International Journal of Computer Sciences and Engineering
Vol.**11**, Issue.**7**, pp.**01-07**, July **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Comparative Performance Analysis of Datamining and Machine Learning Techniques for Diabetes Prediction

Vaishali Sarde[1*], Pankaj Sarde[2]

[1]Govt. J. Yoganandam Chhattisgarh College, Raipur (C.G.), India
[2]Rungta College of Engineering & Technology, Bhilai (CG), India

**Corresponding Author: vaishalimirge2018@gmail.com*

Received: 22/May/2023; **Accepted**: 24/Jun/2023; **Published**: 31/Jul/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i7.17

Abstract: Diabetes is caused by the high blood sugar. Body's main source of energy is glucose. Our body can produce glucose, but glucose also comes from the various foods we eat. One of the hormone called Insulin is generated by the pancreas to help glucose to move into the cells and to be used for energy later. If anyone is diabetic then body doesn't make sufficient, or any insulin, or doesn't usage insulin appropriately. Glucose then remains in the blood and not able to move to cells. Diabetes involves the risk of damage to the eyes, kidneys, nerves, and heart. Early prediction of diabetes can lower the risk of developing diabetes health problems. This paper uses five different techniques from data mining and machine learnings- KNN, Support Vector Machine, decision Tree, Naive Bayes and Artificial Neural Network for the prediction of diabetes. Comparative study based on the performance of these algorithms has been presented. The measures used for the performance analysis of all the five algorithms are Accuracy, Precision, Recall, f1-score and Support. For the experiment purpose the dataset is taken from Mendeley data[1] . It has records of 1000 patients. Result shows that decision tree achieved the best accuracy as compared to the other data mining and machine learning techniques.

Keywords: KNN, Support Vector Machine, decision Tree, Naive Bayes and Artificial Neural Network, Machine Learning

1. Introduction

Diabetes is a condition that controls how blood sugar (glucose) will be processed by the body. For the cells Glucose is an important source of energy. Diabetes can increase the sugar level in the blood which can increase the risk of several critical complications including heart disease, kindly failure and strokes etc. There are two main types of diabetes type 1 and type 2. According to WHO Globally [2], 422 million adults were having diabetes in 2014. 1.5 million deaths in 2012 were caused by Diabetes. An additional 2.2 million deaths were caused due to the higher than optimal blood glucose, which raises the risks of cardiovascular and other diseases. This paper covers the various classification methods from machine learning and data mining for the prediction of diabetes. The dataset is chosen from Mendeley data [1]. Records of 1000 patients are collected and covers three classes Diabetic, Non-Diabetic, and Predicted- Diabetic and 12 parameters. In this paper Predicted- Diabetic and Diabetic are considered as same class. So, two classes Diabetic and Non-Diabetic are used for experimental purpose. Five techniques KNN, decision Tree, Support Vector Machine, Naive Bayes and Artificial Neural Network have been applied to the dataset for the prediction of diabetes. Experiment's result shows decision tree performs well among others. Various measures Accuracy, Precision, Recall, f1-score and Support are taken for performance analysis.

2. Related Work

Debadri Dutta. et. al. [3] discovers the critical factors for the causes of diabetes. They focused in the most important features to predict what are the chances to develop diabetes in a person again in future.
Tao Zheng et. al. [4] set the goal to implement a semi-automated framework using machine learning as a pilot study to relax filtering criteria to improve recall rate with a keeping of low false positive rate.

Nahla Hosny Barakat et. al. [5] uses Support Vector Machines (SVMs) for the prediction of diabetes. They turns the "black box" model of an SVM into an intelligible representation of the SVM's classification decision. Experiments on real datasets proves that intelligible SVMs works as a promising tool for the prediction of diabetes. The rules extracted are clinically sound. Karim M. Orabi et. al. [9] developed a prediction system, by using randomization code in addition with regression technique to predict the person's age. Very promising results were found, the accuracy of the system was 84% to predicts diabetes incidents at what age.

Tarik A. Rashid et. al. [10] proposes a description model for diabetes and chronic disease prediction, comprises of two sub modules to verify the this relationship. In the first sub-Artificial Neural Network (ANN) is used to classify the types of case and to predict the rate of fasting blood sugar (FBS) of patients. The second sub-module shows the effect of the rate of FBS on the patient's health. For the description part of diabetes symptoms decision tree (DT) is used .

3. Dataset

3.1 Source of Dataset

The dataset is chosen from Mendeley data[1]. The data were collected from the Iraqi society, it has been gathered from the laboratory of Medical City Hospital and the Specializes Centre for Endocrinology and Diabetes-Al-Kindy Teaching Hospital. 1000 patient's records are used and cover three classes (Diabetic, Non-Diabetic, and Predicted- Diabetic). For the experimentation purpose Diabetic and Predicted- Diabetic classes are merged and treated as Diabetic class. Total two classes Diabetic and Non-Diabetic are considered in experiment.

3.2 Features used in Dataset

12 parameters given in the table 1 are used for the experimental purpose.

Table1: Features used in experiment

S. No.	Feature
1	Gender
2	AGE
3	Urea
4	Cr
5	HbA1c
6	Chol
7	TG
8	HDL
9	LDL
10	VLDL
11	BMI
12	class

Table 1 shows the various features used for the prediction of diabetes. Description of features is as follows-Gender, Age, Urea, Creatinine ratio(Cr), HBA1C, Cholesterol(chol) , Cholesterol (Chol), Triglycerides(TG), Fasting lipid profile, including total, LDL, VLDL, and HDL, Body Mass Index (BMI), Class (the patient's diabetes disease classes are Diabetic and Non-Diabetic).

3.3 Data Preprocessing

Data preprocessing is most important task for data mining and machine learning process. Data preprocessing improves the quality of data that can give better result. For Mendeley data[1] data preprocessing is performed in two steps.

1. **Converting text data into numerical form.**
   Some of feature values are in the textual form which has been converted into numerical form.
2. **Splitting dataset**
   Dataset is separated between training and testing data set. 67% of data is used for training purpose and remaining 33% of data is used for testing purpose.

3.4 Dataset Statistics

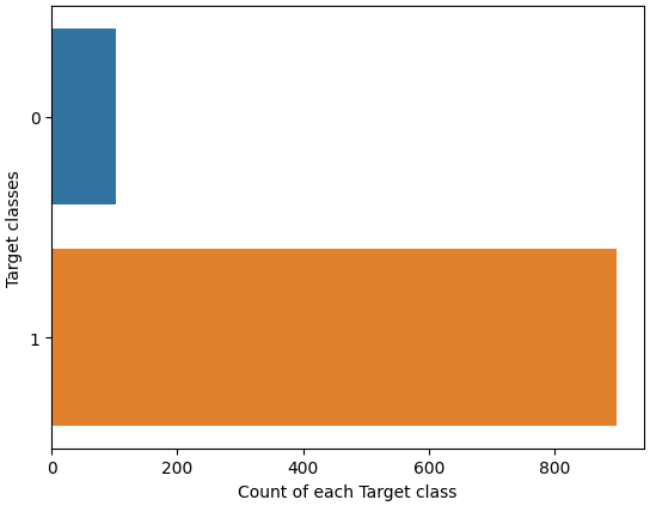

Fig 1 : Counting of each Target Class- Diabetic(class o, non-Diabetic(class-1)

Fig 1 represents number of records corresponding to diabetic and non-diabetic class.

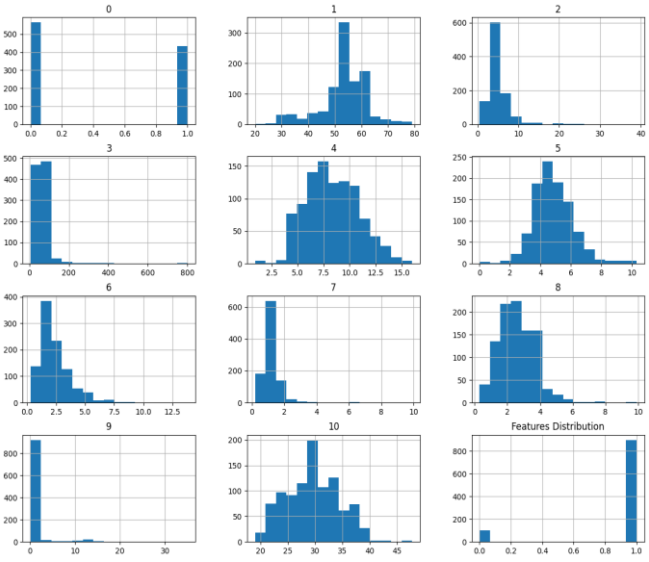

Fig 2 : Histogram of features

Fig 2 represents the distribution of data for each of the feature.

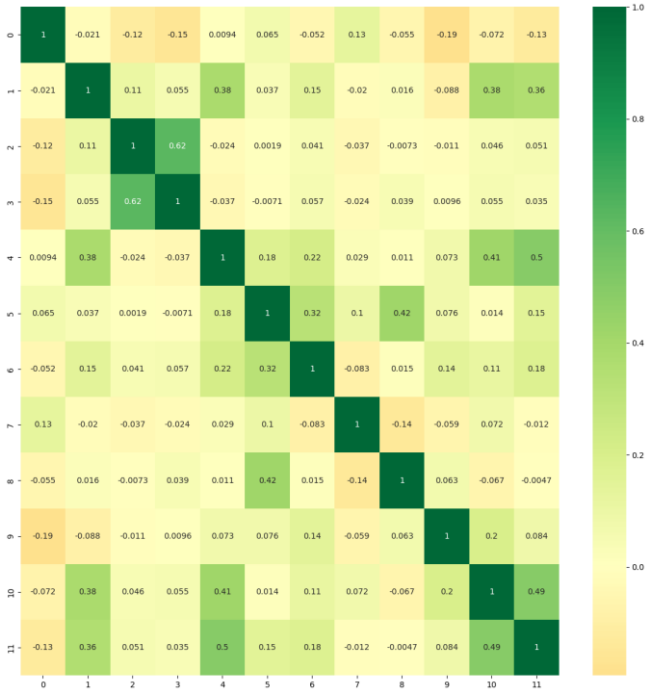

Fig 3: HeatMap

Fig 3 represents the heat map to relate feature values.

3. Experimental Methods

4.1 Overall Process Model

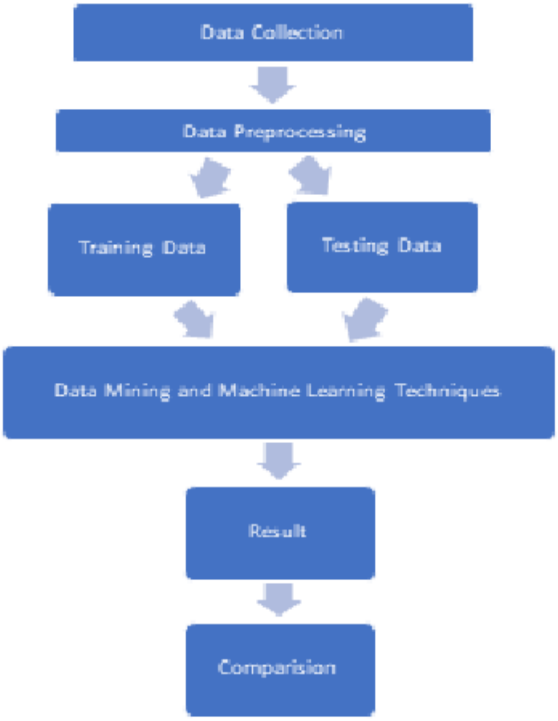

Fig 4: Process Model

Fig 4 shows the overall process model of the proposed system. Firstly data collection is done for that purpose, dataset from Mendeley data[1] is chosen. Next the preprocessing of data is done by converting textual data into numerical form and dividing data for training and testing purpose. Further various techniques from data mining and machine learning have been applied on the real dataset for the prediction of diabetes. Outcomes of different algorithms are compared to find the best result.

4.1 K-Nearest Neighbour (KNN)

KNN is very simple supervised Machine learning algorithm. It is based on the concept that new data is assigned to the category which has the maximum number of similar data available. K-NN is a non-parametric algorithm. Which means for the distribution of data no underlying assumptions is made.

The K-NN can be described using the following steps:

Step 1 – Firstly Choose the value of K for nearest neighbours.

Step 2 – For each test data point do the following steps-

- Calculate the distance between current test data point and all rows of training data. Distance can be: Manhattan, Euclidean or Hamming distance.
- Sort the rows based on distance value.
- Choose top K sorted rows.
- Test data point will be assigned to the class which is having maximum nearest training data.

Step 3-End

4.2 Decision Tree

Decision Tree is a Supervised learning technique. It is used for classification and Regression both. It has a tree-like structure, where internal nodes are represented by the features available in the dataset, branches are represented by the decision rules and leaf nodes are represented by the result.
It is a graphical way to represent the all possible solutions of given problem on the basis of the feature values. Figure 5 represents the decision tree structure.

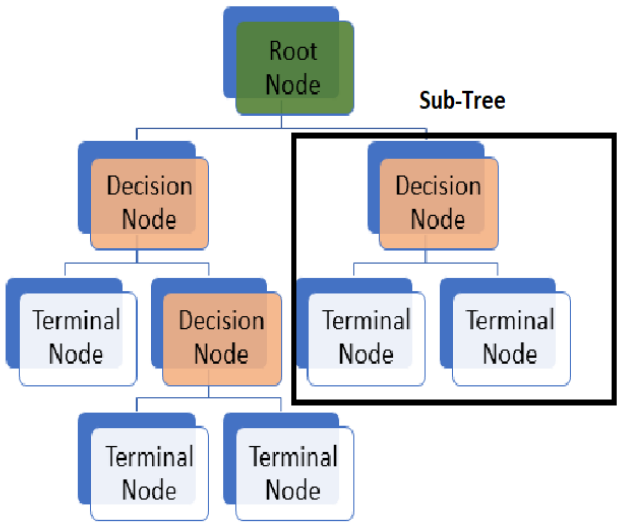

Fig 5: Decision Tree

4.3 Support Vector Machine

Support Vector Machine (SVM) is used for classification. It is a supervised machine learning algorithm. SVM algorithm is used to find the optimal hyperplane to separate the data points in different classes. SVM tries to maintain gap between the nearest points of different classes to be as large as possible. Following figure shows how the hyperplane separates the data point.

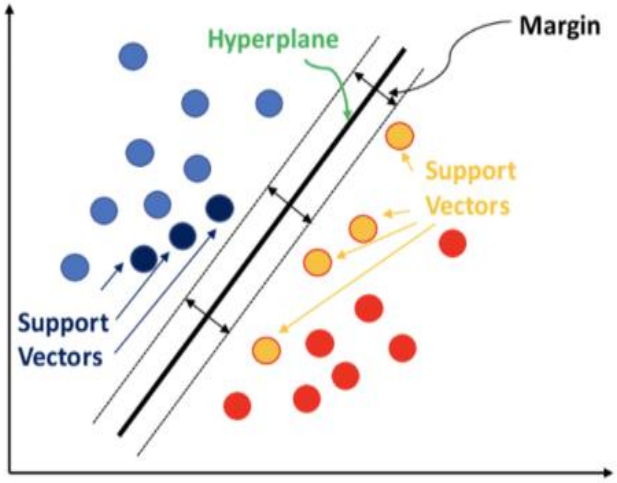

Fig 6: Support Vector Machine

4.4 Navie Bayes

Naive Bayes algorithm is used for classification problem. It is a supervised learning algorithm. It uses the concept of Bayes theorem for classification problem.

Bayes theorem calculate the probability of the current event using the given probability of already occurred event.

It is called as a probabilistic classifier, that means prediction is done on the basis of the probability of an object.

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)\mathrm{P(A)}}{\mathrm{P(B)}}$$

Where,

P(A|B) -Posterior probability- Probability of hypothesis A on the observed event B.

P(B|A) -Likelihood probability-Probability of the evidence

4.5 Artificial Neural Network (ANN)

Artificial Neural Networks consists of various layers of artificial neurons also called units. ANN has three types of layers, input layer, hidden layers and output layers. Input layer takes the real world data and passed it to multiple hidden layers which process and transform the data and passed it to the output layer. Output layer holds the units belongs to the different classes.

The proposed model for this paper is as follows

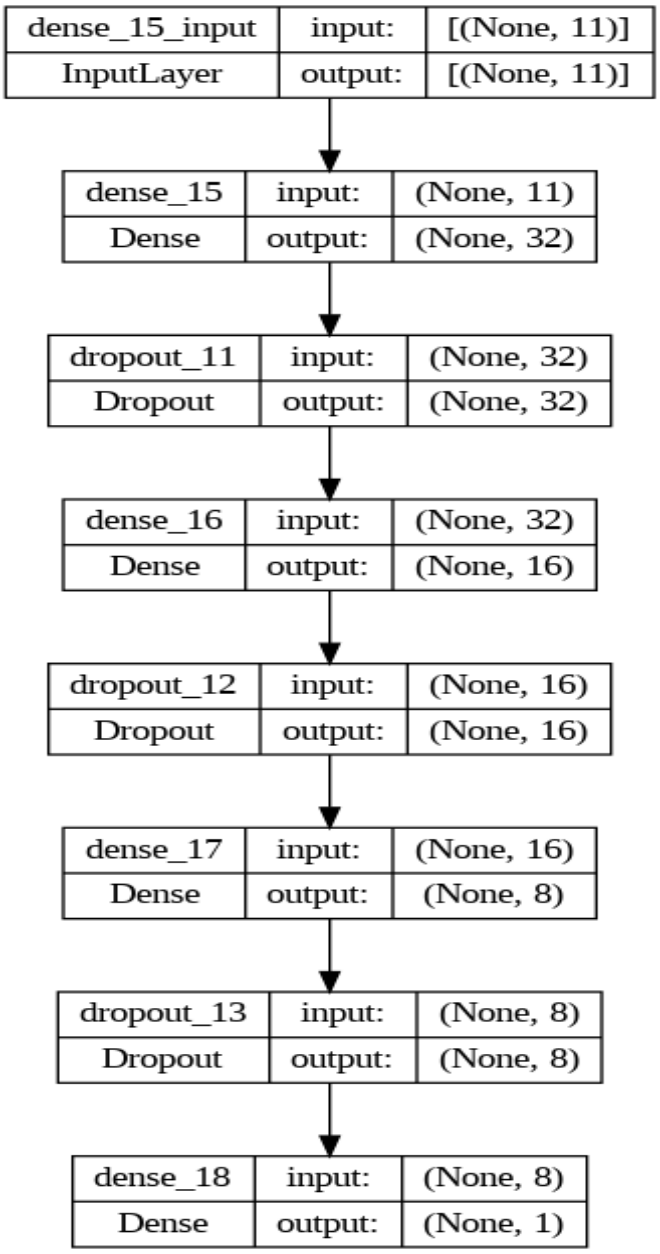

Fig 7: Proposed ANN Model

5. Results and Discussion

5.1 K-Nearest Neighbour (KNN)

Table 2 : KNN Performance

	precision	recall	f1-score	support
0	0.60	0.73	0.66	37
1	0.96	0.94	0.95	293
accuracy			0.92	330
macro avg	0.78	0.83	0.81	330
weighted avg	0.92	0.92	0.92	330

Above table shows the performance of the KNN algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. non-diabetic and diabetic all the four measures are given in the table. Table shows the accuracy of KNN algorithm is 92%.

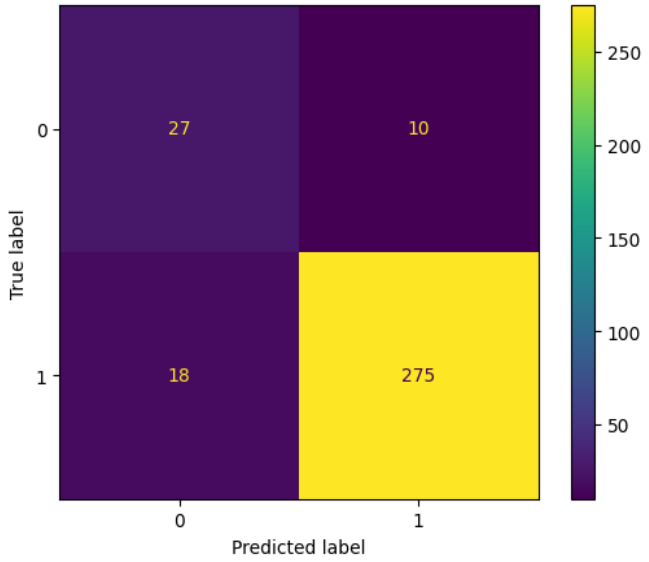

Fig 8: KNN Confusion Matrix

Fig 8 shows the confusion matrix for KNN algorithm. Diagram shows out of 37 non-diabetic test data 27 are correctly classified and out of 293 diabetic test data 275 are correctly classified.

5.2 Decision Tree

Table3: Decision Tree

	precision	recall	f1-score	support
0	0.97	0.95	0.96	37
1	0.99	1.00	0.99	293
accuracy			0.99	330
macro avg	.98	0.97	0.98	330
weighted avg	0.99	0.99	0.99	330

Above table shows the performance of the Decision Tree algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. non-diabetic and diabetic all the four measures are given in the table. Table shows the accuracy of Decision Tree algorithm is 99%. Which is the highest accuracy as compared to all other four algorithms.

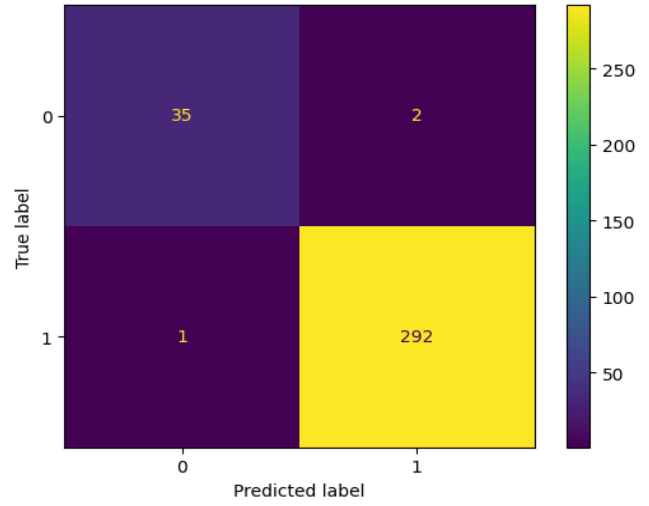

Fig 9: Decision Tree Confusion Matrix

Fig 9 shows the confusion matrix for Decision Tree algorithm. Diagram shows out of 37 non-diabetic test data 35 are correctly classified and out of 293 diabetic test data 292 are correctly classified. It shows the best result in comparison with other algorithms.

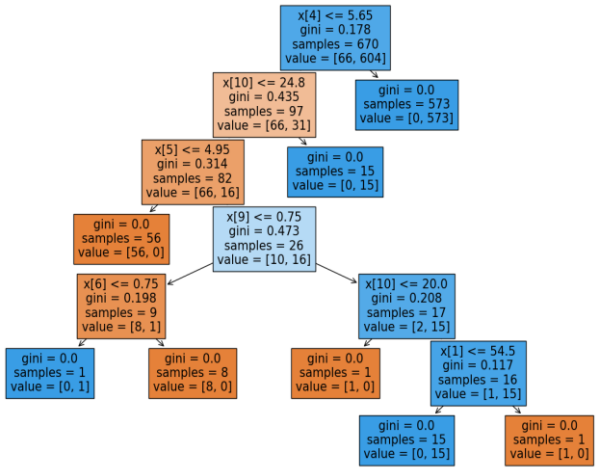

Fig 10: Decision Tree Model

Fig 10 represents the decision tree structure for the real data set[1] used in the experiment.

5.3 Support Vector Machine (SVM)

Table 4: SVM Performance

	precision	recall	f1-score	support
0	0.82	0.86	0.84	37
1	0.98	0.98	0.98	293
accuracy			0.96	330
macro avg	0.90	0.92	0.91	330
weighted avg	0.96	0.96	0.96	330

Above table shows the performance of the Support Vector Machine (SVM) algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. non-diabetic and diabetic all the four measures are given in the table. Table shows the accuracy of SVM algorithm is 96%.

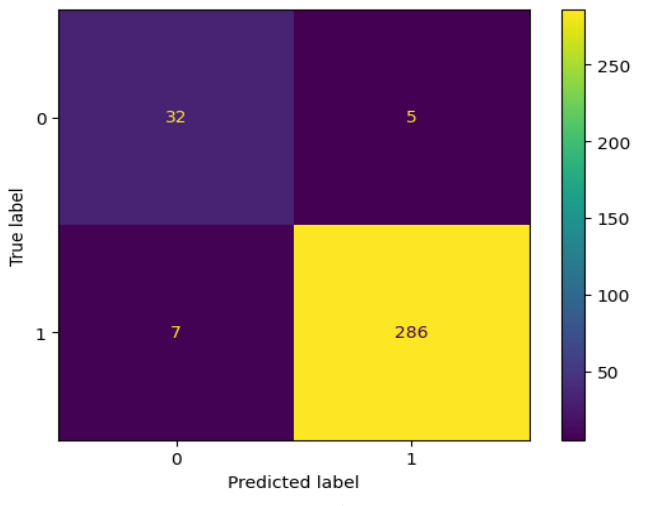

Fig 11: SVM Confusion Matrix

Fig 11 shows the confusion matrix for Support Vector Machine (SVM) algorithm. Diagram shows out of 37 non-diabetic test data 32 are correctly classified and out of 293 diabetic test data 286 are correctly classified.

5.4 Navie Bayes

Table 5: Navie Bayes Performance

	precision	recall	f1-score	support
0	0.62	0.89	0.73	37
1	0.99	0.93	0.96	293
accuracy			0.93	330
macro avg	0.80	0.91	0.85	330
weighted avg	0.94	0.93	0.93	330

Above table shows the performance of the Navie Bayes algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. non-diabetic and diabetic all the four measures are given in the table. Table shows the accuracy of Navie Bayes algorithm is 93%.

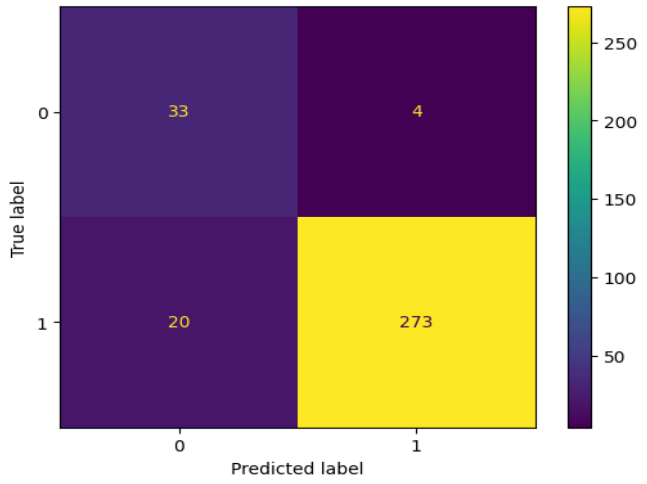

Fig 12: Navie Bayes Confusion Matrix

Fig 12 shows the confusion matrix for Navie Bayes algorithm. Diagram shows out of 37 non-diabetic test data 33 are correctly classified and out of 293 diabetic test data 273 are correctly classified.

5.5 Artificial Neural Network (ANN)

Table 6: Proposed ANN Model Performance

	precision	recall	f1-score	support
0	0.73	0.73	0.73	37
1	0.97	0.97	0.97	293
accuracy			0.94	330
macro avg	0.85	0.85	0.85	330
weighted avg	0.94	0.94	0.94	330

Above table shows the performance of the Artificial Neural Network (ANN) algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. non-diabetic and diabetic all the four measures are given in the table. Table shows the accuracy of Artificial Neural Network (ANN) algorithm is 94%.

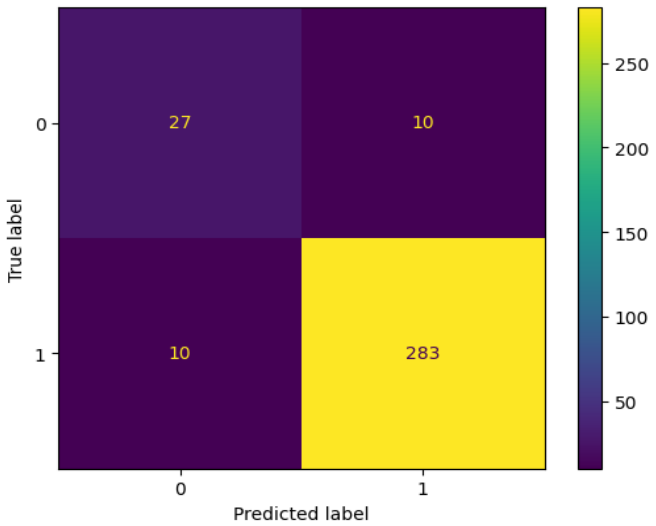

Fig 13: Proposed ANN Model Confusion Matrix

Fig 13 shows the confusion matrix for Artificial Neural Network (ANN). Diagram shows out of 37 non-diabetic test data 27 are correctly classified and out of 293 diabetic test data 283 are correctly classified.

5.6 Comparative analysis of algorithms

Table 7: Comparison Chart of five Algorithms based on measures Precision, recall, f1-score and Accuracy

	Class	precision	recall	f1-score	Accuracy
KNN	**0**	0.6	0.73	0.66	0.92
	1	0.96	0.94	0.95	
Decision Tree	**0**	0.97	0.95	0.96	0.99
	1	0.99	1	0.99	
SVM	**0**	0.82	0.86	0.84	0.96
	1	0.98	0.98	0.98	
Navie Bayes	**0**	0.62	0.89	0.73	0.93
	1	0.99	0.93	0.96	
ANN	**0**	0.73	0.73	0.73	0.94
	1	0.97	0.97	0.97	

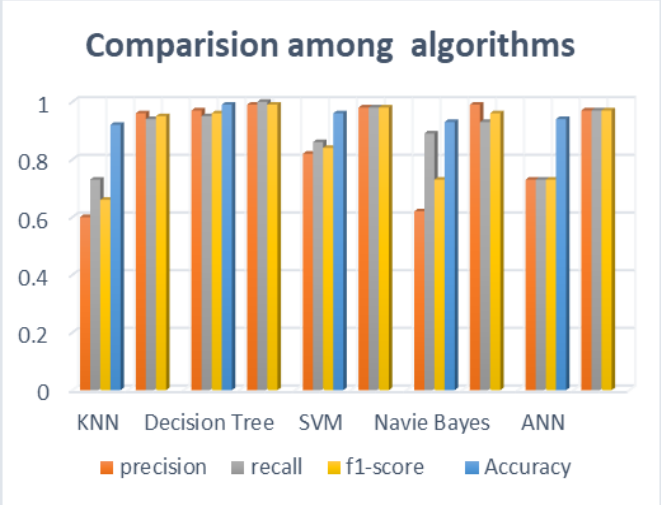

Fig 14: Comparison among algorithms

Fig14 shows the comparative chart of four performance measures of all the five algorithms. As chart shows Decision tree performs best among all other algorithms. Still there is a scope to modify ANN model for the better result.

6. Conclusion and Future Scope

Diabetes is a common health issue facing by most of the people in the world. Early detection of the problem can prevent us from the damage of organs in some extent. In this paper we focused on five different machine learning and datamining techniques KNN, decision Tree, Support Vector Machine, Naive Bayes and Artificial Neural Network. We have taken the Mendeley data [1] for the experimental purpose. As the result shows decision tree gives 99% accuracy for the prediction of diabetes which is the best result as compare to other algorithms. However proposed ANN model can be modified for the more accurate result.

Funding Source
No funding agency or funding source is involved with this research work.

Authors' Contributions
Author – 1 and Author 2 both have done literature study. Experimental part is also the combined efforts of author 1 and author 2. In paper writing also both the authors have equal contribution.

References

[1]. Rashid, Ahlam, "Diabetes Dataset", Mendeley Data, V1 **2020**, doi:10.17632/wj9rwkp9c2.1
[2]. World Health Organization, **2016.** WHO/NMH/NVI/16.3.
[3]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". *Conference: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) IEEE,* pp.**942-928, 2018.**
[4]. Tao Zheng , Wei Xie , Liling Xu , Xiaoying He , Ya Zhang , Mingrong You , Gong Yang , You Chen "A machine learning-based framework to identify type 2 diabetes through electronic health records." *International journal of medical informatics 97*. pp.**120- 127, 2017.**
[5]. Nahla hosny Barakat, Andrew P Bradley and Mohammed N Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus". *Information Technology in Biomedicine", IEEE Transactions*. pp.**1114-1120, 14 July, 2010**.
[6]. V. Veena Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus A machine learning approach". IEEE Recent Advances in Intelligent Computational Systems (RAICS*)*, pp.**122–127, 2015**. doi:10.1109/RAICS.2015.7488400.
[7]. Deeraj Shetty, Kishor Rit, Sohail Shaikh, and Nikita Patil, "Diabetes Disease Prediction Using Data Mining". *International Conference on Innovations in Information, Embedded and Communication Systems* (ICIIECS), **2017.**
[8]. B.P. Shantakumar, and Y.S. Kumaraswamy, "Intelligent and effective heart attack prediction system using data mining and artificial neural networks". European Journal of Scientific Research. Vol.**31,** pp.**642–656. 2009.**
[9]. Karim M. Orabi, Yasser M. Kamal and Thanaa M. Rabah,"Early Predictive System for Diabetes Mellitus Disease". *In Industrial Conference on Data Mining*, Springer, pp.**420–427, 2016.**
[10]. Tarik A. Rashid, Saman M. Abdullah, and Reezhna M. Abdullah, "An Intelligent Approach for Diabetes Classification, Prediction and Description". *In Advances in Intelligent Systems and Computing*, **Vol.424,** pp.**323–335, 2016.**
[11]. Zarita Zainuddin, Pauline Ong, and Cemal Ardil, " A neural network approach in predicting the blood glucose level for diabetic patients". International Journal of Computational Intelligence, Vol.**5,** Issue.**1,** pp.**72-79, 2009.**
[12]. Sadhana Tiwari , Awadhesh Kumar , Aasha Singh, "A Machine Learning Based Diabetes Prediction Using Stacking and Stacking With Hyperparameter Tuning", International Journal of Computer Sciences and Engineering, Vol.**10**, Issue.**6**, June **2022.**
[13]. Pradeep Kumar G., R. Vadivel, "Python Based Diabetes Prediction Using Ensemble Machine Learning Techniques Using LR Algorithm and Hybrid Method", International Journal of Computer Sciences and Engineering, Vol.**10**, Issue**.5**, May **2022**.
[14]. K. Gandhimathi1 , N. Umadevi, " Prediction of Type 2 Diabetics besed on Clustering Algorithm:", International Journal of Computer Sciences and Engineering, Vol.**8**, Issue**.11**, November **2020.**
[15]. B. Vinothkumar1* , M. Ramaswami, "A Novel Prediction of Diabetes by Automatic Insulin Therapy Using Machine Learning Algorithm", International Journal of Computer Sciences and Engineering, Vol.**8**, Issue.**3**, Mar **2020.**

AUTHORS PROFILE

Vaishali Sarde earned her MCA from Govt. Engineering College Raipur, M. Tech(CSE) from Rungta College of Engineering and Technology, Bhilai, Ph. D. from DR. C. V. Raman University, Bilaspaur. She is having more than 15 years of teaching experience. Currently Working in Govt. J. Yoganandam Chhattisgarh College, Raipur (C.G.) India. Her research area includes data mining, machine learning, AI, Blockchain technology etc.

Pankaj Sarde earned his M. Sc. From Govt. J. Yoganandam Chhattisgarh College, Raipur (C.G.) India. He has Completed Ph. D. from Pt. Ravishankar Shukla University, Raipur. He is having more than 20 years of teaching experience. His main research work focuses on Blind Signature, Bilinear Pairings, Digital Signature, Elliptic Curve, Identity based cryptography etc.

