2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)
Bangalore, India. Oct 7-9, 2022

Performance Evaluation of Predictive Machine Learning Models for Diabetic Disease Using Python

Madhubrata Bhattacharya
Department of Physics, The Heritage College,
Heritage Institute of Technology,
Chowbaga Road, AnandaPur, Kolkata-700107
India
madhubrata.bhattacharya@thc.edu.in

Debabrata Datta
Department of Information Technology,
Heritage Institute of Technology,
Chowbaga Road, AnandaPur, Kolkata-700107
India
debabrata.datta@heritageit.edu

2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT) | 978-1-6654-6855-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/GCAT55367.2022.9972220

***Abstract*- The discovery of knowledge from medical database is always beneficial as well as challenging task for diagnosis. For example, patients having high blood glucose are required to diagnose as they fall within a group of Diabetes mellitus. Prediction of diabetes mellitus is an essential research in the domain of medical industry. With the advent of artificial intelligence and machine learning this type of prediction removes the hurdles faced in data mining used for similar task. In case of data mining, extraction of knowledge from information stored in database takes place and an understandable description of patterns is achieved. A large number of researches have been already taken place to predict diabetes using traditional machine learning algorithm such as artificial neural network, Naïve Bayes theorem, decision tree, etc. However, determination of diabetes with a certain degree of confidence is required from the accuracy or any other performance measures point of view. In this context, this research work presents machine learning models such as decision tree, support vector machine, random forest, k-nearest neighbours and Naïve-Bayes as classifier to classify whether a patient is diabetic or prone to diabetic. Performance measures of these algorithms have been carried out in terms of accuracy score. Dataset for training and testing the algorithms mentioned is retrieved from Pima Indian Database. On the basis of their comparative evaluation, most important feature with respect to identification of diabetic is extracted. A complete python code has been developed for this research work.**

Keywords: diabetes, Naïve-Bayes, random forest, k -nearest neighbour, decision tree

I. INTRODUCTION

Diabetes Mellitus is an extensive disease in which the hormone insulin producing capacity of the body becomes affected. It increases the glucose level in the blood due to abnormal carbohydrate metabolism which in turn affects the vital organs of human body causing other health disorders. Diabetes is broadly classified into four categories such as Type-I, Type-II, Gestational and Pre-diabetes. Type-I is sometimes called as the "insulin-dependent" diabetes [1], Type-II is known as "insulin resistance" diabetes [1] which is usually diagnosed later on for most of the cases. During pregnancy, the insulin blocks the hormones and Gestational (Type III) diabetes occurs. When the blood sugar level goes above the normal level, it is called Pre-diabetes.

Canadian Diabetes Association (CDA)' showed that, during the year 2010 to 2020 the number of individuals having the diabetes in Canada has been escalated to approximately 3.7 million from 2.5 million [2]. As specified by the 'International Diabetes Federation', 382 million people are having the disease out of 2013 [3] which are 6.6% of the total grown up population of the world. According to the World Health Organisation (WHO) in 2012, diabetes was one of the leading reasons for death of 1.5 million patients. Early prediction of the disease can save these lives. Therefore, the goal of this research is to have an early prediction of diabetes using machine learning algorithms. Our research will address machine learning algorithms as classifier for such detection.

Recent trends envisaged that machine learning (ML) has become a fast growing technology in the field of medical management and wellness program by enabling the clinical experts for data analysis. This helps to identify the patterns and cautions that can make improved diagnosis and therapy. In recent years, a number of researches are being conducted in this field using various algorithms of classifications in ML. Scientists have shown that ML algorithms [4-6] works well in diagnosing diseases. Orabi et al. in [7] designed a model based on ML to predict diabetes at a particular age. Genetic programming (GP) was used by Pradhan et al. [8] in analyzing the database. Harry Zhang [9] illustrated the performance of Naïve Bayes (NB) to show that it is optimal. K- Nearest neighbour (KNN) algorithm is used only in diagnosis of diabetes mellitus in [10] whereas in [11] complications in diabetes are predicted by Dagliati et al. Aljumah et. al. [12] framed a predictive analysis model using Support Vector Machine (SVM) algorithm. In this work, real diagnostic data has been analysed using ML classification algorithms to predict diabetes mellitus. The objective of this research work is to develop a new classifier algorithm known as Bayesian Belief Network (BBN) to predict diabetic. The task of verification and validation of this new classifier is achieved by comparing its performance with traditional NB and random forest (RF) classifier. Accuracy and sensitivity will be the metrics as performance measure. The remaining structure of the paper is organized into sections. Section 2 describes different classifier models used in this work. In section 3, the statement of the problem of interest, resource of input dataset used for the present work and the results of prediction with models of interest are discussed. Section 4 finally presents conclusions of this research work with future scope.

II. Different Classifier Models Used

A. Naïve Bayes (NB) classifier

NB classifier is a method of classification established on conditional probability with all unrelated and independent features. According to this classifier technique, status of a particular feature in a class remains unaffected by another feature. It is effective for the data with variation problems and problems where some values are missing. NB classifier employs Bayes theorem that is written mathematically as

$$P(Y|X) = P(X|Y)P(Y)/P(X) \quad (1)$$

where, X and Y are events with $P(X) \neq 0$. $P(Y|X)$ is defined as posterior probability, $P(X|Y)$ is known as likelihood. $P(Y)$ is the prior probability and $P(X)$ is defined as evidence. Now introducing Naïve assumption to Bayes theorem (independence of features), according to data set $X = (X_1, X_2 \ldots . X_n)$ Bayes theorem can be written as

$$P(Y|X_1, X_2 \ldots . X_n) = \frac{P(X_1|Y)P(X_2|Y)\ldots P(X_n|Y)P(Y)}{P(X_1)P(X_2)\ldots P(X_n)} = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(X_1)P(X_2)\ldots P(X_n)} \quad (2)$$

We know that for any two independent events (mutually exclusive) A and B, $P(A, B) = P(A)P(B)$. The denominator of Eq. (2) can be removed as for 's' specific input remains constant and we can write

$$P(Y|X_1, X_2 \ldots . X_n) \propto P(Y) \prod_{i=1}^{n} P(X_i|Y) \quad (3)$$

Now to create a model the probability of a specific input set can be evaluated. Here all possible values of the variable Y should be taken into account. Finally the maximum probability of the output Y is taken as $Y = argmax_Y P(Y) \prod_{i=1}^{n} P(X_i|Y)$.

Using the equations (1-3), we can obtain the class provided the predictors are given.

B. Random Forest (RF) Classifier

Random forest (RF) classifier is based on supervised machine learning algorithm. RF can be used for problems pertaining to classification and regression. The basic concept on which this machine learning algorithm is based is ensemble learning [13]. In RF classifier, decision trees are applied on various data points of the input dataset and then average performance of all decision trees is estimated to improve the accuracy of the dataset. Consequently, if the number of trees in the forest is large, higher accuracy can be achieved and over fitting problems can be avoided. Decision rules taken from prior information are used to predict the target class. The function of random forest through a number of trees is as shown in Fig. 1. Decision tree selects each node in each stage by finding the information gain among all the features [13].

The tree begins with the root node consisting the input dataset. In the next step the best feature is selected and the root node is divided into branches containing the possible outcomes of the best feature. Finally the decision node is generated which contains the best feature. The process continues until a stage is achieved where further classification of nodes is not possible and the final node is named as leaf node. Random forest algorithm is based on certain assumptions which are:

a) In the data set, the feature variable should have some real values so that accurate results can be predicted by the classifier avoiding a guessed result.

b) Very low correlations must be maintained between the predictions from each decision tree.

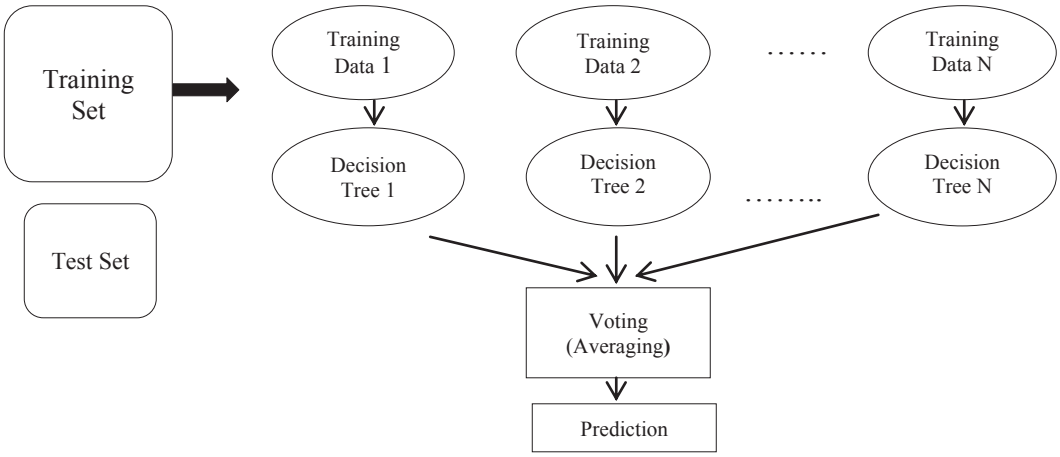

Fig. 1. Structure of Random Forest

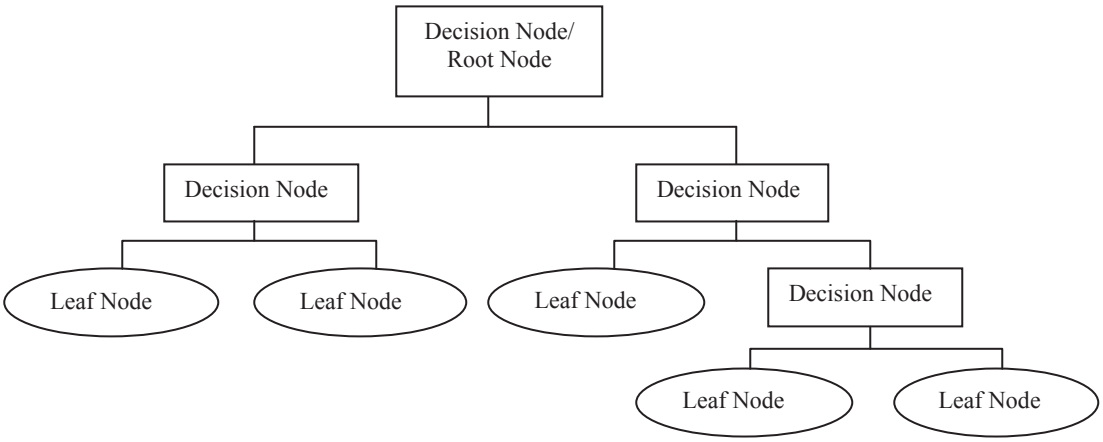

Fig. 2. General structure of a Decision tree

The algorithm of random forest can be represented as:
Step 1: Select K number of data points from the training set {Tr} randomly
Step 2: Construct decision trees linked with the selected data points
Step 3: Choose the number N to build the appropriate decision tree.
Step 4: Repetition of Steps 1 & 2

C. K-Nearest Neighbour (K-NN) Classifier

KNN classifier accepts the similarities between the available data and new data and keeps the new data into the class having the maximum similarities with the available class. The algorithm of K-NN can be written as:
Step 1: Select K as the number of neighbours
Step 2: Compute the Euclidean distance of K neighbours and choose the nearest-neighbours
Step 3: Count the number of the data points in each class amongst these neighbours
Step 4: Store new data, where maximum nearest neighbours are located

D. Decision Tree (DT) Classifier

Decision rules taken from prior information are used to predict the target class. The tree shaped (Fig.1) classifier uses nodes to describe the features of a given dataset, decision rules are shown by the branches and the outcome is illustrated by the leaf nodes. Decision tree selects each node in each stage by finding the information gain among all the features [13].

The algorithm of Decision tree can be represented as:
Step 1: Start from the root node
Step 2: Select the best feature using Attribute Selection Measure (ASM)
Step 3: Divide root node into branches containing the possible outcomes of the best feature
Step 4: Generate decision node with the best feature
Step 5: Repeat step 3 to 5 until leaf node is classified

E. Support Vector Machine (SVM) Classifier

Classification is achieved in SVM classifier by realizing a linear or non-linear surface in the input space.

In Support vector classification, kernels linked with the Support Vectors are linearly combined to express the separating function as

$$f(x) = \sum_{x_j \in S} \alpha_j y_j K(x_j, x) + b \quad (4)$$

Where the training patterns are denoted by x_i, $y_i \in \{+1, -1\}$ explains the class labels of the patterns and S is the set of support vectors. The objective function to be minimized is denoted as dual function and is formulated as

$$\min W = \frac{1}{2}\sum_{i,j} \alpha_i Q_{ij} \alpha_j - \sum_i \alpha_i + b \sum_i y_i \alpha_i \quad (5)$$

Where α_i represents the corresponding coefficients, b denotes the offset, $Q_{ij} = y_i y_j K(x_i, x_j)$ represents a Kernel matrix (symmetric and positive definite) and the parameter C penalizes the error points in typical cases. The Karush-Kuhn-Tucker (KKT) conditions for the dual can be illustrated as $\frac{\partial W}{\partial \alpha_i} = g_i = y_i f(x_i) - 1$ *and* $\frac{\partial W}{\partial b} = \sum y_j \alpha_j = 0$. Details of more about SVM can be found in [14].

III. Problem Statement, Input Dataset, Results and Discussions

Our goal is to build a machine learning based predictive model with a classifier of early detection of diabetes. The patient suffered from diabetic disease depends on many factors such as:

Body Mass Index (BMI), (ii) Insulin Level, (iii) Age, (iv) Diabetes Pedigree, (v) Pregnancy, (vi) Glucose index, (vii) Blood Pressure, (viii) Skin Thickness and (ix) Outcome

The features or covariates are as shown in Fig. 3

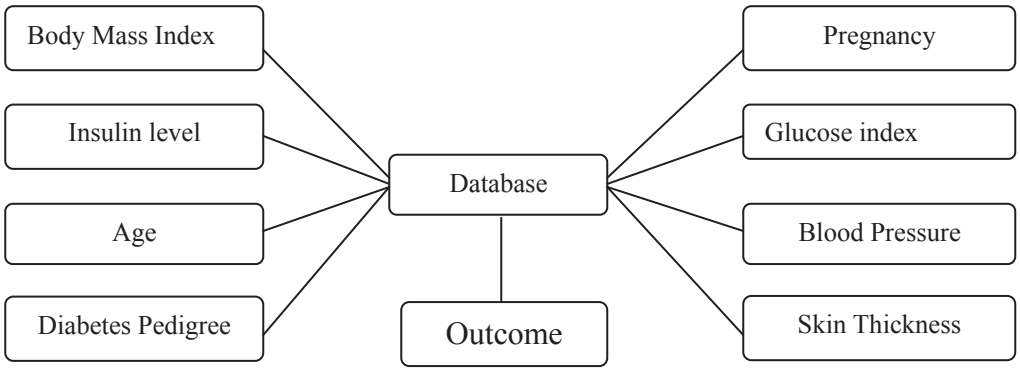

Fig. 3. Attribute Description

All these features are assumed to be normal for the sake of simplicity in computation. The outcome of the model is further compared and also validated with other classifier algorithms.

A. Data Pre-processing

Sample values of all dataset of 768 data have been checked and normalised and there is no missing values found in the dataset (checked by isnull() of PYTHON). Descriptive statistics of main features are shown in Table 1. Probability distribution of individual features are investigated and illustrated in Fig. 4. Correlation of features is explained by the Heat map shown in Fig. 5. A comparative study of importance of all the features has been carried out using the 768 data set and the outcome is presented in Fig. 6.

TABLE I. DESCRIPTIVE STATISTICS OF MAIN FEATURES

	Pregnancy	Glucose	Age	Outcome
Sample Size	768	768	768	768
Mean	3.84	120.89	33.24	0.34
Standard Deviation	3.37	31.97	11.76	0.47
Min	0.00	0.00	21.00	0.00
25%	1.00	99.00	24.00	0.00
50%	3.00	117.00	29.00	0.00
75%	6.00	140.25	41.00	1.00
Max	17.00	199.00	81.00	1.00

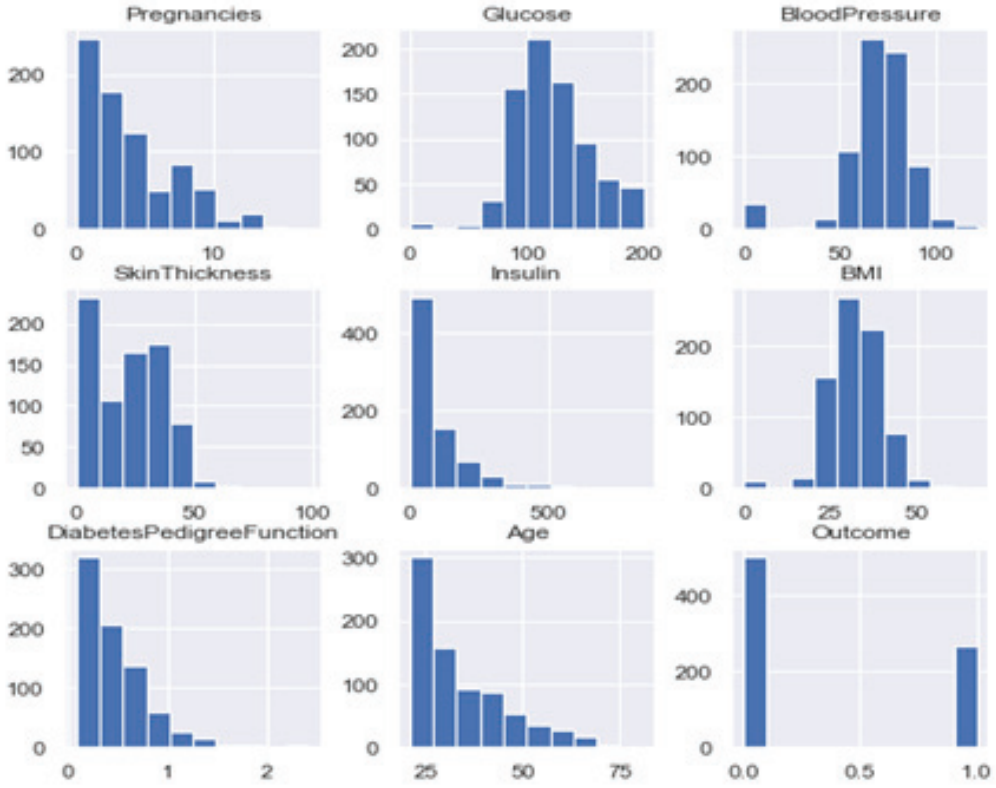

Fig.4: Histogram of individual features

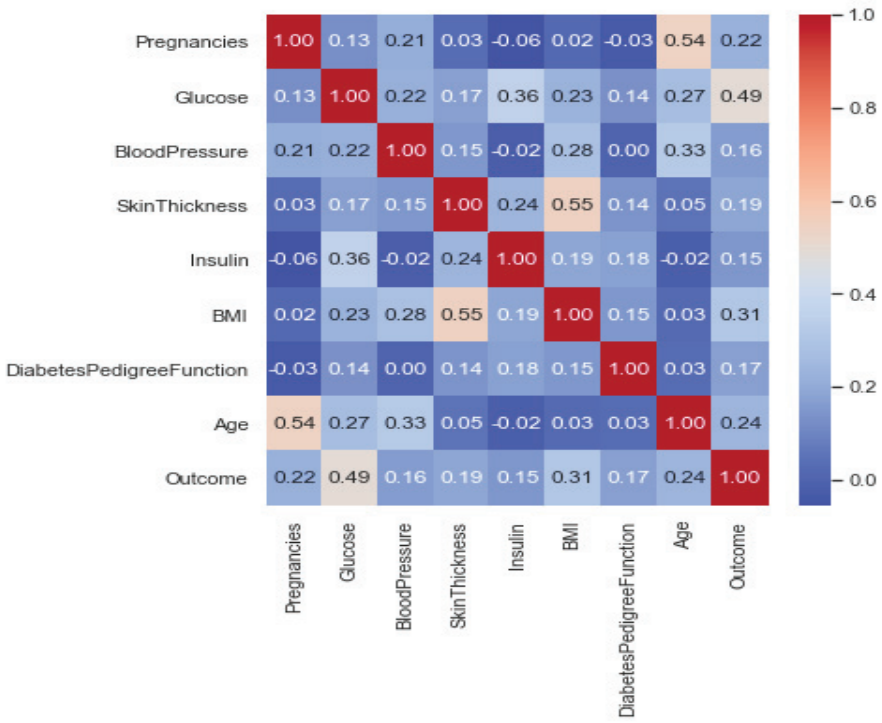

Fig.5: Heat map explaining the correlation of features

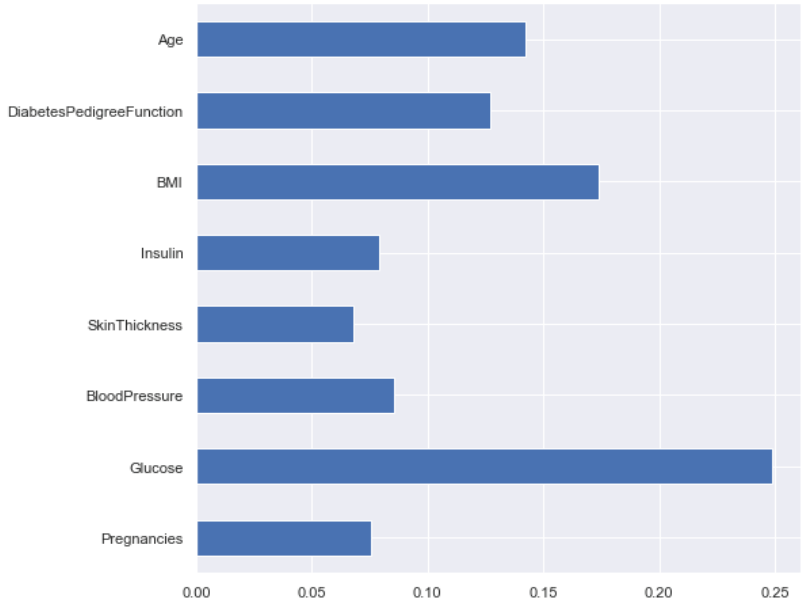

Fig.6: Feature importance

B. *Confusion matrix*

A confusion matrix carries information regarding real and predicted classifications performed by a classifier system. The matrix data evaluate the system performance. For a two class classifier the general format of a confusion matrix is shown below.

	Predicted classes	
Actual classes	True positive	False positive
	False negative	True negative

- If both the real value and the predicted value are positive then it is called true positive.
- If the real value and the predicted value both are negative then it is called true negative.
- If the real value is negative but predicted value is positive it is false positive.
- If the real value is positive but predicted value is negative it is false negative.

Accuracy can be calculated from the confusion matrix as: (TP +TN) / (P+N)

In this work we have used five classifier models such as, Naïve Bayes, random forest, support vector machine, decision tree and k -nearest neighbour to investigate the data. The complete data set is partitioned into training and testing using sklearn module in PYTHON wherein train_test _split function is invoked. The performance report of the classifier models are displayed in details in Table 2.

TABLE II. CLASSIFICATION REPORT OF DIFFERENT CLASSIFIER MODELS

Classifier Model	Confusion matrix	Outcome	Precision	Recall	f1-score
Naïve Bayes	[125 26] [36 44]	0	0.76	0.89	0.82
		1	0.70	0.47	0.57
		Macro avg.	0.73	0.68	0.70
		weighted avg.	0.74	0.75	0.73
Random Forest	[129 22] [35 45]	0	0.79	0.85	0.82
		1	0.67	0.56	0.61
		Macro avg.	0.73	0.71	0.72
		weighted avg.	0.75	0.75	0.75
K-Nearest Neighbour	[130 21] [37 43]	0	0.78	0.86	0.82
		1	0.67	0.54	0.60
		Macro avg	0.73	0.70	0.71
		weighted avg	0.74	0.75	0.74
Decision Tree	[131 20] [39 41]	0	0.77	0.87	0.82
		1	0.67	0.51	0.58
		Macro avg.	0.72	0.69	0.70
		weighted avg.	0.74	0.74	0.73
Support Vector machine	[135 16] [42 38]	0	0.76	0.89	0.82
		1	0.70	0.47	0.57
		Macro avg.	0.73	0.68	0.70
		weighted avg.	0.74	0.75	0.73

The performances of all five classifier models can be compared by evaluating the accuracy. Following figure (Fig.7) shows clearly that among the five classifier models Decision Tree classifier provides the accuracy of highest order. So in this work we can conclude that Decision Tree is the best model in comparison with the other four classifier models.

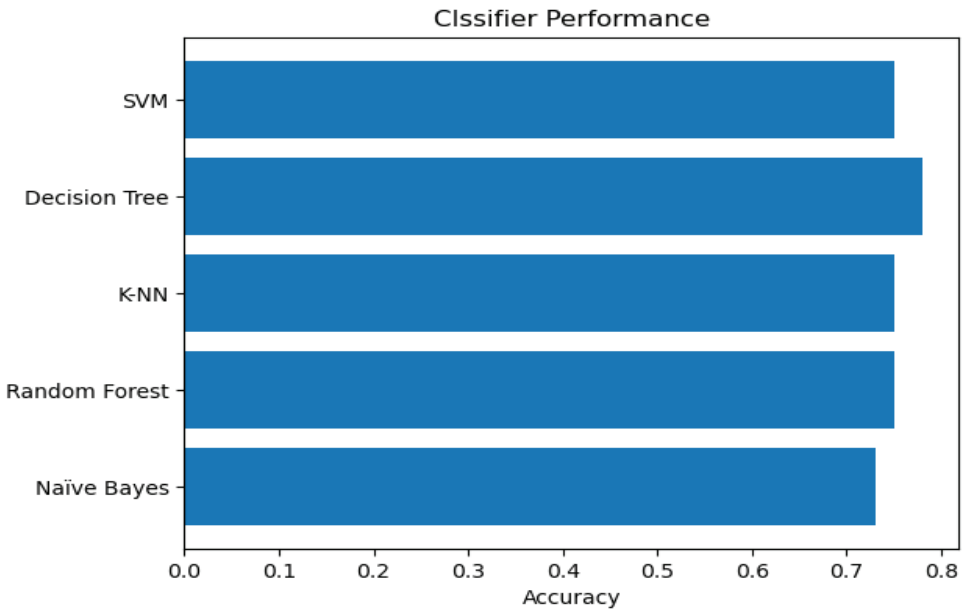

Fig. 7: Performances of Classifier Models

IV. Conclusions and Future Scope

Dataset retrieved from Pima Indian Database are pre-processed by the methods of normalisation and numerical discretization. Five different types of classifiers are applied to the processed data and the performances are investigated towards different features of diabetes attributes using PYTHON. Classification with Decision Tree model presents the best accuracy.

A large size of database is required to compare the various ML models. In addition to this the data may not be available always in the form of a crisp value but in terms of linguistic variable like less, high, very high. So, due to some uncertain factors of some of the diabetes features, fuzzy set approach can be tested to compare the results.

For future study we are planning to introduce Bayesian belief network as well in this work in order to get a better prediction model. A mixture database can be incorporated also.

References:

[1] S. Rani and S. Kautish, "Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction," Second International Conference on Intelligent Computing and Control Systems (ICICCS),2018

[2] Morteza, M., Franklyn, P., Bharat, S., Linying, D., Karim, K. and Aziz G. 2015. Evaluating the Performance of the Framingham Diabetes Risk Scoring Model in Canadian Electronic Medical Records. Canadian journal of diabetes 39, 30, 152-156(April. 2015).

[3] V., A. K. and R., C. 2013. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. 3, 1797-1801(April. 2013).

[4] Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[5] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116, 2017.

[6] DhomseKanchan B., M.K.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, IEEE. pp. 5–10, 2016.

[7] Orabi, K.M., Kamal, Y.M., Rabah, T.M., Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427 2016.

[8] Bamnote, M.P., G.R. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5, 2014.

[9] Harry Zhang, "The Optimality of Naïve Bayes", Published in FLAIRS Conference 2004

[10] D. KratiSaxena, Z. Khan, and S. Singh, "Diagnosis of diabetes mellitus using k nearest neighbor algorithm," International Journal of Computer Science Trends and Technology (IJCST), 2014.

[11] Dagliati, A., Marini, S.,Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., and Bellazzi,. R. "Machine learning methods to predict diabetes complications," Journal of diabetes science and technology, vol. 12, no. 2, pp. 295–302, 2018.

[12] Abdullah A. Aljumah et al., Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2, Pages 127-136,July 2013.

[13] Iyer, A., S, J., Sumbaly, R. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. 2015.

[14] Vapnik V.N., The Nature of Statistical Learning Theory, Springer, New York, 2nd Edition, 2000.