Diagnosis and Classification of the Diabetes Using Machine Learning Algorithms

Prasannavenkatesan Theerthagiri (prasannait91@gmail.com)
GITAM University Bengaluru
Usha Ruby A
GITAM University Bengaluru
Vidya J
GITAM University Bengaluru

Research Article

Keywords: Diabetes prediction, MLP, machine learning algorithm, Classification of diabetes

Posted Date: May 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-514771/v1

Abstract

Diabetes mellitus is characterized as a chronic disease may cause many complications. The machine learning algorithms are used to diagnosis and predict the diabetes. The learning based algorithms plays a vital role on supporting decision making in disease diagnosis and prediction. In this paper, traditional classification algorithms and neural network based machine learning are investigated for the diabetes dataset. Also, various performance methods with different aspects are evaluated for the K-nearest neighbor, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms. It supports the estimation on patients suffering from diabetes in future. The results of this work shows that the multilayer perceptron algorithm gives the highest prediction accuracy with lowest MSE of 0.19. The MLP gives the lowest false positive rate and false negative rate with highest area under curve of 86 %.

1. Introduction

Diabetes (Diabetes Mellitus - DM) is one of the metabolic disorders, with inappropriately raised blood glucose levels. The carbohydrates consumed will be turned into a type of sugar called glucose and it will be released into the bloodstream. Insulin, a hormone helps move glucose from blood to cells. With this chronic condition, the pancreas will produce little or no insulin, sometimes the produced insulin will not be absorbed by the cells this is termed as insulin resistance [1].

At present, diabetes is considered to be one of the lethal diseases across the globe and people are being affected in a huge number. Around 422 million people are diabetic patients and about 1.6 million deaths are attributed to diabetes every year. Over the past few decades, the number of cases and the prevalence of diabetes are steadily increasing [2].

DM is classified as type 1, type 2 and gestational diabetes. The condition where pancreas will produce little or no insulin is type 1 diabetes. If the insulin is not absorbed by the cells or not produced in enough quantity then it is referred to as type 2 diabetes (T2D). High glucose level during pregnancy is gestational diabetes and it returns back to normal after childbirth [1]. A high sugar level would increase the risk of complications like hearing loss, dementia, heart diseases, stroke, depression, vision loss, retinopathy, neuropathy, and so on. Early detection plays a prominent role in disease detection. It is one of the crucial causes of cardiovascular diseases, and there is an immense need to support the medical decision-making process. Many researchers in different medical diagnoses have employed various machine learning techniques [3].

Most of the researchers heed medical expert systems, and there has been much contemplation in this field. The medical experts and data analysts collaborate continuously to make this system more accurate and, thus, useful in real life. Recent surveys by the World Health Organization indicated a tremendous increase in diabetic patients and the demise attributed to diabetes every year. So, early diagnosis of diabetes is a significant concern among researchers and medical practitioners. [19] Multitudinous

computer-based detection systems were designed and outlined for analyzing and anticipating diabetes. The usual identifying process for diabetes takes time. Nevertheless, with the rise of machine learning, we can develop a solution to this intense issue [13].

To accurately predict the disorder a good model that can represent the presence of diabetes through input characteristics is required. With a good model and an accurate detection technique, diagnosis can be made more efficient. Based on the prediction medical practitioner can envision biomedical diagnosis by engineering tools that can automatically adapt to any unexpected future conditions. Along-term prediction algorithm can play a vital role in planning and provisioning. Intelligence systems can learn or adapt and modify the functional dependencies in response to new experiences or changes in functional relationships [16].

2. Literature Survey

Adel Al-Zebari et al have compared performance of various machine learning algorithms for diabetes detection. MATLAB classification Learner Tool has been used in this work that covers Decision tree, Discriminant analysis, SVM (Support Vector Machine), k-NN (k-Nearest Neighbor), Logistic regression and ensemble learners and their variants; totally 26 classifiers are considered. The results are evaluated on a 10-fold cross-validation basis and average classification accuracy is considered for performance measure. To increase accuracy deep neural networks, feature selection techniques can be used in future [10].

G.A. Pethunachiyar used SVM with disparate kernel functions for classification of diabetes. The simulation model of the proposed system includes 5 phases. After collecting the data, selection process is carried out by rectifying the errors (inconsistency in data or missing values or wrong information). Next data will be divided as training (70%) and testing dataset (30%). For efficient prediction SVM technique has been selected and model has been built. Test data is applied to the model in order to make prediction. The Linear, Polynomial and Radial kernel based SVM has been implemented in this work. Confusion matrix is used for calculating prediction accuracy. To evaluate 3 kernel functions ROC (Receiver Operating Characteristic curve) is used. Linear kernel with SVM predicts more accurately compared to other kernels [11].

Pahulpreet Singh Kohli et al, have applied various machine learning techniques on 3 different diseases datasets for disease prediction. Feature selection is carried out by backward modeling using p-value test. The proposed model includes 4 phases: Initially dataset is explored in Python environment. Next during data munging, the missing values are replaced with mean value and mode value for the continuous variable and categorical variable, respectively. Next features are selected very cautiously to improve the performance of the model. The attributes are eliminated using backward selection method (based on p-value it is eliminated) and the model is refitted. After selecting the features 5 algorithms including Decision Tree, Logistic Regression, Random Forest, Adaptive Boosting and Support Vector Machine were compared. The dataset has been divided into training set(90%) and test dataset(10%).In future data

munging, selection of features and model fitting steps can be automated; pipeline structure for preprocessing data would improve results [12].

Samrat Kumar Dey et al have developed a web application using Tensorflow for successful prediction of diabetes. This proposed model requires patient's data for successful diagnosis, and the techniques like SVM, ANN (Artificial Neural Network), KNN and Naive Bayes are used for predicting the disease. The dataset is divided into 2 parts: training and testing dataset. Preprocessing of data and data normalization would increase accuracy of the model. Min Max Scaler normalization model is used to improve accuracy. Deep learning model can be adopted in future to predict diabetes [13].

Sidong Wei et al, have done a comprehensive exploration on DNN (Deep Neural Network), Logistic regression, SVC (Support Vector Classifier), Naive Bayes and Decision tree techniques for identification of diabetes. This work has been carried out in 4 steps: Initially the best preprocessor is identified for the classifier. Next the parameters are optimized. In third step, these techniques are compared by accuracy, later relevance of these features are considered. The features like Plasma glucose concentration, age and number of times pregnant were found to be more significant [14].

S. Hari Krishnan et al, used machine learning techniques to measure the blood glucose level. A Photoplethysmograph (PPG) based system is used determine the glucose parameters which uses light sources of 3 different wavelengths. The light is illuminated, the skin at the wrist along with the reflected light are captured by a photodiode receiver and the same is conditioned, digitalized and sent to Arduino UNO microcontroller. The PPG signal is derived by the microcontroller in accordance with the blood glucose values. The waveform is preprocessed and segmented in order to obtain peak of the signal. To obtain the statistical features like mean, skewness, variance, kurtosis, standard deviation and entropy the Random forest technique is implemented on the acquired signal. The model is designed and trained to estimate the blood glucose from the features extracted. The future would focus on estimation of correlation of the feature sets with different machine learning techniques [15].

M.Shanthi et al, proposed and developed a model for diagnosing T2D using ELM (Extreme Learning Machine) technique. The ELM mathematical model has one hidden layer feed-forward network, which can create hidden nodes at random. Parameters are randomly generated for the hidden nodes initially. The next output matrix is calculated, and then the network's optimal weight is given as the output. From the characteristics, input weight, and activation functions, the output is obtained. The activation functions available are a triangular basis, sine, hard-limit, and sigmoid. This model assists medical experts to forecast T2D [16].

Sajratul Yakin Rubaiat et al, introduced an approach to predict type 2 diabetes using neural network. This analysis is carried out in two methods: The first method involves data recovery followed by selection of features, the selected features are inputted to MLP (multilayer perceptron) neural network classifier. Second approach uses K-means algorithm. Neural network based method involves 3 steps such as data recovery (missing data are replaced with mean value to complete the dataset), selection of features

e selected) and Multilayer Perceptron Classifier

(hyper parameters are selected. K-means reduces noise very effectively and its output has been used as feature for the model. Model can be trained using these 2 methods and predict whether or not a person is diabetic at an early stage. The first method is more efficient and requires less computation compared to k-means [17].

Maham Jahangir et al, presents a novel prediction framework which uses AutoMLP (automatic multilayer perceptron) combined with an outlier detection method. This method involves 2 stages: pre-processing of data with outlier detection following by training of AutoMLP. In the second stage it is used to classify the data. Compared to the other architectures of neural network AutoMLP gives higher accuracy. The attributes like plasma glucose level, blood pressure and number of times pregnant are found to be more relevant [18].

Ali Mohebbi et al used CGM (Continuous Glucose Monitoring) signals for adherence detection in diabetic patients. A considerable amount of signals were simulated using a T2D adapted version of the MVP (Medtronic Virtual Patient) model. Different classification algorithms were compared by using a comprehensive grid search. Logistic regression, Convolutional Neural Network (CNN), and Multi-Layer Perceptron techniques have been used in this work. CNN shows better performance in classification [19].

3. Methodology

3.1 Data Pre-processing and Cleaning

The Diabetes dataset from the Pima Indians Diabetes Database [40] is taken for the predictive analysis in this research work. The considered dataset was cleaned using the data preprocessing and data cleaning methodologies, then the resulted dataset has been considered for several number of experiments over different classification algorithms. The Pima Indians Diabetes Database contains the patient's details with diabetes status (Non-Diabetes and Diabetes). The vital patient's information is used to diagnose and predict the Diabetes Mellitus among the population.

The considered Diabetes Mellitus dataset contains 768 records. The dataset contains features of patients such as 1) Number of times pregnant, 2) Plasma glucose concentration a 2 hours in an oral glucose tolerance test, 3) Diastolic blood pressure in mmHg, 4) Triceps skin fold thickness in mm, 5) 2-Hour serum insulin in muU/ml, 6) Body mass index (weight in kg/(height in m)^2), 7) Diabetes pedigree function, 8) Age in years, and 9) Class variable (0 or 1) [40].

The data preprocessing and cleaning process (data imputation-mean technique) removes the missing and outlier data values from the dataset. The resulted dataset after preprocessing is reduced to 722 records with three required relevant features of patient details. In the dataset, there are 722 patient details, out of which 474 cases are in the class of ‘Non-Diabetes' and 248 cases are in the class of 'Diabetes' with 46 records are missing required essential values. Six numerical features from the dataset are taken as the input attributes, and one feature is considered as the output attribute. The patient's information is

Table.1 Sample record of cleaned dataset

#Pregnant	Glucose	BP	BMI	DPF	Age	Class
8	183	64	23.3	0.672	32	Diabetes
1	89	66	28.1	0.167	21	Non-Diabetes
0	137	40	43.1	2.288	33	Diabetes
5	116	74	25.6	0.201	30	Non-Diabetes
3	78	50	31	0.248	26	Diabetes
2	197	70	30.5	0.158	53	Diabetes
4	110	92	37.6	0.191	30	Non-Diabetes
10	168	74	38	0.537	34	Diabetes
10	139	80	27.1	1.441	57	Non-Diabetes
1	189	60	30.1	0.398	59	Diabetes

The patient features such as number of times pregnant, Plasma glucose concentration, Diastolic blood pressure, Body mass index (BMI), Diabetes pedigree function (DPF), Age is considered as input variables, and the Class is taken as the output variable. Figure.1 illustrates the population with respect to age. Figure.2 and Figure.3 depict the diabetes/non-diabetes population with respect to BMI and BP.

This research work analysis the prediction of diabetes of non-diabetes patients using different machine learning algorithm. Different classification models are applied to the diabetes dataset, and its performance in terms of accuracy, error rates, and area curves are evaluated. This work includes the evaluation of KNN, Naive Bayes, Extra Trees, Decision trees, Radial basis function, and Multilayer perceptron algorithms.

3.2 Machine Learning Algorithms

The KNN is one of the most straightforward supervised machine learning algorithms used to solve regression and classification problems. It assumes that similar things exist close. It assumes the similarity between the new data and the available data and assign the new data to the category that is most similar to the categories that are available. The distance between data points is calculated using Euclidean distance; The distance between two points (X1,Y1) and (X2, Y2) = $\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$, this gives the nearest neighbor [4].

Naive Bayes is one of the popular classification algorithms that are most widely used to get the base accuracy of the dataset. It assumes that all the variables present in the dataset are Navie (not correlated to each other). They are used in real-time prediction, multi-class prediction, spam filtering, sentimental analysis, text classification, a recommendation system, etc. Bayes rule determines the probability of the hypothesis. The formula used is: $(A \mid B) = \frac{P(B|A)\,P(A)}{P(B)}$; where P(A|B) refers to Posterior probability, P(A) refers to Prior probability, P(B) is marginal probability and P(B|A) refers to Likelihood probability [5].

Extra tree (Extremely Randomized tree) classifier is a type of ensemble learning method that is based on decision trees. It will work by creating a huge number of unpruned decision trees from the training dataset. In case of classification prediction is carried out using majority voting and in case of regression

Decision tree is a supervised machine learning technique that splits the data based on a parameter. Tree contains 2 entities, namely leaves and decision nodes. The leaves are the final outcomes and the data is split in the decision nodes. Selection of best attribute as the root node and sub-nodes is the main issue. Information gain and Gini index techniques can be used for attribute selection. Information gain is calculated using as follows:
$Informationgain = Entropy(S) - [(WeightedAvg) * Entropy(eachfeature)]$, Entropy metric is used to measure the impurity in the attribute. Entropy is calculated using this formula, $Entropy(S) = -P(yes)log2P(yes) - P(no)log2P(no)$.; S represents the number of samples, probability of yes and no are represented as P(yes) and P(no) respectively. It tells us the amount of information a feature provides about the class, with this information decision tree can be built. Gini index is calculated as follows: $GiniIndex = 1 - \sum_j P_j^2$. It is the measure of purity or impurity, used for decision tree creation in Classification and Regression Tree (CART) algorithm. An attribute with low gini index is preferred. [7]

A radial basis function will assign a real value to every input from its domain and the outcome will be an absolute value and cannot be negative (it's a measure of distance) $f(x) = f(||x||)$. Mainly they are used to approximate the functions. The sum $y(x) = \sum_{i=1}^{N} w_i \emptyset\left(\left|\left|x - x_i\right|\right|\right)$ represents radial basis function. These functions act as activation functions. [8].

Multi-Layered Perceptron is one of the simple, commonly used neural network model, referred to as "vanilla" neural network. It can be used for variety of applications like spam detection, image identification, election voting predictions and stock analysis [9].

4. Results And Discussions

This section summarizes the prediction results of the KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms. The k-fold cross-validation is one of the resampling procedure used in this work to validate the machine learning models on the limited data sample. In this work, the *'k'* value is chosen as 7. Therefore, it can be called as a 7-fold cross-validation resampling method. The 7-fold cross-validation method intends to reduce the bias of the prediction model [23].

4.1 Performance Evaluation

Typically, the performance of the machine learning prediction algorithms measured by using some metrics based on the classification algorithm. In this work, the prediction results are evaluated by using the metrics such as accuracy, mean square error (MSE), root mean square error (RMSE), Kappa score, confusion matrix, the receiver operating characteristic area under curve (ROC_AUC), classification performance indices, sensitivity, specificity, and f1 score values [18, 20, 23].

In this work, the prediction accuracy (that is, whether the patient is diabetic or non-diabetic) of different

on trees, radial basis function, and multilayer

perceptron algorithms) are determined. Each classification model has a different prediction accuracy based on its hyperparameters and a certain level of improvement over other prediction models. This work considers 70 % dataset for training and 30 % of the data samples for testing in classification algorithms. In this work, each model's accuracy is compared, and its prediction results are summarized in Table.2.

Table.2 Accuracy score of classifiers

S. No	Classifier	Accuracy	Kappa	AU_ROC
1.	K-Neighbors Classifier (KNN)	71.7241	0.3973	0.77
2.	Decision Tree (DT)	66.8965	0.3125	0.64
3.	Naive Bayes	77.2413	0.5112	0.85
4.	Extra Trees	72.4137	0.3770	0.72
5.	Radial Basis Function	68.2758	0.3213	0.70
6.	Multi-Layer Perceptron (MLP)	80.6890	0.5740	0.86

In Table.2, the classification algorithms such as KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms have the prediction accuracy of 71.7241, 66.8965, 77.2413, 72.4137, 68.2758, and 80.6890 respectively. Whereas, the multilayer perceptron algorithm predicts the diabetes cases (based on number of pregnant, glucose level, BP, BMI, DPF, age, diabetes class) more accurately than the other algorithms.

The multilayer perceptron is the feed forward artificial neural network. The MLP processes the diabetes dataset using non-linear activation functions with three layers of neurons such as input, hidden, and output layer. The dataset was tested with several neuron values and classifies the given dataset into two forms of classes as diabetes and non-diabetes patients with reduced errors. It works by mapping the given weighted inputs to the output of each neuron among the test data and training data. For each data points based on the error values, the testing datasets are classified. Such that the multilayer perceptron algorithm produces a higher classification rate than the other algorithms.

Table.3 Error metrics of classifiers

S. No	Classifier	MSE	RMSE
1.	K-Neighbors Classifier (KNN)	0.2827	0.5317
2.	Decision Tree (DT)	0.3310	0.5753
3.	Naive Bayes	0.2275	0.4770
4.	Extra Trees	0.2758	0.5252
5.	Radial Basis Function	0.3172	0.5632
6.	Multi-Layer Perceptron (MLP)	0.1931	0.4394

The Cohen's kappa score for multilayer perceptron algorithm is also high than other algorithms. Cohen's kappa score estimates the consistency of the classification algorithm based on its predictions. Figure.4 depicts the accuracy scores of different classification algorithms. From Figure.4, we can clearly see that the multilayer perceptron algorithm has the highest accuracy of 80.68. The multilayer perceptron algorithm has 3.4 to 13.8 % of improved accuracy as compared to KNN, DT, NB, ET, and RBF algorithms. The MLP algorithm works by classifying the data point of the diabetes dataset based on the similarity.

Table.3 presents the performance error metrics of the various machine learning algorithms. The error metrics mean square error and root mean square error values for each algorithm is evaluated. The KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms have the MSE error rate as 0.2827, 0.3310, 0.2275, 0.2758, 0.3172, and 0.1931 respectively. As per Figure.5, the multilayer perceptron algorithm produces the lowest error rate as 0.1931 on the prediction of accurate diabetes cases than the other algorithm. The multilayer perceptron algorithm classifies the testing dataset by output mapping between the new (testing) instance (x_i) and the existing (training) instance (y_j). Therefore, it results in lower error rates.

Similarly, the MLP's RMSE error rate also very low (0.43) as compared to the error rates of KNN (0.53), DT (0.57), NB (0.47), ET (0.52) and RBF (0.56) algorithms as shown in Table.3. As depicted in Figure.6, the multilayer perceptron algorithm produces the highest consistency (Kappa score) among the evaluated algorithms as 0.57. The KNN, Naive Bayes, extra trees, decision trees, and radial basis function has 0.3973, 0.3125, 0.5112, 0.3770, and 0.3213 respectively.

Figure.7 illustrates the confusion matrix without normalization of the multilayer perceptron algorithm. In all classification algorithm, 30 % of the data samples are taken for testing with the 70 % training dataset. In this Figure, the x-axis represents the percentage of predicted values and the y-axis represents the percentage of true values. It can be seen that the multilayer perceptron algorithm predicts 79 % (true positive) of the diabetes cases correctly, with 12 % (false positive) of misclassification. Similarly, in Figure.8 the confusion matrix with normalization is depicted.

Figure.9 is the pictorial representation between the false positive rate and true positive rate in the form ROC area under the curve. The multilayer perceptron algorithm produces the highest value of 0.86 as compared with KNN, Naive Bayes, extra trees, decision trees, and radial basis function algorithms. Figure.10 summarizes the performance metrics such as precision, recall, and confusion matrix of the multilayer perceptron algorithm.

The multilayer perceptron algorithm produces the precision (true positive rate) value of 0.82 for the non-diabetes cases and 0.78 for the deceased cases. The recall values for the non-diabetes and diabetes cases are 0.89 and 0.67, respectively. Further, the F1 score for non-diabetes and diabetes cases is 0.85 and 0.75, respectively.

5. Conclusion

It is worth of studying much essential to accurately predicting and diagnosing any disease by using machine learning. This work explores different machine learning algorithms and their performances on the diabetes dataset. The results of the KNN, Naive Bayes, extra trees, decision trees, radial basis function, and machine learning algorithms. All above algorithms are experimented with the prediction accuracy, MSE, RMSE, Kappa score, AUROC, precision, recall, and F1-score. The results are show that MLP has a better performance with the Prima diabetes dataset. In addition, by comparing with the results has better AUROC as 86 % among KNN, Naive

Bayes, extra trees, decision trees, and radial basis function. Thus, the results suggest that the MLP algorithm can able to diagnose and classify the diabetic patients.

Declarations

Competing interests:

The authors declare no competing interests.

References

1. https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview accessed on 12/12/2020.
2. https:// accessed on 16/12/2020.
3. https:// accessed on 17/12/2020.
4. https:// accessed on 19/12/2020.
5. https://acadgild.com/blog/naive-bayesian-model accessed on 19/12/2020.
6. https://machinelearningmastery.com/extra-trees-ensemble-with-python/ accessed on 18/12/2020.
7. https://-,Blog,namely%20decision%20nodes%20and%20leaves accessed on 19/12/2020.
8. https://wiki.pathmind.com/radial-basis-function-network-rbf accessed on 19/12/2020.
9. https:// accessed on 19/12/2020.
10. Adel Al-Zebari, Abdulkadir Sengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection", 2019 1st International Informatics and Software Engineering Conference (UBMYK).
11. G. A. Pethunachiyar, "Classification Of Diabetes Patients Using Kernel Based Support Vector Machines", 2020 International Conference on Computer Communication and Informatics (ICCCI – 2020), Jan. 22–24, 2020, Coimbatore, INDIA.
12. Pahulpreet Singh Kohli, Shriya Arora, "Application of Machine Learning in Disease Prediction", 2018 4th International Conference on Computing Communication and Automation (ICCCA).
13. Samrat Kumar Dey, Ashraf Hossain, Md. Mahbubur Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm", 2018 21st International Conference of Computer and Information Technology (ICCIT), 21–23 December, 2018.
14. Usha Ruby, A., Theerthagiri, P., Jeena Jacob, I., Vamsidhar, Y., “Binary cross entropy with deep learning technique for image classification International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(4), pp. 5393–5397
15. S. Hari Krishnan, P. Vinupritha and D. Kathirvelu, "Non-Invasive Glucose Monitoring using Machine Learning", International Conference on Communication and Signal Processing, July 28–30, 2020, India.

16. M.Shanthi, RamalathaMarimuthu, S.N.Shivapriya, R.Navaneethakrishnan, "Diagnosis of Diabetes using an Extreme Learning Machine Algorithm based Model", 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST).
17. Sajratul Yakin Rubaiat, Md Monibor Rahman, Md.Kamrul Hasan, "Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection", 2018 International Conference on Innovation in Engineering and Technology (ICIET).
18. Prasannavenkatesan Theerthagiri, I.Jeena Jacob, A.Usha Ruby et al. Prediction of COVID-19 Possibilities using KNN Classification Algorithm, 03 November 2020, PREPRINT (Version 2) available at Research Square.
19. Ali Mohebbi, Tinna B. Aradottir, Alexander R. Johansen, "A Deep Learning Approach to Adherence Detection for Type 2 Diabetics", 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).
20. Theerthagiri, P., 2020. Forecasting Hyponatremia in hospitalized patients Using Multilayer Perceptron and Multivariate Linear Regression Techniques. arXiv preprint arXiv:2007.15554
21. Maham Jahangir, Hammad Afzal, Mehreen Ahmed, Khawar Khurshid, Raheel Nawaz, "An Expert System for Diabetes Prediction using AutoTuned Multi-Layer Perceptron", Intelligent Systems Conference 20177-8 September 2017 | London, UK.
22. Sidong Wei, Xuejiao Zhao, Chunyan Miao, "A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification", 2018 IEEE 4th World Forum on Internet of Things (WF-IoT).
23. Prasannavenkatesan Theerthagiri. Probable Forecasting of Epidemic COVID-19 in Using COCUDE Model for the State of Tamilnadu, India, 16 July 2020, PREPRINT (Version 1) available at Research Square.

Figures

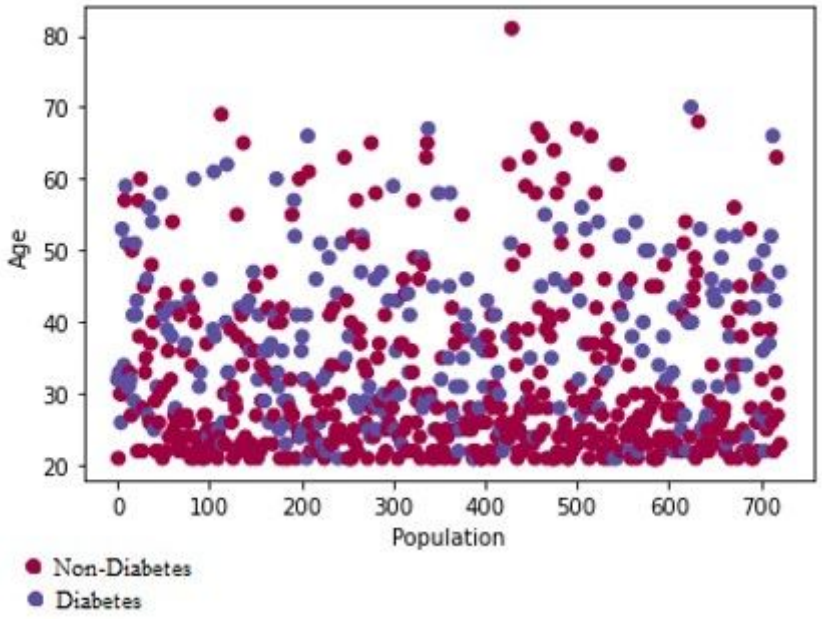

Figure 1

Population vs Age

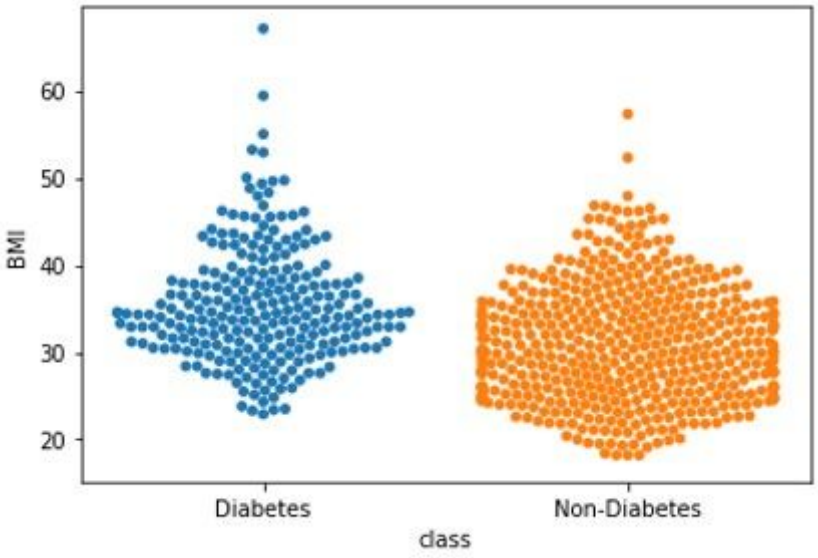

Figure 2

BMI vs Diabetes Classes

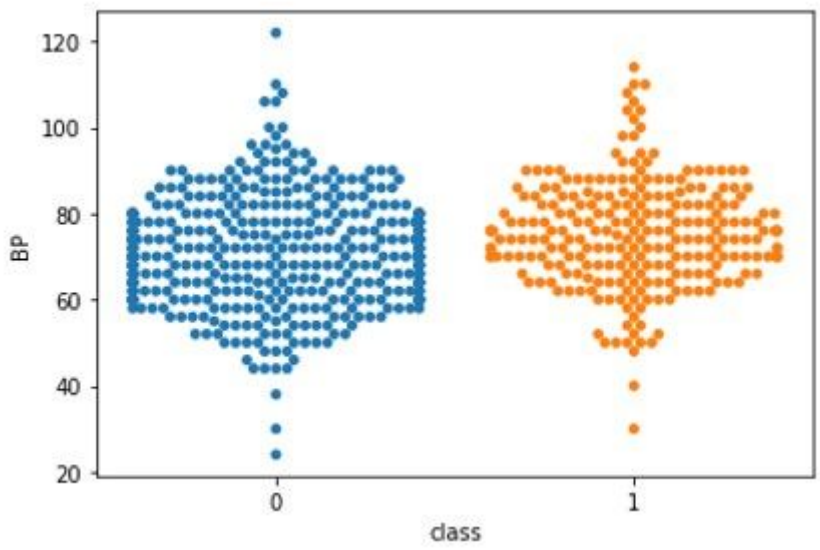

Figure 3

BP vs Diabetes Classes

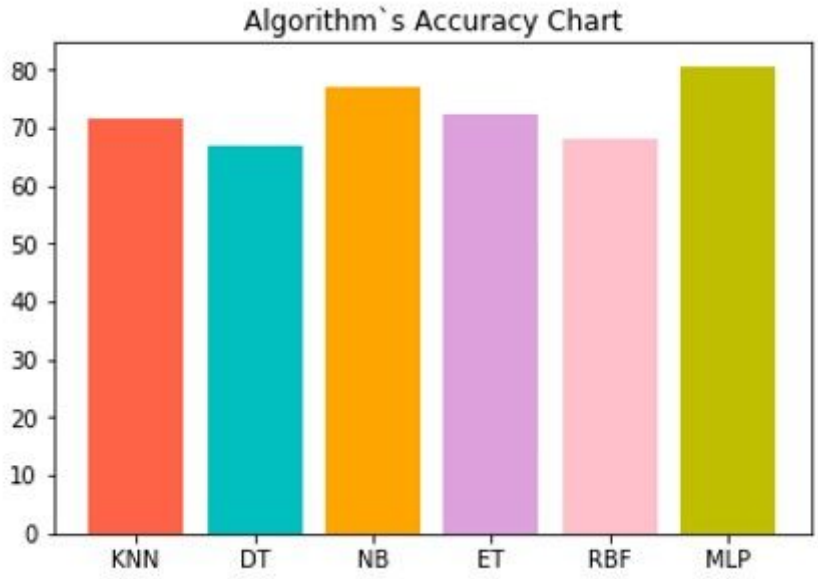

Figure 4

Prediction Accuracy

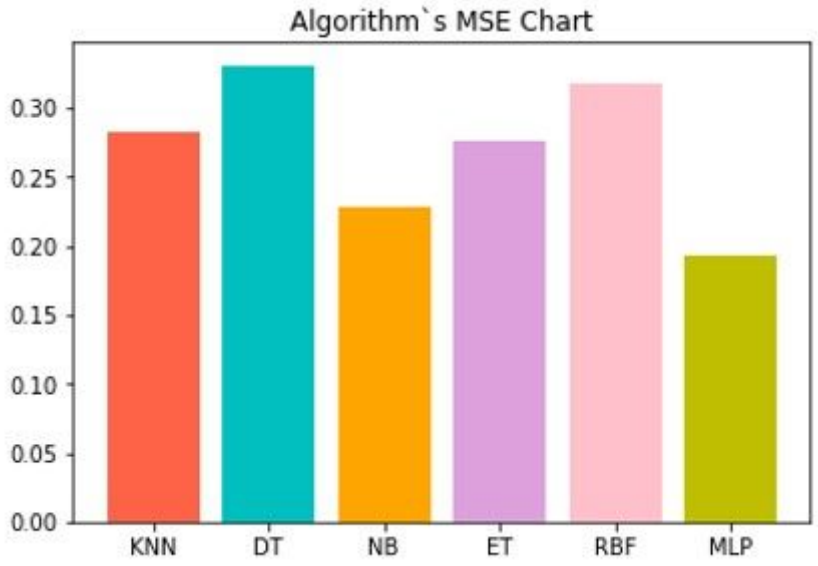

Figure 5

MSE rates

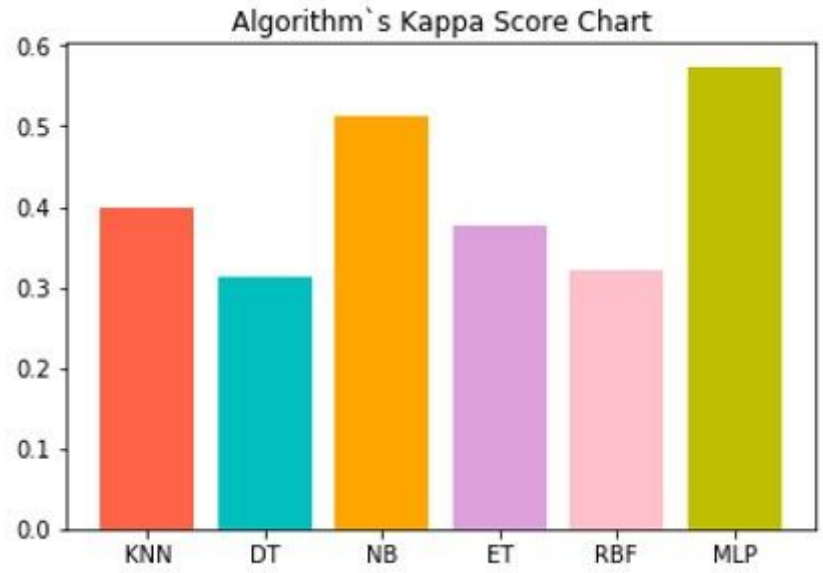

Figure 6

Cohen’s kappa scores

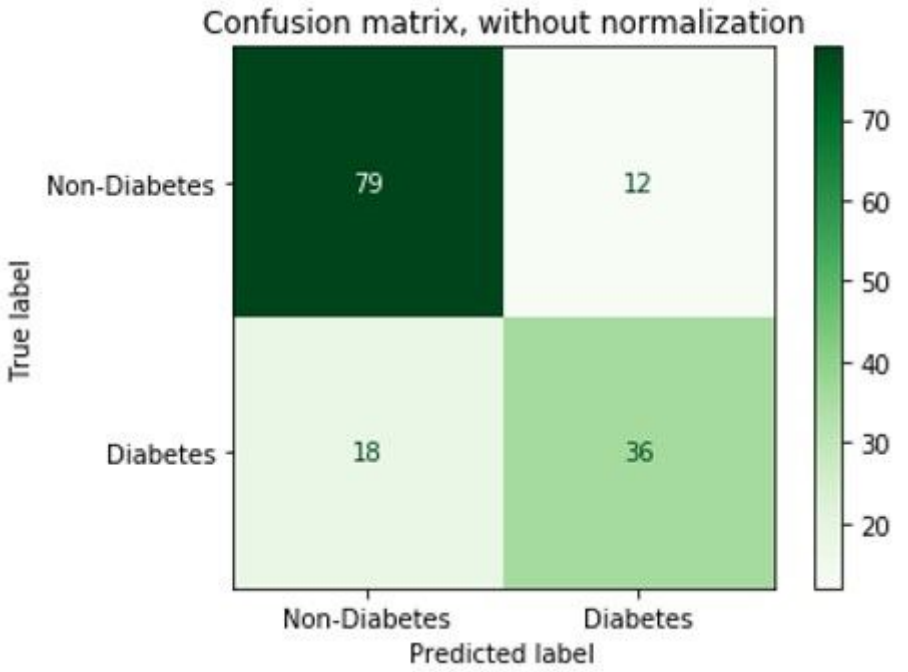

Figure 7

Normalized Confusion matrix

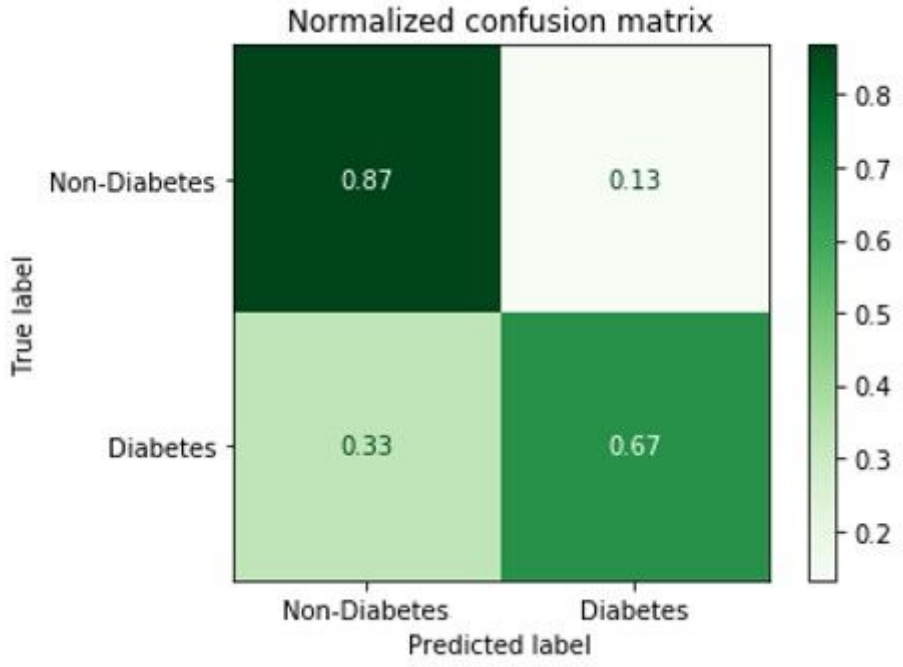

Figure 8

Confusion matrix (no normalization)

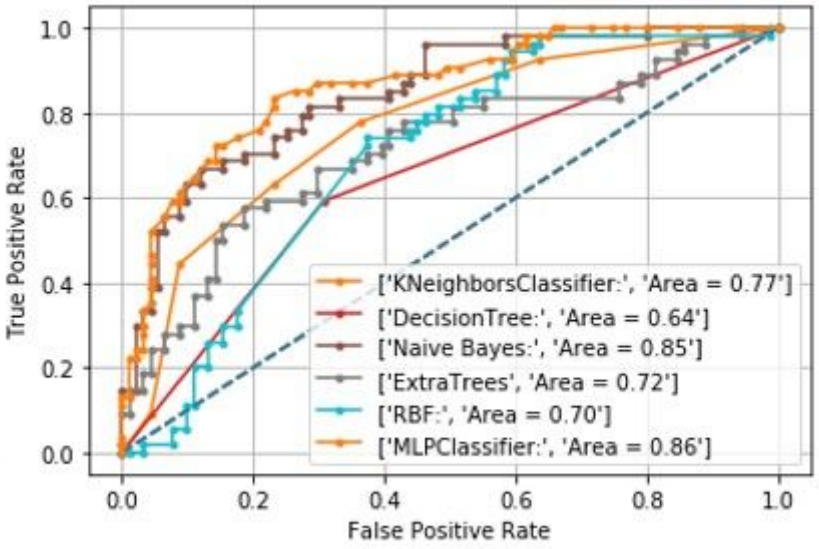

Figure 9

ROC_AUC Curve

```
MSE: 0.19310344827586207
RMSE: 0.4394353744020411
MAE: 0.19310344827586207
R2 score: 0.17
Kappa_score: 0.5740663029794377
Accuracy: 80.6896551724138

Classification Report:

              precision    recall  f1-score   support

Non-Diabetes       0.82      0.89      0.85        91
    Diabetes       0.78      0.67      0.72        54

    accuracy                           0.81       145
   macro avg       0.80      0.78      0.79       145
weighted avg       0.80      0.81      0.80       145

Confusion matrix, without normalization
[[79 12]
 [18 36]]
Normalized confusion matrix
[[0.86813187 0.13186813]
 [0.33333333 0.66666667]]
```

Figure 10

Summary of Performance metrics scores of MLP algorithm