Diabetes Prediction using Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Logistic Regression Classifiers

S Peerbasha [a*], A Saleem Raja [b], Praveen K.P [a], Y. Mohammed Iqbal [a], Mohamed Surputheen [a]

[a]Department of Computer Science, Jamal Mohamed College, Affiliated to Bharathidasan University, Trichy-20, Tamil Nadu, India.

[b]Information Technology Department, University of Technology and Applied Sciences-Shinas, Sultanate of Oman

Abstract

One of the world's deadliest diseases is diabetes. It is an additional creator of different assortments of problems. Ex: Coronary disappointment, Visual impairment, Urinary organ illnesses, and so forth. In such cases, the patients are expected to visit a hospital to get a consultation with doctors and their reports. They must contribute their time and cash every time they visit the hospital. Yet, with the development of AI techniques, we have the adaptability to search out a response to the present problem. We have progressed an advanced framework for handling data that can figure regardless of whether the patient has polygenic sickness. In addition, being able to foresee the onset of the disease is crucial for patients. Data withdrawal has the adaptability to eliminate concealed information from an enormous amount of diabetes-related data. The most important outcomes of this research are the establishment of a theoretical framework that can reliably predict a patient's level of risk for developing diabetes. We have utilized the existing categorization methods such as DT (Decision Tree), RF (Random Forest), SVM (Support vector Machine), LR (Logistic Regression) as well as K-NN (K-Nearest Neighbors) for predicting the severity of Type-II Diabetes patients. We got an accuracy of 99% for the Random Forest, 98.40% for the Decision Tree, 78.54% for Logistic Regression, 77.94% for SVM (Using RBF Kernal SVM), and 77.64% for KNN.

Keywords: Logistic Regression, Data Mining, K-Nearest Neighbors, Decision Tree, Machine Learning, Random Forest, Support Vector Machine.

*Corresponding author Email: bashapeer2003@gmail.com

1. Introduction

Diabetes is what is occurring, which results in a lack of enough insulin levels in the blood. Cautioning indication of high blood sugar brings about incessant peeing, feeling parched, and expanded hunger. If it isn't cured, it will prompt numerous hardships. This trouble leads to death. Heart disease, foot sores, and hazy vision are all side effects of extreme stress. A rise in levels of blood sugar is a sign of early diabetes. The earlier diabetes is not in this manner incredible than the customary worth. Diabetes is caused by the exocrine organ working together but not assembling enough hypoglycemic specialists to respond to the hypoglycemic specialist created. Different data mining calculations present different choices of emotionally supportive networks for helping wellbeing-trained professionals. The adequacy of choice emotionally supportive network is perceived by its Precision. In this way, the objective is to fabricate a choice emotionally supportive network to foresee and determine a specific infection with an outrageous accuracy measure. Man-made intelligence includes Machine Learning (ML), a branch of it that addresses verifiable problems by giving workstations the capacity to learn without the benefit of program creation.

1.1 Types of Diabetes

1) Type 1 diabetes is the consequence of the pancreas' failure to produce sufficient hypoglycemic experts. This type is stated as "Insulin-subordinate polygenic illness mellitus" (IDDM) or"Adolescent diabetes". The description is unknown. Type-1 polygenic illness was discovered in children under the age of twenty. Individuals struggle throughout their lives as a result of the kind of diabetes they have and depend on their insulin intake.

2)Type-2 diabetes begins with hypoglycemia specialist instruction, a condition in which cells fail to respond to the hypoglycemic specialists efficiently. The condition develops as a result of the shortage of hypoglycemia experts that have also gathered. This type is stated as "non-insulin-subordinate polygenic illness mellitus". The standard reason is outrageous weight. By 2025, the total population affected by type 2 will be known. The presence of diabetes mellitus is consolidated by 3% in the country zone when contrasted with the metropolitan zone. The investigation discovered that an individual who Joined the Country's organization has a customary crucial sign.

3) Blood sugar levels increase in a pregnant woman with Type-3 gestational diabetes even though she has never had diabetes before. Hence, viewed as altogether 18 percent of pregnant ladies have diabetes. Thus, there is a risk of developing gestational diabetes during pregnancy. Being overweight is a significant contributor to getting type 2 diabetes. Type 2 polygenic infection is

taken care of by legitimate exercise and taking proper system. When more advanced methods fail to reduce the aldohexose level, medications are usually recommended. As per the polygenic disease static data, 29.1 million US citizens have diabetes.

2. Literature Review

Veena Vijayan and V. Anjali C has examined, the diabetes sickness created by ascent of sugar level in the plasma. Different modernized data frameworks were illustrated using classifiers for expecting and diagnosing diabetes utilizing decision tree, SVM, Naïve Bayes and ANN algorithms [1].

P. Suresh Kumar and V. Umatejaswi has introduced the calculations like Decision Tree, SVM, Naive Bayes for recognizing diabetes utilizing information mining methods [2].

RidamPal, Dr.JayantaPoray and Mainak Sen has introduced the Diabatic Retinopathy (DR) which is one of the driving reason for sight failure for diabetic patients. In which they explored the exhibition of a bunch of machine learning calculations and check their exhibition for a specific informational collection [3].

Dr. M. Renuka Devi and J. Maria Shyla a has examined about the investigation of different abilities of mining to figure diabetes utilizing Naive Bayes, Random Forest, Decision Tree furthermore, J48 calculations [5].

Rahul Joshi and Minyechil Alehegn has examined the ML strategies which are utilized to figure the datasets at an beginning stage to save the life. Utilizing KNN and Naive Bayes calculation [6].

ZhilbertTafa and Nerxhivane Pervetica has examined the consequence of calculations that are carried out to progress the determination dependability [7].

Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. has examined the investigation of AI Calculations for example, Support Vector Machine, Credulous Bayes, Decision Tree, PCA for Extraordinary Infection Expectation utilizing Head of Part Investigation [11].

3. Proposed System

The suggested framework is based on using the calculation mixtures shown in the below block graph. The base characterization calculations are DT, RF, SVM (RBF Kernel-based), KNN and Logistic Regression for accuracy authentication.

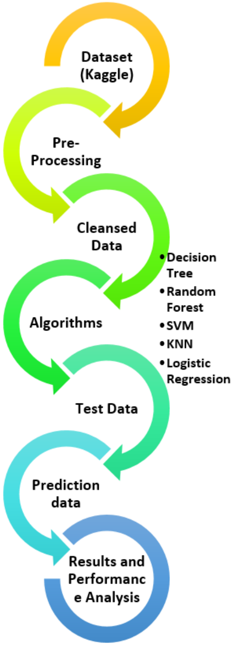


Fig. 1 Process Flow of Proposed System

4. Dataset

The phase of training is over. Two thousand cases and 9 features are included in the dataset. Features of the dataset include:

- Total number of times pregnant
- Glucose/sugar level
- Diastolic Blood Pressure
- Body Mass Index (BMI)
- Skin fold thickness in mm
- Insulin value in 2 hrs
- Hereditary factor- Pedigree function
- Patient’s age in years

Testing and training may be done using the percentage split option.Out of 2000 instances, 60% are utilized for testing, and 40% are for training. [1]

4.1 Training Data and Testing Data

Training data and testing data are two important sets of data in data analysis for machine learning. The machine learning paradigm is trained by the training data, and as it learns from the data, it begins to recognize patterns and correlations.The trained model's performance is assessed using the testing data, where the predictions made by the model are contrasted with the actual results in the test data. It is crucial to keep the training and test data separate to avoid overfitting, where the model performs well on the test data simply because it has already seen the data during training rather than because it has learned to make accurate predictions. The quality and representativeness of both the training and test data have a major effect on the accuracy and reliability of the machine learning paradigm's results.

4.2 Pre-processing

It refers to the alterations made to our information before sending it on to the classifiers. Information Pre-processing techniques are used to transform the unjustified informational collection into a justifiable informational collection. As a result, if the data is combined from several sources, it is done so in an unsuitable manner for inspection. Fig. 2 depicts the Data Pre-processing.

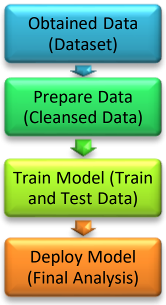

Obtained Data (Dataset)
Prepare Data (Cleansed Data)
Train Model (Train and Test Data)
Deploy Model (Final Analysis)

Fig. 2 Data Preprocessing

4.3 Cleansed Data

Cleansed data in machine learning analysis refers to the process of preparing and pre-processing the data for analysis. Any machine learning project must include this stage since the data's quality and structure have a significant influence on the model's accuracy and efficacy. Data cleaning can include tasks such as handling missing values, removing duplicate data, correcting inconsistent data, handling outliers, and transforming variables to suitable formats. For the machine learning model to give correct results, the data must be accurate, consistent, and fit for analysis. This is the aim of data cleaning.

4.4 Machine Learning Algorithms Used:

4.4.1 Decision Tree:

It's the broad, figure-displaying apparatus that has applications crossing various assorted zones. Decision trees, for the most part, are constructed as an algorithmic methodology that detects methods of dividing an informative collection in light of numerous conditions. It is among the managed learning strategies that are used the most often. The goal is to make a model that accurately predicts the value of an objective variable using unambiguous decision tree criteria that have been learned. It is appropriate for a reveal of the knowledge since it doesn't call for any boundary setup. The principles that the choice tree follows are, for the most part, else articulations. The decision tree performs arrangement without requiring much calculation. Choice trees are proficient in dealing with constant as well as all-out factors.

4.4.2 Random Forest:

It is a popular machine learning method that belongs to the ensemble learning family of algorithms. It is used for both classifications as well as regression problems. The term "Random Forest" refers to the way that several decision trees are combined in a Random Forest to create a forest of trees. In order to create each tree in the forest, a DT model is trained using a subset of the training data and characteristics that have been randomly chosen. The average of all the forest's trees' projections is used to determine the outcome. Random Forest has several advantages over single decision trees, such as improved accuracy, reduced overfitting, and improved interpretability. It also can handle a mix of continuous and categorical features and can handle missing values.

4.4.3 Support Vector Machine:

The SVM calculation is then shown, resulting in isolated zones of strength for the classes, which represent the events of importance in the region. The goal is to choose the edge hyperplane with the most severe edge that provides the greatest class separation. Help vectors are the events that are closest to the largest edge hyperplane. The vectors are selected on the basis of the section of the dataset that represents the preparation set. Support vectors of 2 classes empower the formation of two equal hyperplanes. Hence, the classifier's speculation error will be more favorable the wider the fringe between the two hyperplanes. SVMs are carried out extraordinarily as contrasted and other ML calculations.

4.4.4 K-Nearest Neighbours:

The fundamental principle behind KNN is that target variables are likely to be similar when data points are near to one another in the feature space. In other words, the algorithm assumes that alike observations are possible to have a similar class label or continuous target value. The value of "K" is a hyperparameter that needs to be set by the practitioner, and it determines the number of nearest neighbors that will be used to make the prediction. A common heuristic is to set "K" to an odd number to avoid ties. In summary, KNN is a simple, flexible, and effective algorithm that can be used for various types of problems, including classification, regression, and anomaly detection.

4.4.5 Logistic Regression

Based on one or more input features, it is a classification technique used in ML to forecast the likelihood of a binary result. It models the relationship between the input variables and the output by estimating the probabilities using a logistic function and then makes a prediction based on a threshold value. Logistic regression is a popular and widely used algorithm in various domains such as healthcare, finance, and marketing.

4.5. Algorithm Formulas

Precision: The number of TP divided by the "number of TP '+' number of FP" is defined as Precision. When a model incorrectly labels a result as positive when it should be negative, this is called a false positive.

Precision = TP/ (TP+FP)

Recall: The number of true TP divided by the "TP '+' FN" is defined as Recall.

Recall=TP/(TP + FP)

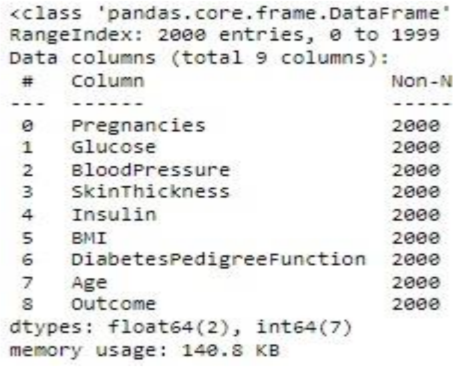
F1-Score: The function of Precision as well as Recall is F1. When there is uneven class distribution and a balance between Precision and Recall is required, there is a need for F1 Score.

F1= (2*Precision*Recall)/ (Precision*Recall)

5. Experimental Results and Discussion

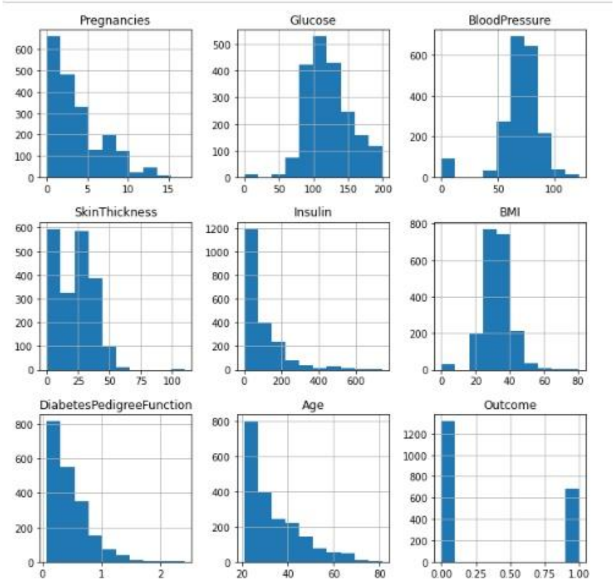
In this research contribution, experiments are performed on the original data set taken from the ***"National Institute of Diabetes and Digestive and Kidney Diseases***" which is available in Kaggle. The dependability of the suggested methodology has been assessed using the Precision, Recall, F1 Score, &Accuracy metrics that are often used in medical research. After obtaining the information dataset, the model will forecast the data using ML calculations and give the finest kind of correlation between predictions for the highest accuracy in treating diabetes.

The experimental configuration is done using Python Language and implemented in Jupyter Notebook with advanced machine learning packages. In this contribution, five different machine learning algorithms are used. The number of records used in this experiments are 2000. Totally, nine features are used. The results are shown in the figure … and table … The performance of the work are evaluated on the basis of accuracy, precision, recall and f1-score.



```
<class 'pandas.core.frame.DataFrame'
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                     Non-N
---  ------                     -----
 0   Pregnancies                2000
 1   Glucose                    2000
 2   BloodPressure              2000
 3   SkinThickness              2000
 4   Insulin                    2000
 5   BMI                        2000
 6   DiabetesPedigreeFunction   2000
 7   Age                        2000
 8   Outcome                    2000
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Table 1. Data Frames



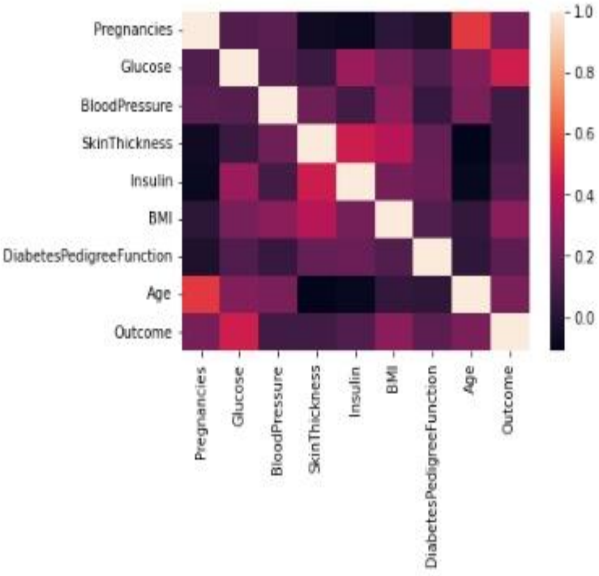


Fig. 3 Summarization of Data measured using Histograms Fig. 4 Correlation of variables using Heatmap

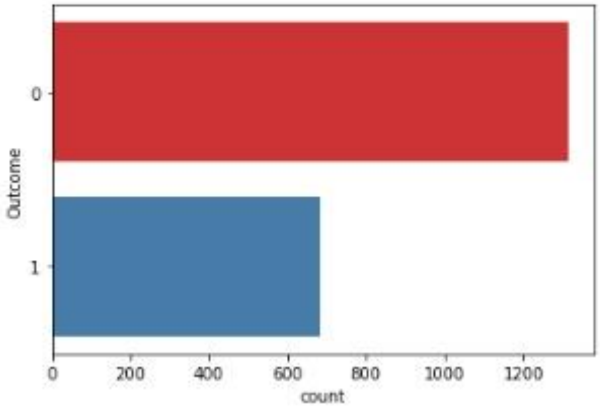


Fig. 5 Count of Observations using CountPlot

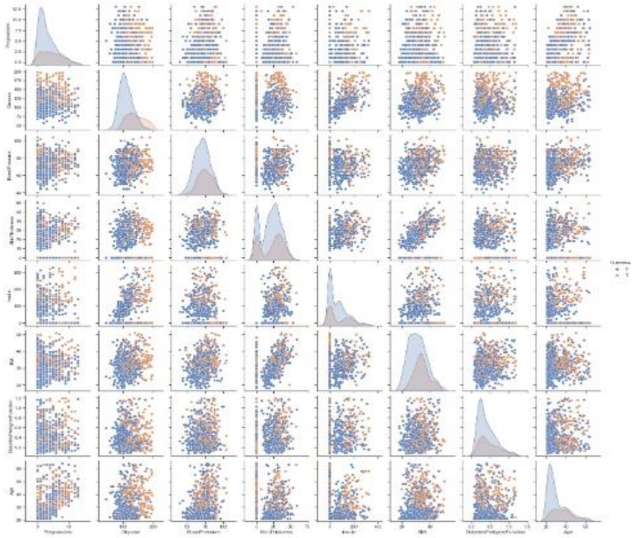
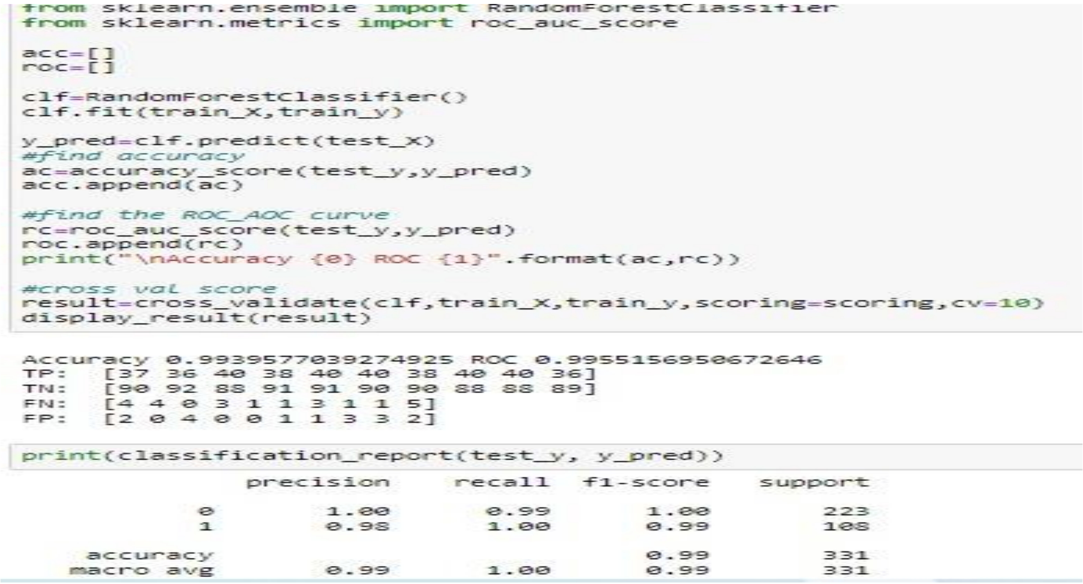


Fig. Count of Observations using PairPlot



```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score

acc=[]
roc=[]

clf=RandomForestClassifier()
clf.fit(train_X,train_y)

y_pred=clf.predict(test_X)
#find accuracy
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

#find the ROC_AOC curve
rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

#cross val score
result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.9939577039274925 ROC 0.9955156950672646
TP:   [37 36 40 38 40 40 38 40 40 36]
TN:   [90 92 88 91 91 90 90 88 88 89]
FN:   [4 4 0 3 1 1 3 1 1 5]
FP:   [2 0 4 0 0 1 1 3 3 2]
```

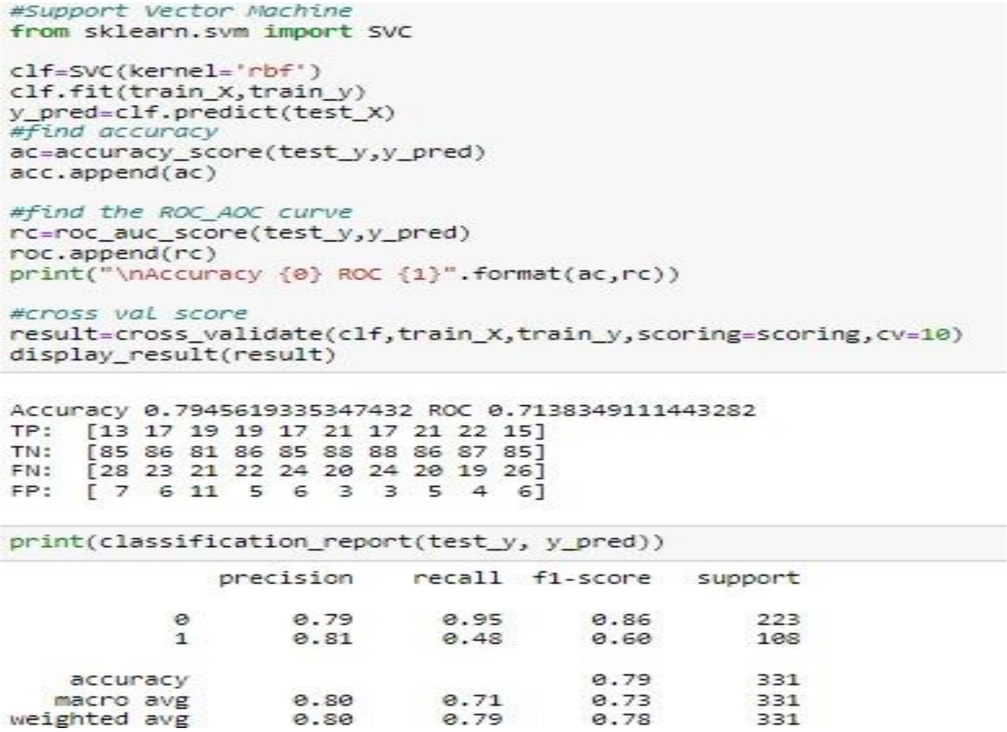
```
print(classification_report(test_y, y_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       223
           1       0.98      1.00      0.99       108

    accuracy                           0.99       331
   macro avg       0.99      1.00      0.99       331
```

Fig. Classification Report of RF (Random Forest Classifier)



```
#Support Vector Machine
from sklearn.svm import SVC

clf=SVC(kernel='rbf')
clf.fit(train_X,train_y)
y_pred=clf.predict(test_X)
#find accuracy
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

#find the ROC_AOC curve
rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

#cross val score
result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.7945619335347432 ROC 0.7138349111443282
TP:  [13 17 19 19 17 21 17 21 22 15]
TN:  [85 86 81 86 85 88 88 86 87 85]
FN:  [28 23 21 22 24 20 24 20 19 26]
FP:  [ 7  6 11  5  6  3  3  5  4  6]
```

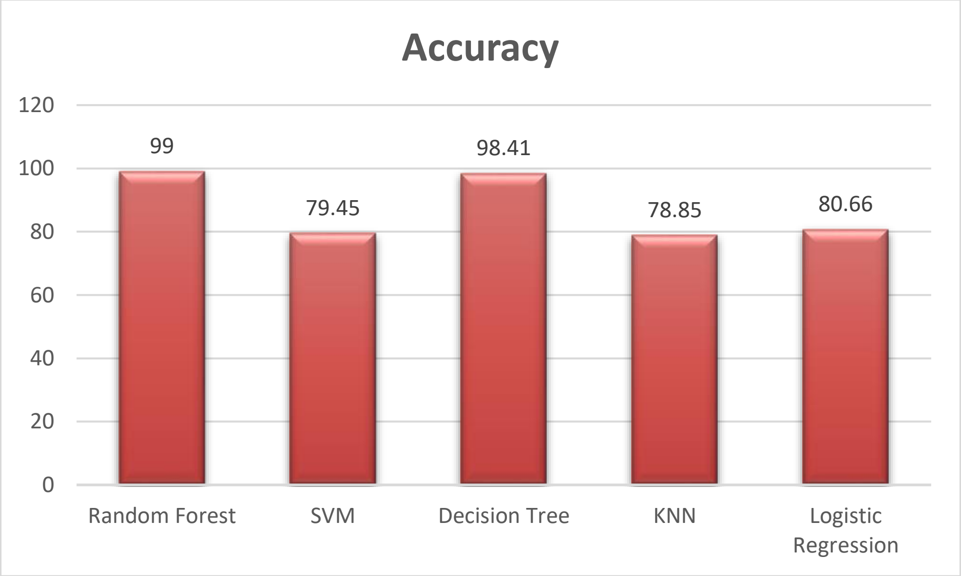
```
print(classification_report(test_y, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.79      0.95      0.86       223
           1       0.81      0.48      0.60       108

    accuracy                           0.79       331
   macro avg       0.80      0.71      0.73       331
weighted avg       0.80      0.79      0.78       331
```

Fig. Classification Report of SVM (Support Vector Machine Classifier)



6. Conclusion

Random Forest and Decision tree classifiers are generally excellent when we have no clue about the information. Indeed, even if it is unstructured along with semi-organized information such as text, and pictures, the Random Forest and SVM calculation functions admirably. The drawback of the Scale Vector Machine computation is that it is difficult to get the best order results for indeterminate issues.A few key boundaries are required to have been set accurately. Decision tree: It is not difficult to comprehend and control the choice tree. The decision tree exhibits instability because a significant change in the information structure of the ideal decision tree should be obvious to a modest modification. They are usually somewhat incorrect. KNN: It is effective at filling in the gaps by ignoring computations for likelihood assessments. The result drawn from using these techniques demonstrates that the RF classifier and DT classifiers perform better in this study effort in predicting diabetes when compared to RBF Kernel-based SVM, KNN, and Logistic Regression.

7. References

[1]. Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach",2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10- 12 December 2015 | Trivandrum. https://doi.org/10.1109/RAICS.2015.7488400

[2]. P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.

[3]. Ridam Pal, Dr. Jayanta Poray, and Mainak Sen, "Application of Machine Learning Algorithms on Diabetic Retinopathy", 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19-20, 2017, India.

[4]. BerinaAlic, LejlaGurbeta and AlmirBadnjevic, "Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases", 2017 6th Mediterranean Conference On Embedded Computing (MECO), 11-15 JUNE 2017, BAR, MONTENEGRO. https://doi.org/10.1109/MECO.2017.7977152

[5]. Dr. M. Renuka Devi and J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications. http://www.ripublication.com

[6]. Rahul Joshi and MinyechilAlehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm": Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017

[7]. ZhilbertTafa and NerxhivanPervetica, "An Intelligent System for Diabetes Prediction", 4th Mediterranean Conference on Embedded Computing MECO - 2015 Budva, Montenegro.

[8]. Sumi Alice Saji and Balachandran K, "Performance Analysis of Training Algorithms in Diabetes Prediction", International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India 2015.

[9]. Aakansha Rathore and Simran Chauhan, "Detecting and Predicting Diabetes Using Supervised Learning". International Journal of Advanced Research in Computer Science, Volume: 08, May June 2017.

[10]. April Morton, EmanMarzban and Ayush Patel, "Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients, 13th International Conference on Machine Learning and Applications",2014. https://doi.org/10.1109/ICMLA.2014.76

[11]. Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis". International Conference on Global Trends in Signal Processing, Information Computing and Communication 2016.