data[16]. The R language is well developed for statistics, data mining algorithms, and analysis, and can be used for healthcare, credit risk scoring, CRM, and most predictive analytics. However, due to the deluge of data that must be analyzed and processed today, many organizations have some delegation about deploying R beyond research into health care applications [17].

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems. There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming. Python community has developed many modules to help programmers implement machine learning

## 2.5 Machine Learning Algorithms

In this section, we split the algorithms of classification for the dataset samples into target classes. The classification techniques to predict the target class for each data point. The classification data approach is a supervised learning approach that must have known class categories[18]. Data is divided into training and testing datasets. Then using the training dataset, researchers trained the classifier. The accuracy of the classifier must be tested; researchers have tested it using the test dataset. Classification is one of the most used methods in healthcare organizations. However, the accuracy of these methods is different and that's because of the classification algorithm that is used in data analytics. Determining the best classification algorithm between all available is a challenging task. This research suggests a comprehensive analysis, to test different classification algorithms and evaluate them by applying the diabetes dataset. We used different classification methods such as support vector machine (SVM), ensemble approach, and decision tree to analyze diabetes data [19].

Data mining techniques can be separated by their different functionality, standard preference, representation, and algorithms. The main function of a model and a short overview of the classification algorithms that have been used in this research.

A) The decision tree technique is one of the most common and efficient in the data mining field. This technique has been well explored and determined by many other researchers. However, some decision tree techniques may produce a huge tree structure in size and complexity in terms of understanding, moreover, the miss classification of the data usually happens in the learning process phase. Therefore, a decision tree that can produce a high accuracy simple tree structure is a requirement to work with a large dataset. Pruning methods are methods to reduce the complexity of the tree structure without affecting the accuracy of the classification[20].

B) The naïve Bayes technique is a simple classification algorithm that is multiclass and with the assumption of individuality between every pair of features. It can be used to train the dataset in a professional way, within one pass of the training data, and the uses for prediction[20].

C) Random forest technique, this technique is a set of the decision tree. It’s one of the most successful machine learning for regression and classification. It runs well on a huge dataset, but it is comparatively slower than other algorithms, it can handle missing values and estimate them, therefore, it is suitable for a large database that needs handling of missing values[20].

D) Support vector machine or (SVM) for short, is a supervised machine learning technique that can be used for either regression or classification challenges. However, it’s particularly used in classification problems. In the SVM algorithm, we first visualize all items of the data as points in n-dimensional space (where n is the number of the features in the dataset) with the value of all features being the value of a specific coordinate.

E) The K-nearest neighbors (K-NN) algorithm is a type of supervised machine learning algorithms. K-NN is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all the data for training while classifying a new data point or instance. K-NN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data.

## 2.6 Evaluation Metrics in ML/AI for Classification Problems

During the dealing with the missing data where different data mining was used, the researchers choose some measurement parameters to compare between the different algorithms. In this section, we present an overview of the chosen measurement parameters to illustrate what they represent, and later the evaluation and its result will be presented in section 4.8.the results and its evaluation will be shown.

While there are different types of classification algorithms, the evaluation process for the classification models almost all share similar principles. In a supervised classification problem, the model predicts an output for each data point as model-generated predicted output and there is also a true output. For this reason, the results of each algorithm for each point in the dataset can be assigned to one of four categories:

True Positive (TP) - the label is positive, and prediction is also positive.
True Negative (TN) - the label is negative, and prediction is also negative.
False Positive (FP) - the label is negative, but the prediction is positive.
False Negative (FN) - the label is positive, but the prediction is negative.

These four numbers are the building blocks for most classifier evaluation metrics. A fundamental point when considering classifier evaluation is that pure accuracy (i.e. was the prediction correct or incorrect) is not generally a good metric[21]. The reason for this is because a dataset may be highly unbalanced. Metrics like precision and recall are typically used because they consider the type of error. In most applications, there is some desired balance between precision and recall, which can be captured by combining the two into a single metric, called the F-measure.

### 2.6.1 Precision

Precision is used to determine how well the proposed algorithm matches the truth ground. Some researchers use recall and precision. Precision is also known as (PPV) or Positive predictive Value [22]. Precision is defined by an equation that is illustrated in 4.2 and the measures of the number of true positives relative to the sum of the true positives and the false positives. Precision is also a fraction of detected items that are correct. The lower the result of the precision the greater the value of the precision.

$$\boldsymbol{Precision = PPV = \frac{TP}{TP + FP}} \qquad (4.1)$$

### 2.6.2 Recall

The recall is also a measure that is used to quantify how the suggested algorithm matches the truth. Recall or equality or sensitivity True Positive Rate (TPR) is defined in equation 4.3 and it measures the number of true positives relatives to the sum of the false negatives and true positives [22]. The recall is the fraction of items that were correctly detected among all the items that were the scope of the test. The value 100 is the ideal percentage of recall.

$$\boldsymbol{Recall = TPR = \frac{TP}{P} = \frac{TP}{TP+FN}} \qquad \textbf{(4.2)}$$

### 2.6.3 F-Measure

F-measure or F1-score of sensitivity, specificity (i.e., harmonic mean) is defined by the equation of 4.3 and it measures overall how well we have been able to identify the ground truth foregrounds and backgrounds. The value of 100 is the ideal percentage value of the f1-score of sensitivity and specificity:

$$\boldsymbol{F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall}} \qquad \textbf{(4.3)}$$

### 2.6.4 Accuracy

Accuracy or percentage correct classification as it is called. This is a statical measure describe as the ratio of correctly classified instances and that it is equal to the sum of TN and TP to the total number of instances. Upon reaching 100% accuracy, it means that the value obtained from the proposed algorithm is exactly the same as the value of the ground truth[23]. 100% is the ideal value of accuracy and is defined by equation 4.4.

Accuracy (ACC) is the percentage of the total number of predictions that was correct. And it is determined by using the following equation:

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}} \qquad (4.4)$$

### 2.6.5 Training time

The machine learning algorithms are to be measured as an accomplished. The time required for the model to be built is called training time. This time varies on the implementations of the algorithms. From the previously mentioned parameters of measurement, the researcher will compare the accuracy that is provided by the applied algorithms on the dataset. Here, the focus will be on comparing the accuracy and training time parameters to choose which algorithm is better and suitable for a selected type of dataset.

## 2.7 Experimental Result and Analysis

According to the previously mentioned techniques in the previous section, these classification techniques are used and then applied to the diabetes dataset in a specific order so that a comparison study can be done. Analysis can rely on different standards, however the most necessary is the accuracy and time for the model to be done. Once the deployment is done sorts of dataset in terms of the training time, accuracy, false-positive rate (FPR), true positive rate (TPF), precision, recall, and F-measure, this analysis can help to decide the best-studied machine learning algorithm model is best suited to a specific of data. These studies were operated on a Lenovo Laptop using Linux operating system has the following specifications, the number of core processors 1 with Intel® Core ™ i7-5500U CPU, 2.40 GHz Processor, and 16 GB RAM.

## 2.8 Classification Results Using Decision Tree by Method C5.0

First, in order to measure the model performance of an algorithm, several Binary-classification evaluation metrics, the dataset was split into two parts, training, and testing sets, the training set was 80% of the whole dataset to train the dataset, and the testing set is the remaining 20% and it was used as cross-validation to be used in the training model as testing set for the model to be trained. Table 1 present the values of accuracy, F-measure, recall, precision, true positive, and time taken to build a model per second, and they are classified as the following dataset. This shows that the decision tree had the highest value of accuracy at 0.7581699% and the training time for the same algorithms was the best at 0.3408051 (seconds).

### 2.9 Classification Results Using Naïve Bayes algorithm

In the diabetes dataset that is used in this research, as shown in table 1, to measure the model performance of specific algorithms by several Binary-classification evaluation metrics. Table 1 presents the value of the accuracy, F-measure, Recall, true positive precision, and time taken to build model per second, which are classified to the following dataset. The result shows that the Naïve Bayes algorithm had the highest value of accuracy at 0.7662338 % and the training time for the same algorithms was the best at 0.1878929 (seconds).

### 2.10 Classification Results Using Random Forest algorithm

In this study results, the random forest classification algorithms result, first, in order to measure the model performance we split the data set into training and testing sets where the training was 80% of the entire dataset and the remaining 20% were for testing to measure the performance of the algorithms by several binary class classification evaluation and to know the accuracy and the required time to build the model per second which is defined by equations by means of the confusion matrix.

Table 1 represents the values of evaluation parameters for the random forest algorithm, and these parameters are TP, FP, TN, FN, precision, recall, F-measure, accuracy, and time needed to build the model. The result shows that Random first reached the peak for the accuracy at 0.7313% and peak for the time was at 0.45174 (seconds)

### 2.11 Classification Results Using SVM algorithm

In the diabetes dataset, the result of using several Binary classification evaluation metrics to measure the performance of the algorithm. Table 1 shows the different parameters for evaluation that is used for testing the SVM algorithm, and these parameters were Accuracy, F-measure, recall, precision, FN, TN, FP, and TP. the results show that the SVM algorithm highest accuracy score was at 0.7662338%, and the best value of time was at 0.1878929 (seconds).

Table 1 presents the value of all the parameters that are used in the evaluation, the table shows that the SVM algorithm had the highest value of precision at 0.6667, SVM had the highest value of recall at 0.5556 respectively, and SVM had the highest value of F-measure at 0.6061.

### 2.12 Classification Results Using K-NN algorithm by Method "Minkowski"

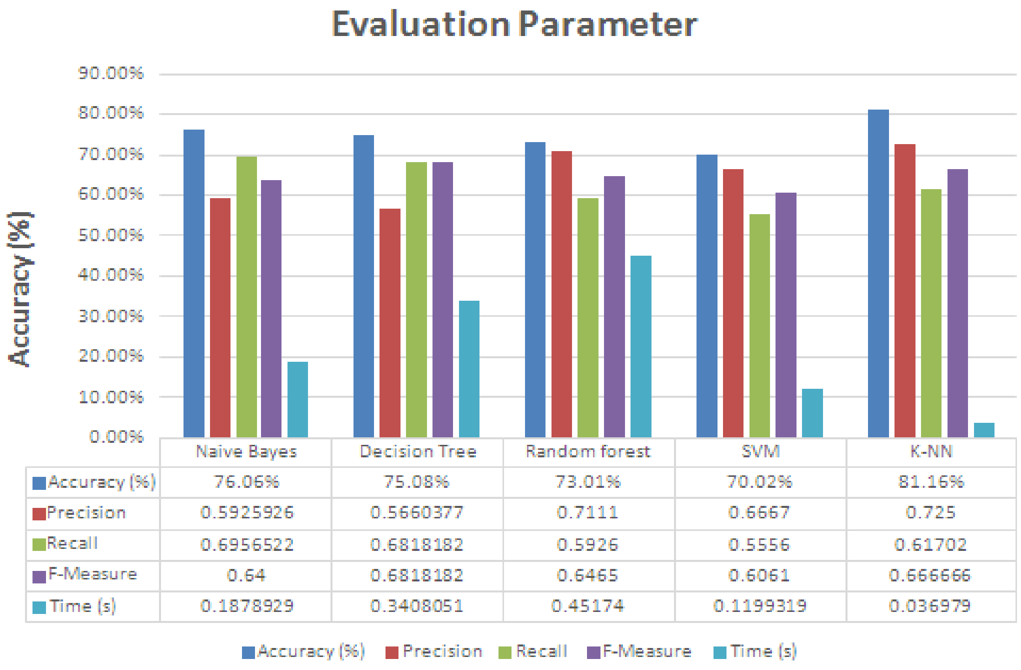
In the diabetes dataset, as shown in table 1 the performance of the algorithm was recorded by using different Binary-classification evaluation metrics. Table 1 illustrates the values of accuracy, F-measure, recall, precision, FN, TN, FP, TP, and time needed to build the model. The result of the K-NN algorithm reached the peak of the accuracy at 81.168831%, and the peak for the time 0.036979 (seconds).

### 2.13 Classification Results and Performance Evaluation

In the diabetes dataset, in Table 1 a comparison between four different classification algorithms, which are K-NN, naive Bayes, decision tree, random forest, and SVM, to evaluate the performance of each algorithm and doing so by using several Binary-classification evaluation metrics and regression metrics. Table 1 presents the values of accuracy, TP, FP, TN, FN, precision, recall, F-measure, and time taken to build the model. Fig.2 shows that the K-NN algorithm had the highest value of accuracy 81.16%, while the SVM algorithm had the lowest value of accuracy at 70.02%.

**Table 1. Comparison of five classifiers using Diabetes dataset**

| Classifier | TP | FP | TN | FN | |
|---|---|---|---|---|---|
| Naive Bayes | 32 | 22 | 86 | 14 | |
| Decision Tree | 0.1960784 | 0.1503268 | 0.5620915 | 0.09150327 | |
| Random forest | 0.5926 | 0.8700 | 0.7111 | 0.7982 | |
| SVM | 0.5556 | 0.8500 | 0.6667 | 0.7798 | |
| K-NN | 29 | 11 | 96 | 18 | |
| **Classifier** | **Accuracy (%)** | **Precision** | **Recall** | **F-Measure** | **Time (s)** |
| Naive Bayes | 76.06% | 0.5925926 | 0.6956522 | 0.64 | 0.1878929 |
| Decision Tree | 75.08% | 0.5660377 | 0.6818182 | 0.6818182 | 0.3408051 |
| Random forest | 73.01% | 0.7111 | 0.5926 | 0.6465 | 0.45174 |
| SVM | 70.02% | 0.6667 | 0.5556 | 0.6061 | 0.1199319 |
| K-NN | 81.16% | 0.725 | 0.61702 | 0.666666 | 0.036979 |



| | Naive Bayes | Decision Tree | Random forest | SVM | K-NN |
|---|---|---|---|---|---|
| 🟦Accuracy (%) | 76.06% | 75.08% | 73.01% | 70.02% | 81.16% |
| 🟥Precision | 0.5925926 | 0.5660377 | 0.7111 | 0.6667 | 0.725 |
| 🟩Recall | 0.6956522 | 0.6818182 | 0.5926 | 0.5556 | 0.61702 |
| 🟪F-Measure | 0.64 | 0.6818182 | 0.6465 | 0.6061 | 0.666666 |
| 🟦Time (s) | 0.1878929 | 0.3408051 | 0.45174 | 0.1199319 | 0.036979 |

**Fig. 2. Accuracy of dataset using R Language data mining tool algorithm**

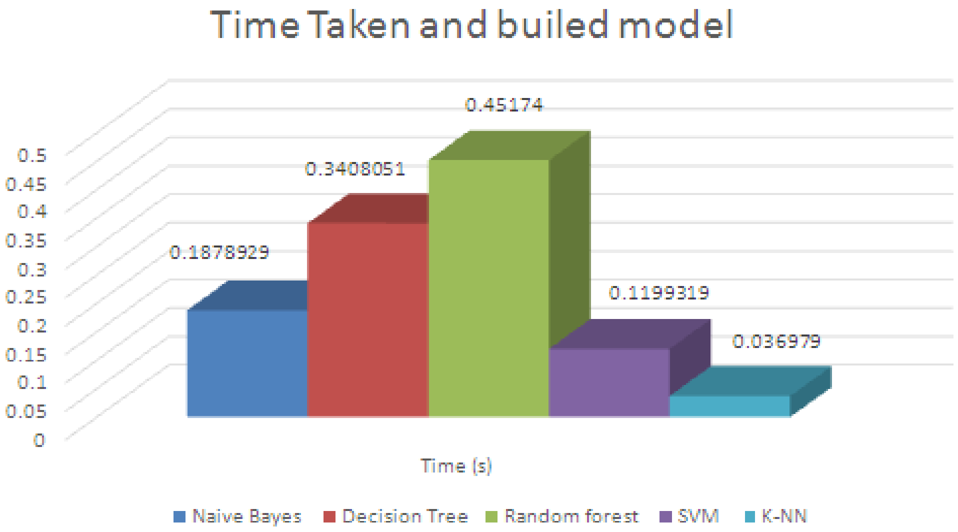


**Fig. 3. Time analysis for dataset one**

Finally, the dataset used data mining to achieve the best accuracy while using different classification algorithms. The results show that the best classification algorithm that fit in with our dataset was K-NN where the accuracy peak was at 81.168831% and the time peak was at 0.036979 (seconds).

## 3. CONCLUSION AND FUTURE WORK

This paper investigated diabetes datasets using different machine learning algorithms on the data sets using “Python and R Language”. This comparison has led to a fair judgment to determine the best fit for each field. Different algorithms were applied to the diabetes dataset. The result suggests that K-NN was the most suitable algorithm, as its highest accuracy score was 81.168831% and time was 0.036979 seconds. The experiment was implemented locally on specific data sets. Therefore, future work needs to implement more classification algorithms, neural network algorithms that automate the parameters to obtain the best parameters that fit the data set. On the other hand, we will use a larger data set from a different source.

## COMPETING INTERESTS

Authors have declared that no competing interests exist.

## REFERENCES

1. Ahuja R, Sharma SC, Ali M. A diabetic disease prediction model based on classification algorithms. Annals of Emerging Technologies in Computing (AETiC), Print ISSN, 2019;2516-0281.
2. Hafez MM, Shehab ME, El Fakharany E. Effective selection of machine learning algorithms for big data analytics using apache spark. in International Conference on Advanced Intelligent Systems and Informatics;2016. Springer.
3. Fiarni C, Sipayung EM, Maemunah S. Analysis and prediction of diabetes complication disease using data mining algorithm. Procedia Computer Science, 2019;161:449-457.
4. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Computer Science. 2018;132:1578-1585.
5. Anitha J, Pethalakshmi DA. Comparison of classification algorithms in diabetic dataset. International Journal of Information Technology (IJIT)–Volume. 2017;3.
6. Sarkar D, Bali R, Sharma T. Practical machine learning with Python. A Problem-Solvers Guide To Building Real-World Intelligent Systems. Berkely: Apress;2018.
7. Emoto T, et al. Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease. Heart and vessels, 2017;32(1):39-46.
8. Sharma G. Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes. Int. J. Advanced Networking and Applications.2020;12(01): 4509-4518.
9. Hafez MM, Redondo RPD, Vilas AF. A Comparative Performance Study of Naïve and Ensemble Algorithms for E-commerce. in 2018 14th International Computer Engineering Conference (ICENCO);2018. IEEE.
10. Orabi KM, Kamal YM, Rabah TM. Early predictive system for diabetes mellitus disease. in Industrial Conference on Data Mining;2016. Springer.
11. Pradhan M, Sahu RK. Predict the onset of diabetes disease using Artificial Neural Network (ANN). International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004), 2011;2(2):303-311.
12. Kumar DA, Govindasamy R. Performance and evaluation of classification data mining techniques in diabetes. International Journal of Computer Science and Information Technologies, 2015;6(2):1312-1319.
13. Alkaragole MLZ, Kurnaz A. Comparison of data mining techniques for predicting diabetes or prediabetes by risk factors;2019.
14. Woldemichael FG, Menaria S. Prediction of diabetes using data mining techniques. in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI);2018. IEEE.
15. Ratner, B., Statistical and Machine-Learning Data Mining:: Techniques for Better Predictive Modeling and Analysis of Big Data2017: CRC Press.
16. Prajapati V. Big data analytics with R and Hadoop2013: Packt Publishing Ltd.
17. Nisbet R, Elder J, Miner G. Handbook of statistical analysis and data mining applications2009: Academic Press.
18. Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 2013;5(5):241-266.
19. Hu H, et al. A comparative study of classification methods for microarray data analysis. in Proceedings of the 5th Australasian Data Mining Conference (AusDM 2006): Data Mining and Analytics 2006;2006. ACS Press.

20. Maimon O, Rokach L. Data mining and knowledge discovery handbook; 2005.
21. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation, in AI 2006: Advances in Artificial Intelligence, Springer. 2006;1015-1021.
22. Minnen D, et al. Performance metrics and evaluation issues for continuous activity recognition. Performance Metrics for Intelligent Systems, 2006;4.
23. Nguyen, T.T. and G. Armitage, A survey of techniques for internet traffic classification using machine learning. Communications Surveys & Tutorials, IEEE, 2008;10(4):56-76.

*Peer-review history:*
*The peer review history for this paper can be accessed here:*
*https://www.sdiarticle4.com/review-history/76525*