https://doi.org/10.36785/BUITEMS.JAES.418

Machine Learning-based Web Application for Early Diagnosis of Diabetes

[a]Farhad Hassan, [a]Maryam Wardah, [b]Muhammad Yasir, [c]Hamayoun Shahwani, [c]Syed Attique Shah, [c]Mohammad Imran, [c]Muhammad Ashraf , [c]Muhammad Qasim, [c]Muhammad Akram , [c]Zahid Rauf

a: Department of Computer Science, Air University Multan Campus, Multan, Pakistan
b: Department of Computer Science, University of Engineering and Technology Lahore, Faisalabad Campus, Pakistan
c: Faculty of Information and Communication Technology, Balochistan University of IT, Engineering and Management Sciences, Quetta, Pakistan

Abstract-- **Diabetes has become a chronic disease that seriously threatens human health. It is a group of metabolic diseases characterized by hyperglycemia and there is no role of the age factor involved. The long-term of diabetes disease causes chronic damage and dysfunction of various tissues, especially the eyes, kidneys, heart, blood vessels, and nerves. Most of the time people are not sure about this common disease at the early stage and unluckily the patient moves to a critical situation to meet with major disease due to the continuous effect of diabetes. This research is conducted to build the machine learning-based web application platform for the early diagnosis of the disease, freely accessible anywhere anytime. We used the benchmark dataset named PIDD (Prima Indian Diabetes Dataset) and performed the comparative analysis among the Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forest and Support Vector Machines. Based on the classification performance, we found that SVM performed the best among the pool of mentioned algorithms and, therefore, adopted for the development of the intelligent web application for the diabetes diagnosis.**

Keywords- **Classification, Support Vector Machine, Diabetes diagnosis, Diabetes prediction**

Date Received 15-10-2020
Date Accepted 24-10-2020
Date Published 18-12-2020

I. INTRODUCTION

Diabetes is characterized by a high blood sugar level which may result in diseases like heart attack, kidney failure, and stroke[1]. There are three types of diabetes: “Type-1 Diabetes”, “Type-2 Diabetes”, and “Gestational diabetes”. T1D is normally present in young adults who are under 30 years of age. It needs regular insulin injection because the pancreas of these patients does not produce insulin [2]. The signs of Type 1 include weight loss, polyuria, constant hunger, eyesight weakness, and tiredness [1].

In T1D the body kills the cells that are responsible for processing insulin to consume the sugar for energy production. And that sort of diabetes will contribute to obesity. Obesity is a rise in the body mass index (BMI) over an individual's usual BMI value [3]. Type 2 diabetes is a condition where cells fail to respond to insulin properly and its most common cause is lack of exercise and obesity and it happens in people who are above 45 age and they suffer from indications like obesity, dyslipidemia, arteriosclerosis, asthma, etc. [1] The third type occurs in pregnant women who have no previous history of diabetes but they develop high blood sugar levels and this type is known as “Gestational diabetes” [2]. In 2019, according to “IDF Diabetes Atlas Ninth edition 2019” approximately 463 million adults have diabetes and by 2045 this figure will rise to 700 million. Diabetes has left 4.2 million people dead [4]. Diabetes is increasing in adults and children and as a result death rate is also increasing. [1]. Data analysis is difficult because the data is complex and non-linear [4]. Health care infrastructure requires vast quantities of medical records where latent trends may be derived from data mining. It also helps to define the association in clinical evidence between the various trends. Diabetes is a severe health condition in which sugar content intake cannot be regulated. Irrespective of the fact that the primary cause is the accumulation of sugar, different variables such as height, weight, genetic factor, and insulin often bear a prominent role to influence diabetes. The early detection and resolution of these issues help to identify and keep away [6][7][8]. In the field of medical healthcare machines, learning-based systems are dominating. Different Machine Learning techniques were used to support medical experts to use various data mining algorithms. The efficacy of

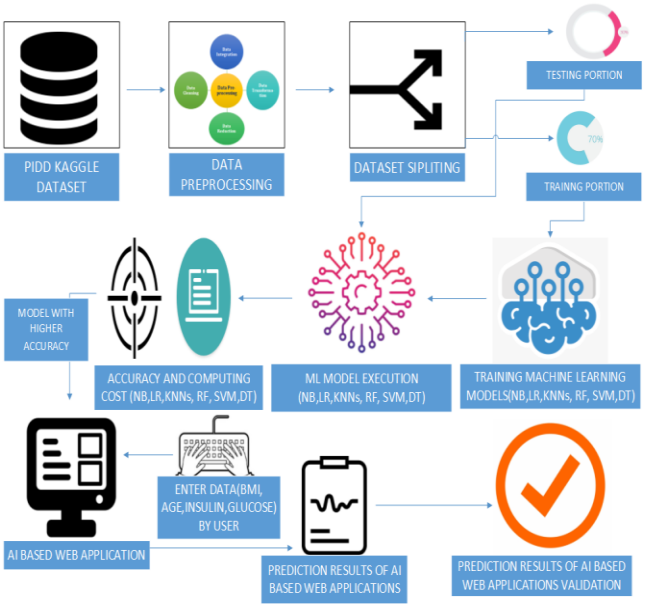
the decision support device is recognized by its precision. So the key aim of developing a decision support framework is to anticipate and diagnose a particular condition with a greater degree of accuracy [9][10][11]. Feature selection [12] and classifiers are used as machine learning models. It helps in better diagnosis of diabetes by using only a few attributes. In building, this web application dataset we used is "PIDD (Pima Indian Diabetes Data Set)" which consists of 769 records.

The rest of the paper is structured as under. In section II, a detailed literature review is given. Section III, explains the methodology. The results are discussed in Section IV. Finally, the paper is concluded in section V.

II. Literature Review

Sarwar and Sharma proposed different algorithms for pattern recognition [13] from given information and decision making techniques. Artificial Intelligence (AI) is an emerging technology nowadays in every field of life that includes robotics, industries, medical field, business, etc. In the medical field, its benefits are seeing in detecting brain tumors, diagnosis of cancer, lung diseases, heart diseases, etc. The major aim of their paper is to introduce different algorithms that will be helpful in medical fields in predicting and diagnosis diseases. They chose diabetes for research purposes. They choose 10 parameters that were the backbone for all prediction algorithms they worked on. Authors implemented 3 algorithms by keeping in mind these training data. "Naive Bayes algorithms", neural "artificial networks (ANN)", and K-nearest (KNN)" and developed Layout projections. The findings are to measure efficiency. The obtained system was compared with the actual diagnosis of the medical record. ANN has an accuracy of 96%, Naive Bayes has an accuracy of 95% and KNN has the least accuracy of 91% [14]. Whereas Han and his co-authors presented a machine-learning algorithm SVM which was rule-based extraction of features, Random Forest for prediction of diabetes. The proposed system gives an accuracy of 94.2% [15]. In the same way, Shetty and Joshi also gave a tool for diabetes prediction using data mining techniques. The idea they gave was to build a device that uses information mining techniques to predict diabetes. They also planned to find a new example that has valuable data that helps the clients to predict their diabetic state. They used the ID3 algorithm [16] for this purpose. From the dataset, a tree was created to show the working of the model. Results show that the error rate was only 6%, accuracy was 94%, specificity 22%, and affectability was 55% [17]. Ahmed also used different algorithms for the prediction of Type-2 Diabetes only. He used different data mining techniques to build a model based on the medical records of the patients. He used three models that are Naïve Bayes, Logistic, and J48 [18]. He used the WEKA application for this purpose. Logistic gave an accuracy of 74%, exactness was 0.73, a review was 0.744, F proportions of 0.653. Naïve Bayes's accuracy was 74%, exactness was 0.717, F proportions was 0.653 and review was 0.742. J48 accuracy was 73.5%, exactness was 0.54, F proportions were 0.623 and review was 0.735. It shows strategic calculations were more important than precision. The restriction was that only Type 2 Diabetes was under consideration [19]. Younus et al proposed an algorithm that is based on random forest and attempted to detect the complicated areas of patients with type 2 diabetes. In this paper, the authors studied that people suffer from chronic diseases due to their lifestyle, food intake, and reduced physical activity. Diabetes is one of the most common chronic illnesses that people of all ages suffer from. Complex and very heterogeneous data were collected from different resources. The solution for this is to convert this complex and meaningless data into useful data. The purpose of this paper is to identify the occurrence of diabetes in patients with type-2 diabetes mellitus [20] concerning long-term complications. They identify a higher percentage where the HbA1c level is higher than 7 and the BMI value is higher than 20. This model, then, can be recommended for researching medical data for controlling HbA1c and BMI, which could further serve to improve participant's knowledge and awareness. Further work is needed to make low-cost tools which would be cost-effectiveness. Improving the present condition in the future will be more advantageous [21]. Maniruzzaman, Rahman, Ahammed, and Abedin described that diabetes is a disease that is caused when the sugar level of blood increases. It can cause many diseases such as kidney failure, stroke, heart attack, etc. In 2014 422 million people suffered from diabetes around the globe. In 2040, the figure will reach 642million. The key goal of this learning is to build a system based on ML to predict diabetic patients. "Logistic regression (LR)" is used for knowing the risk aspects for diabetes based on P-value. 4 classifiers are used for predicting diabetes "Decision tree", "Naive Bayes", "Random forest" and "AdaBoost". They have used the National Health and Nutrition Diabetes dataset, conducted in the 2009–2012 survey of the examination. The sample is comprised of 6561 respondents with 657 diabetic controls and 5904 controls. The LR model shows that the risk factors for diabetes are 7. ML-based system's average is 90%. The combination of LR and RF- gives K10 protocol 94% ACC and 0.95 AUC. LR and RF performance-enhanced when combined [1]. Karatsiolis and Schizas also done prediction and diagnosis of diabetes are done by using different clustering and classification algorithms. PIMA dataset is used for the training of SVM [22]. Vijayan and Anjali suggest that better accuracy for cancer and diabetes can be achieved by using the "Adaptive Neuro-Fuzzy Inference system". Vijayan and Anjali also show that Naïve Byes and K-means achieved 80% accuracy by using these methods [8].

Figure 1. Research Methodology



III. Methodology

The adopted methodology for this research work as shown the Figure 1. In which PIDD dataset is utilized that it is consists of 769 records with 9 different attributes and the performance of machine learning algorithms depends on the data. In preprocessing of data, ensure the dataset should be in CSV format by the operation of CSV file conversion, and to deal with the missing values with adopts the NAN values value-based approach of the respective feature. The dataset is further divides into two categorized by the ration of 70/30 training data and testing data. Trains the model and also applied different (NB, LR, KNNs, RF, SVM, DT) machine learning algorithms and observed the best one among these achieved higher accuracy as compared to others and use it as a model for the in the AI-based web application which is developed by the Flask. For the appropriate interaction with AI-based web application the user needs to provide the information about glucose, BMI, age, and insulin with the help of the best model predicts either the user is patient or not.

A. Native Bayes Algorithm

It is based on the Bayesian system [24] and is used when the number of inputs is too big. It is mostly used in mathematics and statistical fields. The fundamental concept in the NB approach is that any aspect of a class is irrelevant to some other function of that class. This approach reaches strong precision while the underlying statement isn't accurate. A Naive Bayesian model can be effectively built without difficulty and has a parametric calculation that ultimately provides usefulness for broad datasets.

B. Logistic Regression Algorithm

It is a supervised learning algorithm [25], based on one or more predictors binary response is estimated. It uses probability logit function to check the relationship between response and predictors.

C. K Nearest Neighbors Classification

It is a non-parametric technique [26] and it stores all the present cases and calculates the new cases based on the distance measures [27]. It is a kind of instance establishing a learning model and its results may contain a member's group.

D. Decision Tree Algorithm

The indecision analysis decision tree classifier works fit. It is a flowchart tree-like structure that has nodes, root, and leaves. It is a supervised learning algorithm and a classification tree can be constructed when the response variable is categorical and can be used as a regression tree [28] when the response variable is continuous. The input may be of any type. It is a tree-like structure based on the input features. It is a type of system that has only conditional control [29].

E. Random Forest Algorithm

It is a machine learning algorithm that constructs a decision tree. This algorithm was given by "Breiman". Regression and Classification techniques can be used in Biomedical science and diabetes prediction. It can give the estimates of variables that which variables are important for our processing. It efficiently works on larger data.

F. Support Vector Machine Algorithm

SVM is a supervised learning algorithm that gives high accuracy with the least computation power. Classification and regression techniques can be used. Its major task is to find a hyperplane in an N-dimensional area that classifies the data points. It is used for mapping large data into high dimensional space. The aim is to find a plane with the maximum margin, that is, the maximum distance between the data points of both classes. Maximizing the gap from the margins gives some clarification such that potential data points can be identified with better trust.

IV. Results

Some interesting results were extracted with the implementation of several machine learning algorithms (NB, LR, KNNs, RF, SVM, DT) for detailed comparative analysis deals with the accuracy of the machine learning algorithm and the computing cost of classification while for the curial decision of the best machine learning algorithm selection among the pools of above-mentioned algorithms considered the accuracy factor because such type of AI-based prediction systems [30] is mainly focused on accuracy but in a real-time medical system, the importance of computing cost of classification has been a great impact.

A. Accuracy of a Machine Learning Algorithm

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) innovations have been developing rapidly, and predictable results continue to evolve as accessibility increases while the accuracy of the model having the major impact and

importance for the advancement of this field. For the measurement of the accuracy of a machine-learning algorithm uses the Eq. 1.

$$ACC = ((TP+TN)/ (TVT)) \quad (1)$$

Here, TP and TN are total correctly classified positive and negative values while TVT indicates the total values of a particular confusion matrix.

B. Classification of Computing Cost of a Machine Learning Algorithm

Machine learning can be a powerful analysis tool for processing large amounts of data while the computing cost of the classification of every algorithm has a variation in values and it can be calculated by the Eq. no. 2.

$$CCA = TVT [X-(X-Y) * ACC] \quad (2)$$

Here TVT is the total values of the model's confusion matrix and X, Y both values are obtained from the cost matrix where Y is the value of POSITIVE|POSITIVE and NEGATIVE|NEGATIVE while the X is the value of POSITIVE| NEGATIVE and NEGATIVE| POSITIVE.

C. Naïve Bayes Algorithm

The Naive Bayes Algorithm is the simplest classification technique that is based on the Bayesian system and this model originated from classical mathematical theory and has stable classification efficiency. It also performs well on small-scale data, can handle multi-classification tasks, and is suitable for incremental training, especially when the amount of data exceeds the memory, we can perform incremental training batches. The fundamental concept in the NB approach is that any aspect of a class is irrelevant to some other function of that class. This approach reaches strong precision while the underlying statement isn't accurate. A Naive Bayesian model can be effectively built without difficulty and has a parametric calculation that ultimately provides usefulness for broad datasets. It's mathematically represented as shown in Eq. no. 3.

$$P(c|x) = P (x|C) P(C) / p (x) \quad (3)$$

- P(c|x) = probability of class c of given predictor x
- P(c) = probalility of class
- P(x|c) = probabilty of predictor given class
- P(x) = probabltiy of predictor

1) Naïve Bayes Algorithm Accurancy

The algorithm accuracy depends on the agreement between the assumed probability distribution and the real data degree. The confusion matrix is a table that is used to envision the performance of the algorithm and each row represents the actual category and each column represents the predicted value as shown in Table 1.

Table 1. Naïve Bayes Algorithm's Confusion Matrix

	Positive	Negative
Positive	TP = 262	FN = 40
Negative	FP = 70	TN = 128

Uses the above-mentioned Eq. no. 1 to calculate the accuracy of the Naive Bayes algorithm by putting the values TP and TN from Table 1. While TVT indicates the sum of all values of the matrix as Table no. 1.

Accuracy of Naive Bayes algorithm = ((262+128)/(500))

Accuracy of Naive Bayes algorithm = 0.78

Multiply by 100 to get the value of accuracy in percentage so,

Accuracy of Naive Bayes algorithm in percentage = 78.0%.

2) Naïve Bayes Algorithm Computing Cost

The calculation of the computing cost of classification uses the cost matrix and confusion matrix of a Naïve Bayes machine learning algorithm while the cost matrix in machine learning is similar to the confusion matrix, except that in cost matrix major concerns with incorrect or correct predictions.

Table 2. Cost matrix for the Naive Bayes Algorithm.

	Positive	Negative
Positive	390	110
Negative	110	390

Both X and Y values are obtained from table 2 where Y is the value of POSITIVE|POSITIVE and NEGATIVE|NEGATIVE while the X is the value of POSITIVE| NEGATIVE and NEGATIVE| POSITIVE. So, X = 103 and Y=397 and TVT is a total value of confusion matrix table 1 which is 500, and ACC is the achieved accuracy of the model.

Put these values into the above-mentioned Eq. no. 2.

= TVT [X– (X-Y)* ACC]

= 500[110-(110-390)*0.78]

=500[110-(-280)*0.78]

=500[110+280*0.78]

=500[110+218.40]

=500[328.40]

=164200

Hence, computing cost of Naive Bayes algorithm =164200

D. Logistic Regression Algorithm

Logistic Regression is a machine learning method that is used to solve two classifications (0 or 1) problems and is used to estimate the possibility of something just deals with the binary response. The calculation sum is very small when listed, the speed is very high, and the storage resources are low but if the space of the function is that, the logistic regression output is not very good. It uses probability logit function to check the relationship between response and predictors and also requires the dependent variable to be a discrete variable while the variable used for response is Y and d X indicates the linear predictor. Its formulas are mentioned in equation no. (4), (5) and (6) respectively.

logit(p) = b0 + b1X1 + b2X2 + b3X3 +…..bkXk (4)

odds = p/1-p = probabiity of presence of characteristics / probabiity of
absence of characteristics (5)

logit(Pj) = =loge ((Pj/1-pj) = $\sum^{K}_{i=0}$ βiXi (6)

Pj is a probability for a diabetic that is the value of Y will be equal to 1 and 1-Pj is used for non-diabetic that is the value of Y will be 0. βi 's are unknown regression constants while I is equal to 0,1,2,…..K. The total number of predictors is K and Xi's are predictors where X0=1. Unknown coefficients can be estimated by using "Maximum Likelihood Estimator". We can easily select the features whose p values are less than 0.05.

1) Logistic Regression Algorithm Accurancy

The accuracy of the logistic regression algorithm examines the level correctness in the prediction's result and it can be obtained by the confusion matrix that visualizes the performance of the algorithm while each column represents the predicted value while each row represents the actual category as shown in the below Table 3.

Table 3. Logistic Regression Algorithm's Confusion matrix

	Positive	Negative
Positive	TP = 107	FN = 73
Negative	FP = 30	TN = 290

Uses the above-mentioned Eq. no. 1 to calculate the accuracy of the Logistic Regression Algorithm by utilizing the values TP and TN of Table 3. While TVT indicates the sum of all values of the Confusion matrix.

Accuracy of Logistic Regression algorithm = ((107+290)/ (500))

Accuracy of Logistic Regression algorithm = 0.794

Multiply by 100 to get the value of accuracy in percentage

So,

Accuracy of Logistic Regression algorithm in percentage = 79.40%.

2) Logistic Regression Algorithm Computing Cost

Calculating the computational cost of classification for the Logistic Regression machine learning algorithm uses the cost matrix and confusion matrix, while the cost matrix and confusion matrix are nearly similar but the cost matrix concerns with inaccurate or accurate predictions.

Table 4. Cost matrix for the Logistic Regression Algorithm.

	Positive	Negative
Positive	397	103
Negative	103	397

Both X and Y values are obtained from table 4 where Y is the value of POSITIVE|POSITIVE and NEGATIVE|NEGATIVE while the X is the value of POSITIVE| NEGATIVE and NEGATIVE| POSITIVE.

So, X = 103 and Y=397 and TVT is a total value of confusion matrix table 3 which is 500, and ACC is the achieved accuracy of the model.

Put these values into the above-mentioned Eq. no. 2.

= TVT [X– (X-Y)* ACC]

= 500[103-(103-397)*0.794]

=500[103-(-294)*0.794]

=500[103+294*0.794]

=500[103+233.436]

=500[336.436]

=168218

Computing cost of Logistic Regression algorithm =168218

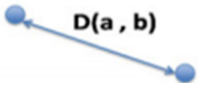
E. K Nearest Neighbors Classification

The principle working of kNN is very simple. By calculating the distance between the sample to be classified and the sample of the known category, it finds the K samples of known categories that are closest to the sample to be classified and then counts the K samples according to the minority "subject to the majority" decision principle. The number of occurrences of various types of samples, the sample with the most occurrences in the category of the sample to be classified. It is a non-parametric technique and it stores all the present cases and calculates the new cases based on the distance measures [27]. It is a kind of instance establishing a learning model and its results may contain a member's group. Then the group is selected based on data that is if K=1 then it has the closest nearest neighbor and if K=2 then the class has a double nearest neighbor and so on. Different distance functions are given in Eq.no. 7,8, and 9:

Euclidean $\sqrt{\sum_{i=1}^{k}(xi-y)2}$ (7)

Manhattan $\sum_{i=1}^{k}|\,xi-yi\,|$ (8)

Minkowski $(\sum_{i=1}^{k}(|\,xi=yi\,|)q)1/q$ (9)



If we choose a smaller value of k, it will mean that our overall model will become complicated and prone to over fitting.

1) K Nearest Neighbors Algorithm Accuracy

To measure the accuracy of K Nearest Neighbors machine learning algorithm utilizes the confusion matrix which is a table and uses to visualize the performance of the algorithm and each column represents the predicted value while each row represents the actual category as shown in the below Table 5.

Table 5. K Nearest Neighbors Algorithm's Confusion matrix.

	Positive	Negative
Positive	TP = 167	FN = 69
Negative	FP = 35	TN = 229

From the above-mentioned Eq. no.1 calculates the accuracy of the K Nearest Neighbors algorithm by utilizing the valuesTP and TN of Table 5. While TVT indicates the sum of all values of the above-mentioned matrix.

Accuracy of K Nearest Neighbors algorithm = ((167+229)/(500))

Accuracy of K Nearest Neighbors algorithm = 0.792

Multiply by 100 to get the value of accuracy in percentage so,

Accuracy of K Nearest Neighbors in percentage = 79.20%.

2) . K Nearest Neighbors Algorithm Computing Cost

The K Nearest Neighbors algorithm's computing cost of classification obtained by the cost matrix and confusion matrix and cost matrix also recognized by the Cost of misclassifying that contains the correctly and incorrectly prediction values.

Table 6. Cost matrix for the K Nearest Neighbors Algorithm.

	Positive	Negative
Positive	396	104
Negative	104	396

Both X and Y values are obtained from table 6 where Y is the value of POSITIVE|POSITIVE and NEGATIVE|NEGATIVE while the X is the value of POSITIVE| NEGATIVE and NEGATIVE| POSITIVE.So, X = 103 and Y=397 and TVT is a total value of confusion matrix table 5 which is 500, and ACC is the achieved accuracy of the model.

Put these values into the above-mentioned Eq. no. 2.

= TVT [X– (X-Y)* ACC]

= 500[104-(104-396)*0.792]

=500[104-(-292)*0.792]

=500[104+294*0.794]

=500[104+231.264]

=500[335.264]

=167632

Computing cost of K Nearest Neighbors algorithm =167632

F. Accuracy of a Machine Learning Algorithm

Several decision trees are consisting of random forests, and there is no connection between specific trees. New input samples are entered as we perform classification tasks, and each decision tree in the forest is independently evaluated and graded. Every decision tree gets the product of its classification. That of the decision tree's classification results is graded At most, this outcome will be viewed by the random forest as the end result. This algorithm was given by "Breiman". Regression and Classification techniques can be used. In Biomedical science and diabetes prediction, it can be used. It can give the estimates of variables that which variables are important for our processing. It efficiently works on larger data. Its generic form is shown in Figure 2.

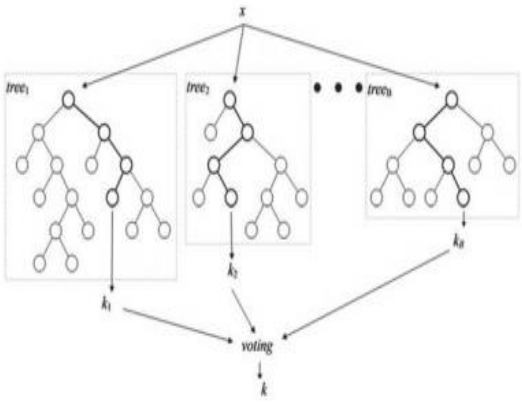


Figure. 2 Random Forest

1) Random Forest Algorithm Accuracy

The assessment of the accuracy of the Random Forest algorithm uses the confusion matrix approach that expresses the performance of the algorithm while each column represents the predicted value while each row represents the actual category as shown in the below Table 7.

Table 7. Random Forest Algorithm's Confusion matrix.

	Positive	Negative
Positive	TP = 196	FN = 38
Negative	FP = 26	TN = 240

From the above-mentioned Eq. no.1 calculates the accuracy of the K Nearest Neighbors algorithm by utilizing the valuesTP and TN of Table 7. While TVT indicates the sum of all values of the Confusion matrix for the Random Forest Algorithm Table 7.

Accuracy of Random Forest algorithm = ((196+240)/(500))

Accuracy of the Random Forest algorithm = 0.872

Multiply by 100 to get the value of accuracy in percentage so,

Accuracy of Random Forest in percentage = 87.20%.

2) Random Forest Algorithm Computing Cost

The classification computing cost for the Random Forest algorithm extracted by the cost matrix and total value of confusion matrix, while in the cost matrix major concerns with appropriate and erroneous predictions.

Table 8. Cost matrix for the Random Forest Algorithm.

	Positive	Negative
Positive	436	64
Negative	64	436

Both X and Y values are obtained from table 8 where Y is the value of POSITIVE|POSITIVE and NEGATIVE|NEGATIVE while the X is the value of POSITIVE| NEGATIVE and NEGATIVE| POSITIVE.So, X = 103 and Y=397 and TVT is a total value of confusion matrix table 7 which is 500, and ACC is the achieved accuracy of the model.

Put these values into the above-mentioned Eq. no. 2.

= TVT [X– (X-Y)* ACC]

= 500[64-(64-436)*0.872]

=500[64-(-372)*0.872]

=500[64+372*0.794]

=500[64+324.384]

=500[388.384]

=194192

Computing cost of Random Forest algorithm =194192.

G. Support Vector Machine Algorithm

Support Vector Classifier (SVM) is a supervised learning algorithm and it belongs to the category of classification. In the application of data mining, it corresponds to and distinguishes clustering [31] with unsupervised learning. It is widely used in machine learning, computer vision, and data mining and gives high accuracy with the least computation power. Its major task is to find a hyperplane in an N-dimensional area that classifies the data points. It is used for mapping large data into high dimensional space. The aim is to find a plane with the maximum margin, that is, the maximum distance between the data points of both classes. Maximizing the gap from the margins gives some clarification such that potential data points can be identified with better trust. In two dimensional areas, the plane separates an area or group, and each group either lies on one side or another side [32]. Possible Hyperplanes of support vector machine are shown in Figure 3.

$$c(x,y,f((x)) = \begin{cases} 0, if y * f(x) \geq 1 \\ 1 - y * f(x), else \end{cases} \quad (10)$$

Loss function for SVM:

$$\min_w \lambda \| w \| 2 + \sum_{i=1}^{n} (1 - yi(xi, w)) + \quad (11)$$

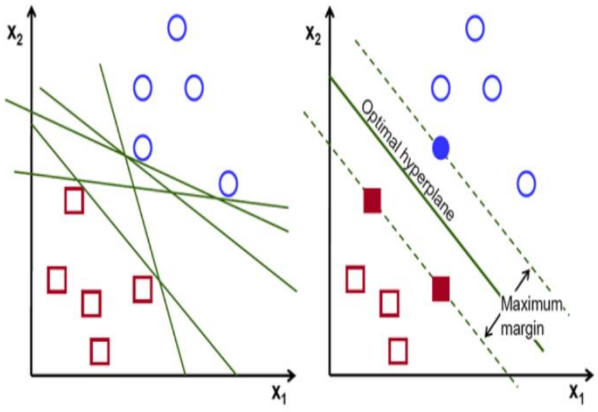


Figure 3. Possible hyper planes

The formulas used to maximize the margins and the loss function are shown in equation 10 which helps to optimize the margin is a loss of the hinge and the Loss function for SVM is given in equation 11.

1) Support Vector Classifier Algorithm Accuracy

Support Vector Machine Algorithm formally finds a hyper plane while ensuring the classification accuracy and from the accuracy determine the performance of the algorithm. Uses the confusion matrix for the calculation of confusion matrix that contains the rows and columns combination while each column indicates the predicted value but each row indicates the actual category as shown in Table 9.

Table 9. Support Vector Machine Algorithm's Confusion matrix.

	Positive	Negative
Positive	TP = 232	FN = 12
Negative	FP = 5	TN = 251

From Eq. no. 1 calculates the accuracy of the Support Vector Machine algorithm by utilizing the values TP and TN of Table 3. While TVT indicates the sum of all values of the matrix.

Accuracy of Support Vector Machine algorithm = ((232+251)/ (500))

Accuracy of Support Vector Machine algorithm = 0.966

Multiply by 100 to get the value of accuracy in percentage so,

Support Vector Machine Accuracy in percentage = 96.60%

2) Support Vector Classifier Computing Cost

The calculation of the computing cost of classification uses the cost matrix and confusion matrix of a Support

Vector Machine algorithm while the cost matrix is similar to the confusion matrix, except that in cost matrix major concerns with incorrect or correct predictions.

Table 10. Cost matrix for the Support Vector Machine algorithm.

	Positive	Negative
Positive	438	17
Negative	17	438

Both X and Y values are obtained from table 10 where Y is the value of POSITIVE|POSITIVE and NEGATIVE|NEGATIVE while the X is the value of POSITIVE| NEGATIVE and NEGATIVE| POSITIVE.So, X = 14 and Y=386 and TVT is a total value of confusion matrix table 9 which is 500, and ACC is the achieved accuracy of the SVM algorithm as mentioned above.

Put these values into the above-mentioned Eq. no. 2

= TVT [X– (X-Y)* ACC]

= 500[17-(17-438)*0.966]

=500[17-(-466)*0.966]

=500[17+466*0.966]

=500[17+450.156]

=500[467.156]

=233578

Computing cost of Support Vector Machine algorithm =233578.

H. Decision Tree Algorithm

The decision tree algorithm uses a tree structure and uses layers of reasoning to achieve the final classification. When predicting, a certain attribute value is used for judgment at the internal node of the tree, and the branch node to enter according to the judgment result is determined until the leaf node is reached, and the classification result is obtained. This is a supervised learning algorithm based on if-then-else rules. These rules of the decision tree are obtained through training instead of the manual formulation. A decision tree is the simplest machine learning algorithm. It is easy to implement, highly interpretable, fully in line with human intuitive thinking, and has a wide range of applications and a classification tree can be constructed when the response variable is categorical and can be used as a regression tree when the response variable is continuous. The input may be of any type. It is a tree-like structure based on the input features. It is a type of system that has only conditional control [29].

1) Decision Tree Algorithm Accuracy

The decision tree algorithm is based on the known probability of occurrence of various situations while the accuracy this algorithm is obtained by the confusion matrix is a table that contains rows and columns and each column represents the predicted value while each row represents the actual category as shown in the below Table 11.

Table 11. Decision Tree Algorithm's Confusion matrix.

	Positive	Negative
Positive	TP = 147	FN = 33
Negative	FP = 67	TN = 253

From the above-mentioned Eq. no. 1 calculates the accuracy of the Decision Tree Algorithm by utilizing the values TP and TN of Table 5. While TVT indicates the sum of all values of the matrix.

Accuracy of Decision Tree Algorithm = ((147+253)/ (500))

Accuracy of Decision Tree Algorithm = 0.80

Multiply by 100 to get the value of accuracy in percentage so,

Decision Tree Algorithm Accuracy in percentage = 80.0%.

2) Decision Tree Algorithm Computing Cost

Uses cost matrix and confusion matrix to calculate the computing cost of a Decision Tree machine learning algorithm.

Table 12. Cost matrix for the Decision Tree Machine algorithm.

	Positive	Negative
Positive	400	100
Negative	100	400

Both X and Y values are obtained from table 12 where Y is the value of POSITIVE|POSITIVE and NEGATIVE|NEGATIVE while the X is the value of POSITIVE| NEGATIVE and NEGATIVE| POSITIVE.So, X = 100 and Y=400 and TVT is a total value of confusion matrix table 5 which is 500, and ACC is the achieved accuracy of the Decision Tree Algorithm as mentioned above.

Put these values into the above-mentioned Eq. no. 2.

= TVT [X– (X-Y)* ACC]

= 500[100-(100-400)*0.80]

=500[100-(-300)*0.80]

=500[100+300*0.80]

=500[100+240]

=500[340]

=170000

Computing cost of Support Vector Machine algorithm =170000.

I. Comparative Analysis of Machine Learning Algorithms

The experimental results show that the classification accuracy of NB is 78% with 164200 computing cost of classification, LR

is 79.4% with 168218 computing cost of classification, KNNs is 79.2% with 167632 computing cost of classifications, RF is 87.2% with 194192 computing cost of classification, SVM 96.6% with 233578 computing cost of classifications, and DT is 80% with 170000 computing cost of classification as shown in Figure 4. There are many machine learning algorithms for prediction of the diabetic state that has been chosen in this study as mentioned above but for the AI-based Web application, the selection of machine learning algorithm depends on the parameter of higher accuracy because it promotes the use of accurate AI-based web application for the prediction of Diabetes that is very important for individual users in the field of healthcare. In this study, we found the classification accuracy of the constructed Support Vector Machine prediction model is 96.6% which is higher as compared to other machine learning algorithms and ignores computing cost of classification while in a real-time medical system, the importance of computing cost of classification has been a great impact as compared to the accuracy parameter but it doesn't mean the accuracy is negligence able in those systems.

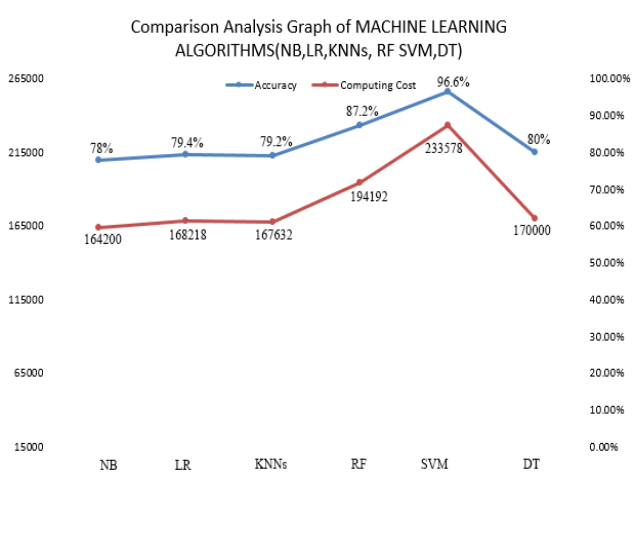


Figure 4. Comparative Analysis of Machine Learning Algorithms

J. AI-based Web Application for Diabetics Predictions

With the recent technological leap, AI has quickly become the mainstream technology of online systems, and designers can apply it to the web application to get the prediction at runtime by giving input to a few parameters. The AI-based web application is designed by using the Flask which is a very flexible and popular python based web framework. The GUI interface of AI-based web application takes values of GLUCOSE, BMI, AGE, and INSULIN from the user as shown in Figure 5, and gets the prediction results according to entered data with the help of the SVM model.

V. Conclusion

This research is conducted for the prediction of diabetics disease through AI-based Web Application with the

Figure. 5. AI-based Web Application for Diabetics Predictions

analysis approach [33] of machine learning algorithms (NB, LR, KNNs, RF, SVM, DT) and we observed each algorithm secured different accuracy with the utilization of Pima Indians Diabetes Data Set. Diabetes is characterized by hyperglycemia [34] and if this most common disease not diagnosed at the early stage then the chance of chronic damage and dysfunction of various tissues, especially the eyes, kidneys, heart, blood vessels, and nerves are at a high level in near future. The results of the comparative analysis of Machine Learning Algorithms NB, LR, KNNs, DT, RF, and SVM are based on two factors that are accuracy and computing cost of a classification. Among all the algorithms compared, SVM achieved the highest accuracy of 96.6% with relatively high computing cost. Considering the significance of accuracy for the disease diagnostics, we choose SVM for the building of AI-based web application to predict the diabetic's status more accurately.

As a future work, we are intended to predict the likelihood of diabetes in the coming future given the current state of the user. We are also interested in suggesting the diet plans for the people to avoid the diabetic conditions. For future work, from the current state of the user, we can predict the likelihood of diabetes and suggest the best diet plan for the people to avoid due to diabetic condition.

VI. References

[1] M. Maniruzzaman, N. Kumar, M. M. Abedin, M. S. Islam, H. S. Suri, A. S. El-Baz, and J. S. Suri "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm", *Computer methods and programs in biomedicine*. 2017.

[2] "Diabetes", *Who.int*, 2020. [Online].

[3] A. H. Mohammad, T. Alwada'n, and O. Al-Momani,"Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network," *GSTF J. Comput.*, 2016.

[4] N. D. S. Report, "National Diabetes Statistics Report, 2020," *Natl. Diabetes Stat. Rep.*, 2020.

[5] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines,"

International Journal of Medical Informatics. 2008.

[6] M. Otoom, H. Alshraideh, H. M. Almasaeid, D. López-de-Ipiña, Bravo and J. Bravo "A Real-Time Insulin Injection System," In *International Workshop on Ambient Assisted Living*. 2013.

[7] V. V. Vijayan and C. Anjali, "Decision support systems for predicting diabetes mellitus-A Review," in *Global Conference on Communication Technologies, GCCT 2015*, 2015.

[8] S. Kumari and A. Singh, "A data mining approach for the diagnosis of diabetes mellitus," in *7th International Conference on Intelligent Systems and Control, ISCO 2013*, 2013.

[9] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm," in *2018 21st International Conference of Computer and Information Technology, ICCIT 2018*, 2019.

[10] 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 2015*, 2016.

[11] K. De Silva, D. Jönsson, and R. T. Demmer, "A combined strategy of feature selection and machine learning to identify predictors of prediabetes," *J. Am. Med. Informatics Assoc.*, 2020.

[12] F. Emmert-Streib and M. Dehmer, "A Machine Learning Perspective on Personalized Medicine: An Automized, Comprehensive Knowledge Base with Ontology for Pattern Recognition," *Mach. Learn. Knowl. Extr.*, 2018.

[13] S. Mukherjee and N. Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction," *Procedia Technol.*, 2012.

[14] C. Han *et al.*, "Subclinical hypothyroidism and type 2 diabetes: A systematic review and meta-analysis," *PLoS One*, 2015.

[15] E. E. Ogheneovo and P. A. Nlerum, "Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification and Predictive Analysis," *Int. J. Adv. Eng. Res. Sci.*, 2020.

[16] S. R. P. Shetty and S. Joshi, "A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique," *Int. J. Inf. Technol. Comput. Sci.*, 2016.

[17] J. H. J. C. Ortega, M. R. Resureccion, L. R. Q. Natividad, E. T. Bantug, A. C. Lagman, and S. R. Lopez, "An analysis of classification of breast cancer dataset using J48 algorithm," *Int. J. Adv. Trends Comput. Sci. Eng.*, 2020.

[18] T. M. Ahmed, "Using data mining to develop model for classifying diabetic patient control level based on historical medical records," *J. Theor. Appl. Inf. Technol.*, 2016.

[19] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Sci. Rep.*, 2020.

[20] M. Younus, M. T. A. Munna, M. M. Alam, S. M. Allayear, and S. J. F. Ara, "Prediction Model for Prevalence of Type-2 Diabetes Mellitus Complications Using Machine Learning Approach," *Data Management and Analysis*. 2020.

[21] S. Karatsiolis and C. N. Schizas, "Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset," in *IEEE 12th International Conference on BioInformatics and BioEngineering, BIBE 2012*, 2012.

[22] Y. Wang, "Iteration-based naive Bayes sentiment classification of microblog multimedia posts considering emoticon attributes," *Multimed. Tools Appl.*, 2020.

[23] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, 2020.

[24] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A new k-nearest neighbors classifier for big data based on efficient data pruning," *Mathematics*, 2020.

[25] M. Nirmaladevi, S. A. Alias Balamurugan, and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," in *2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, ICE-CCN 2013*, 2013.

[26] S. Shabani, H. R. Pourghasemi, and T. Blaschke, "Forest stand susceptibility mapping during harvesting using logistic regression and boosted regression tree machine learning models," *Glob. Ecol. Conserv.*, 2020.

[27] A. A. AlJarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *2011 International Conference on Innovations in Information Technology, IIT 2011*, 2011.

[28] A. Moraru, D. Costin, R. Moraru, and D. Branisteanu, "Artificial intelligence and deep learning in ophthalmology - present and future (Review)," *Exp. Ther. Med.*, 2020.

[29] M. Farhadian, P. Shokouhi, and P. Torkzaban, "A decision support system based on support vector machine for diagnosis of periodontal disease," *BMC Res. Notes*, 2020.

[30] T. Santhanam and M. S. Padmavathi, "Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," in *Procedia Computer Science*, 2015.

[31] K. J. Lo, S. S. Lin, C. W. Lu, C. H. Kuo, and C. Te Liu, "Whole-genome sequencing and comparative analysis of two plant-associated strains of Rhodopseudomonas palustris (PS3 and YSC3)," *Sci. Rep.*, 2018.

[32] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: Literature review," *Journal of Medical Internet Research*. 2018.