Early Detection of Type 2 Diabetes using supervised machine learning

ISSN (e) 2520-7393
ISSN (p) 2521-5027
Received on 3rd Mar 2021
Revised on 10th Mar, 2021
www.estirj.com

Abdul Qayyum[1], Dr Shahnawaz Talpur[2], Dr. Moazzam Jawaid

[1]qayyum.laghari20@gmail.com,[2]mirshan35@gmail.com,[3]moazzam.jawaid@faculty.muet.edu.pk

[1]Department of Computer Information Engineering Mehran University of Engineering Technology Jamshoro
[2,3]Department of Computer Systems Engineering Mehran University of Engineering Technology Jamshoro,

Abstract: Diabetes mellitus is a disease that carries the risk of stroke, renal failure, kidney complications, cardiac disease, peripheral artery disease, and death. There are 3 primary diabetes types: type 1, type 2, and gestational diabetes. type 2 diabetes is the most common and accounts for 90% to 95% of all cases. Type 2 diabetes can be cured if it is predicted timely. Type 2 diabetes is a condition that can be anticipated based on factors such as glucose level, Body mass index etc. The advancement in machine learning proved to ground breaking in the field of medicines a lot of prediction models are in production use for various fatal diseases. In this research early prediction of type 2 diabetes is performed using features such as Pregnancies, Glucose Level Blood Pressure, Skin Thickness, Body Mass Index, Diabetes Pedigree Function and Age . The machine learning algorithms used in this research are Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Decision Tree and Random Forest. The evaluation parameters used in this study are Accuracy, precision, recall, ROC AUC and precision recall curve. The results of evaluation proved logistic regression the best model with accuracy, precision and recall of 80.2%, 75% and 58%.

Keywords: *Diabetes, Machine Learning, PIDD, Type 2 Diabetes and Supervised machine learning*

1. Introduction

Diabetes, also referred to as diabetes mellitus (DM), is a combination of metabolic complications detected over a prolonged period of time by elevated blood glucose levels. Excessive urination, constantly feeling thirsty and elevated appetite are indicators of high glucose [1]. It is estimated that the overall number of people with diabetes will increase from 171 million in 2000 to 366 million in 2030 [2]. The statistics suggest that the vast majority of patients with diabetes suffer from type 2 diabetes [3]. According to [4] More than half of middle-aged and older people do not have a good understanding of disease prevention. It is shown the efficacy of different preventive interventions such as dietary improvements, weight loss, daily physical activity and pharmacological agents in significantly reducing the growth of open Type 2 Diabetes.

Therefore, development of a sound early diabetes type 2 prediction system would offer aid to middle-aged and elderly people to avoid diabetes. A new and commonly accepted approach to health risk prediction is machine learning. Various algorithms for machine learning have been suggested, ranging from traditional to more sophisticated approaches to ensemble machine learning. The majority of machine learning models are based on the existence of biomarkers. The blood glucose level, for example, is a biomarker widely adopted in several machine-learning models for the purpose of prediction of type 2 diabetes [1]. In this study the early diabetes type 2 prediction models are created with the help of various supervised classical machine learning algorithms such as logistic regression or decision tree and supervised ensemble learning algorithms like Random Forest. All early diabetes prediction models are evaluated against each other to see which one is most suitable for the use case.

2. Related Work

Globally, diabetes and its complications are significant causes of premature mortality. In 2017, 451 million people worldwide were estimated to have diabetes, and this number is expected to increase by 35 percent to 693 million by 2045 [5]. Internationally, predicting the risk of diabetes in adults has become the primary focus [6]. Numerous researchers have proposed different methods, such as machine learning and data mining, to help predict diabetes [7]. Initial models for the purpose of prediction of diabetes were based on statistical algorithms such as linear regression [8]. For diabetes prediction, various research studies have used the Pima Indians Diabetes Dataset (PIDD) [9][10].

Three machine learning methods i.e. decision tree (DT), naïve-based (NB) and support vector machine (SVM) were used on PIDD to predict diabetes in [9] A naïve Bayes classifier was discovered to be 76.3% accurate. In [10] researchers have also performed comparison of various machine learning algorithms for the purpose of diabetes prediction as in [11] Four classification techniques were studied: Decision tree, ANN, logistic regression, and naïve

Corresponding author Email address: qayyum.laghari20@gmail.com

Bayes. For the purpose of feature reduction numerous techniques are used as in [12] Principal Component Analysis (PCA) is applied for feature reduction and random forest is used to train the model to make predictions about diabetes. Our research focuses on creating models that can perform the early prediction of diabetes type 2 using features such as glucose level and body mass index etc. Using a standard Pima Indian Diabetes Dataset is used for training and evaluating our models, based on evaluations the decision will be made regarding the performance of all the models based on parameters such as accuracy precision, ROC and Precision recall curve.

3. Dataset:

The Pima Indian Diabetes Dataset (PIDD) comprises data regarding 768 patients. Every record is made up of eight attributes, all of which are numbers. These documents contain personal health information as well as the outcomes of medical tests. The features in PIDD are Number of times pregnant, Plasma glucose concentration at 2h in an oral glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2-h serum insulin, Body mass index, Diabetes pedigree function, Age, Class variable. The class is basically the label i.e. it represents if a person is positive for diabetes type or not. Fig 1 represents the number of records per class i.e how many of them are positive for diabetes type and how many are negative.

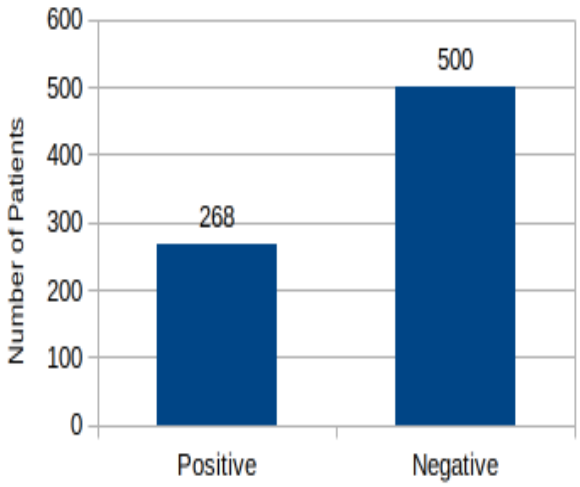

Figure.1. Patient Frequency Per class

From fig 1 it can be seen that there 268 people in the dataset which are labeled as diabetes type 2 positive and 500 people are type 2 negative.

4. Methodology

4.1 System model

In this section methodology employed in the research is discussed. Fig 2 summarizes all the steps used in the research. Each step of research for the is discussed as follow:

Dataset Collection:

For this research a Pima Indian Diabetes Dataset. The dataset contain 8 Features each of the feature is number class is used as label it shows that whether a record is a tested positive or negative

Dataset Preprocessing:

After Dataset collection the next step is to clean it from any missing values, outliers and any data which is scaled out of proportion. The dataset doesn't contain any missing values but all features have their own specific scale. To give a clear understanding of data it is must to scale them up evenly. Here in this step all features are scaled uniformly. Fig 2 Represents the statistics of dataset before being scaled and Fig 3 shows the dataset statistics after being scaled

count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure.2. Dataset Statistics Before Scaling

count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure.3. Dataset Statistics After Scaling

Splitting the Dataset:

After we have the dataset ready to be fed to the machine learning algorithm we will split the dataset into train and test sets so that training and evaluation can be performed fairly.

Apply Machine Algorithms:

After splitting the dataset the training set is fed to the machine learning algorithm so that machine learning model can be created. In this study 5 machine learning algorithms namely logistic Regression, SVM, Decision Tree, Random Forest and are used so after splitting the dataset each algorithm is given the same training set to get trained.

Evaluate the Model:

The test set is used to evaluate the model created using all 5 algorithms. Each model is evaluated using accuracy, precision, recall and F1 Score. All these are standard classification metrics. Along with these parameters Receiver operating characteristic (ROC) Curve and Precision Recall Curve are also used to better understand the evaluation of the model. At the end of this step it can be established which of the models performed well and which didn't.

All the steps in the experiments are performed using python programming language. The libraries used in this research are numpy, pandas, sklearn. Numpy is used to get insight into data from statistics point of view, Pandas used to deal with dataframes, Sklearn is used to perform all machine learning related operations, Matplotlib is used to create graphs and give visual understanding.

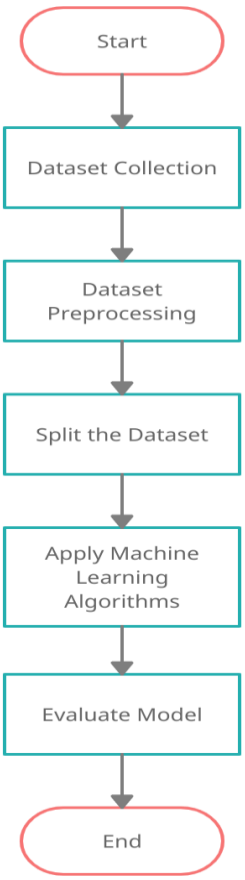

Figure.4. Patient Frequency Per class

4.2. Description of the algorithms used

4.1.1. Decision Tree:

Decision trees (DTs) are tree-structured classifiers that sort instances based on feature values. A node in a decision tree represents a function of an instance to be classified, and each branch a value. The root node serves as the starting point for classifying instances, and they are arranged based on their feature values. Most decision tree classifiers use post-pruning techniques, since they are pruned by using a testing set. Each node can be deleted and allocated the most frequently encountered training instance class [13].

4.1.2. Logistic Regression:

Logistic Regression is a classification algorithm. It classifies based on class, but rather than using a one-stage multinomial logistic regression model, it utilizes a specific multinomial logistic regression model with just one estimator. Logistic regression almost always shows where the distinction between the categories is, and the classification probabilities depend on how far away from the threshold a given object is. Logistic regression is similar to Ordinary Least Squares (OLS) regression, but with the exception that it uses logistic functions in the model formula. On the other hand, in logistic regression, the result of prediction is a binary outcome (prediction results in a yes/no outcome). Applied statistics and discrete data analysis and machine learning are the areas in which logistic regression is widely used [13].

4.1.3. Random Forest:

The Random Forest (RF) is a machine learning approach that is used to address regression and classification related problems. Increasing the number of trees in the category and settling on class form through a vote has greatly increased accuracy rate. For example, arbitrary vectors are assembled so these ensembles can develop. A random vector produces each tree. The classification trees in RF represent the following: Classification problems are addressed by performing class research on the production of trees. The plurality vote of the class decides the RF result. Adding more trees to the RF does not lead to overfitting, but generalization error leads to a limiting value [14].

4.1.4. Support Vector Machine:

Support Vector Machine (SVM) models are similar to multilayer perceptron neural networks that employ multiple layers. The core concept of SVMs revolves around the idea of a margin, each side of a hyperplane that divides two groups of data. An upper bound on the predicted generalization error can be found by maximizing the margin and thereby establishing the maximum possible separations in between separating hyperplanes and the occurrences on each side of it. It is one of the most common classification techniques for the use case in hand [13].

4.1.5. K-Nearest Neighbor:

One of the most common classification algorithms is the k-nearest neighbors (kNN) classifier. It is a non-parametric process, and the classifier uses a similarity function to pick the correct test sample. Because of the usefulness of the k-Nearest Neighbor algorithm, it is now one of the most commonly used machine learning algorithms. The kNN classifier operates in two stages. First, it determines the nearest k neighbors for an unknown set, and then differentiates between the unknown sample's category and the k nearest neighbors' categories based on the maximum likelihood (maximum frequency) of the k nearest neighbors.

5. Results

5.1. Evaluation Parameters

5.1.1. True Positive (TP):

True positive is the prediction that was originally positive and is predicted positive as well, in this use case true positive is if patient has type 2 diabetes and our model predicts the same

5.1.2. True Negative (TN):

True negative is the prediction where the model correctly predicts the patient doesn't have type 2 diabetes.

5.1.3. False Positive (FP):

If a model predicts a negative label on originally positive data it is called false negative. As in this use case, if a model predicts a diabetes positive patient as non-diabetic that prediction will be called false positive.

5.1.4. False Negative (FN):

If a model predicts a non-diabetic patient as diabetic that type of prediction which is giving false results on negative data is called as false negative.

5.1.5. Accuracy:

The percentage of correct predictions for the test data is known as the accuracy. By dividing the number of correctly predicted labels by the overall number of the predictions, it is possible to determine how likely it is to correctly predict future events. It is represented by eq(1). Mathematically accuracy is defined as the sum of TP and TN divided by the sum of TP, TN,FP and FN [15].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

5.1.6. Precision:

Precision is measured as the percentage of valid examples (true positives) compared to all of the instances that were expected to adhere to a specific group i.e. it is ratio of TP to sum of TP and FP [15] it is represented by equation 2

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

5.1.7. Recall:

The term recall is measured as the fraction of instances that were predicted to adhere to a label (i.e., where the classifier guessed right) relative to all of the examples that genuinely belong in the class. It is the ratio of TP to the sum of TP and FN [15] it is represented by equation 3

$$Precision = \frac{TP}{TP+FN} \quad (3)$$

5.1.8.Confusion Matrix:

Confusion matrix is one of the ways to look at the performance of a classifier. It can be created for binary classification and multiclass classification. For binary classification it is a 2 by 2 matrix, the first element of first row represents TP, second element of first row is FP, first element of second row is FN and 2 element of second row is TN. Fig 5 gives graphical understanding of confusion matrix [16].

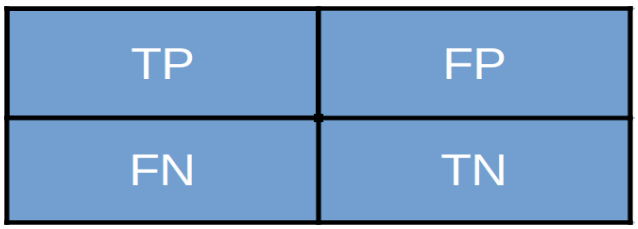

Figure.5. Confusion Matrix

5.1.9. ROC Curve:

A helpful method when evaluating the likelihood of a binary result is the ROC curve, also known as the Receiver Operating Characteristic (ROC) curve. For a variety of different possible threshold values between 0.0 and 1.0, the plot shows the false positive rate on x-axis versus the true positive rate on y-axis. In other words, it portrays the false alarm rate (FAR) vs the hit-rate [15].

5.1.10 Precision Recall Curve:

In the precision recall curve, the positive predictive instance precision is plotted on y axis against the true positive rate i.e. Recall on the x axis for multiple thresholds. These metrics can also be used to test binary classification models in advanced machine learning. In this research it will plot the precision and recall for different thresholds to give understanding that which model is performing well [15].

5.2. Accuracy, Precision Recall for all models:

The evaluation of all models is done based on numerous factors. The first three factors are accuracy, precision and recall and table 1 represents accuracy, precision and recall for the classification models. Based on the table 1 the best performing model is KNN with highest accuracy and recall of 80.7% and 64.5%. The second-best model is Random forest which has an accuracy of 80.2%, precision of 75% and recall of 59.6%. The next best model is logistic regression followed by SVM and Decision Tree each have accuracy of 80.7%, 77.8% and 71.3% respectively.

Table 1: Accuracy, Precision Recall

Algorithm	**Accuracy %**	**Precision %**	**Recall %**
Decision Tree	71.3	55.5	56.1
SVM	77.6	71.1	51.6
KNN	80.7	72.7	64.5
Random Forest	80.2	74	59.6
Logistic Regression	80.2	75	58.1

5.3. Confusion matrix for all the algorithms:

Confusion matrix as discussed in the section of evaluation parameter is a performance metric for classification. Its dimension depends on the number of classes. Here in this study there are 2 class positive and negative diabetes patients. Fig 6, 7, 8, 9 and 10 represents the confusion matrix for decision tree, SVM, KNN Random Forest and Logistic Regression respectively. Each confusion matrix represents true positive, true negative, false positive and false negative values for each model. For instance in Fig 8 TP are 155 FP are 15 TN 22 and FN are 40. Similarly TP, TN, FP and FN can be perceived for any model by looking at each element of the matrix.

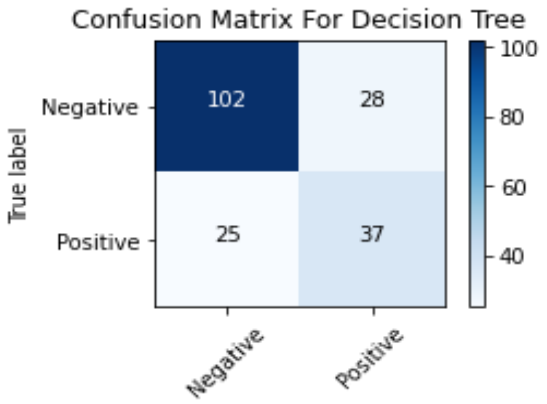

Figure.6. Confusion Matrix for Decision Tree

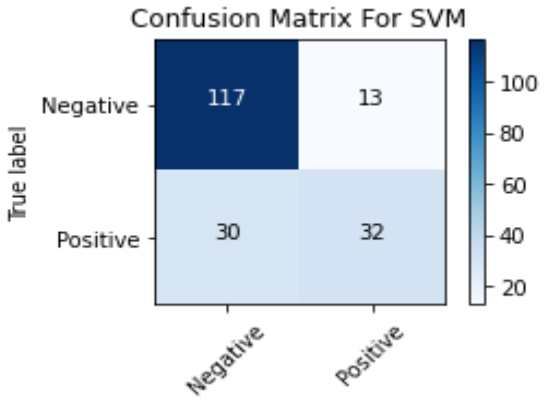

Figure.7. Confusion Matrix for SVM

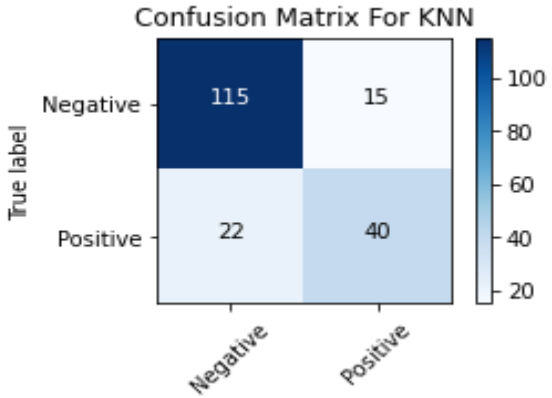

Figure.8. Confusion Matrix for KNN

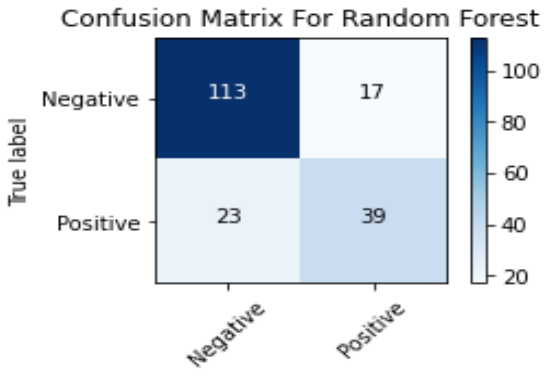

Figure.9. Confusion Matrix for Random Forest

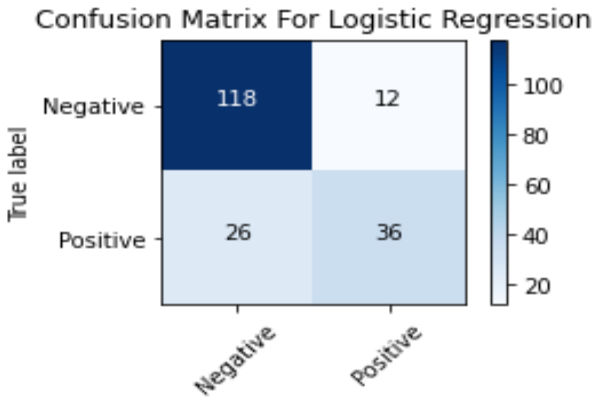

Figure.10. Confusion Matrix for Logistic Regression

.

5.4. Area Under ROC Curve and Precision Recall Curve:

Table 2 contains 2 column AUC ROC and AUC. AUC ROC represents the area under the curve for ROC curve and AUC represents the area under the curve for precision recall curve based on the table 2 it can be extracted that logistic regression has the highest area under the curve for both ROC and Precision and Recall. AUC ROC value for Decision Tree, SVM, KNN, Random Forest and Logistic Regression are 77.3, 85.3, 77.9, 86.3, 86.6 respectively and AUC for Decision Tree, SVM, KNN, Random Forest and Logistic Regression are 59.7, 65.8, 57.1, 65.4, 67 respectively.

Table 2 AUC for ROC and Precision Recall Curve

Algorithm	**AUC ROC %**	**AUC %**
Decision Tree	77.3	59.7
SVM	85.3	65.8
KNN	77.9	57.1
Random Forest	86.3	65.4
Logistic Regression	86.6	67

5.5. Precision and Recall Curve for all the models:

Fig 11 and Fig 12 represents the ROC and Precision Recall Curve for all the models. By looking at the graph it can be interpreted that which of the model is covering the most area under that curve and it is evident from auc values and fig 11 and 12 that the logistic regression model takes most area under the curve and proved to be the best model.

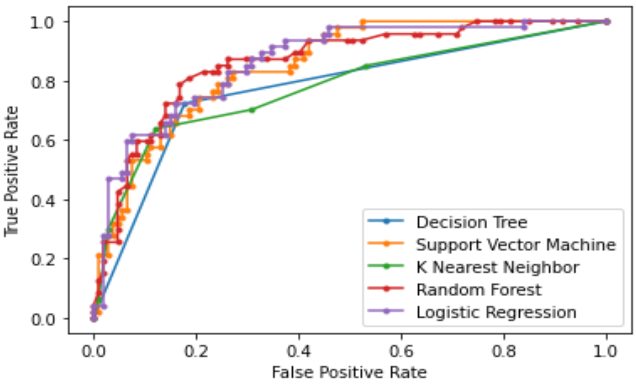

Figure.11. ROC Curve for all the models

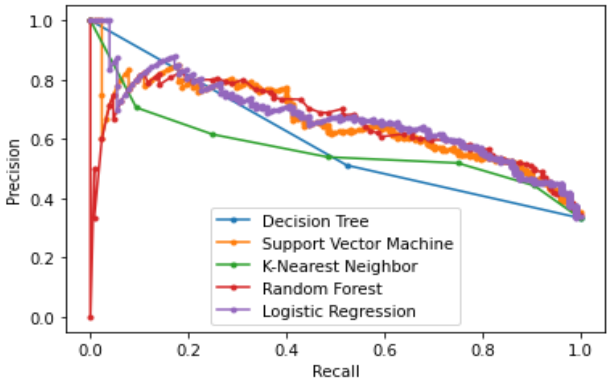

Figure.12. Precision Recall Curve for all the models

6. Conclusion:

Diabetes is one of the most fatal diseases. There are different types of diabetes. Diabetes type 2 is the most common form of diabetes cause of numerous fatalities around the globe. Early detection of diabetes can help fight it and can save many precious lives. In this study machine learning is employed for early prediction of type 2 diabetes. The dataset used in the study is PIMA INDIAN DIABETES DATASET (PIDD). Five machine learning algorithms are applied on training data and classifiers are created. All the classifiers are evaluated using accuracy precision recall confusion matrix and ROC and Precision Recall Curve. Based on the evaluation parameters KNN performed well in terms of accuracy, precision and recall and logistic regression was at number 2 but based on other features logistic regression is considered the best model so keeping all the parameters in check the overall best performing model is logistic regression. This model will be used to create a web UI where anyone can input data to get the prediction regarding type 2 diabetes. This type of system has the ability to work as an early warning system and is a big leap toward machine learning in medical science.

References

[1] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," Procedia Computer Science, vol. 167, pp. 706–716, 2020.

[2] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030," Diabetes Care, vol. 27, no. 5, pp. 1047–1053, 2004.

[3] Y. Liu, A. G. Wheaton, D. P. Chapman, T. J. Cunningham, H. Lu, and J. B. Croft, "Prevalence of Healthy Sleep Duration among Adults — United States, 2014," MMWR. Morbidity and Mortality Weekly Report, vol. 65, no. 6, pp. 137–141, 2016.

[4] B. Farran, R. Alwotayan, H. Alkandari, D. Al-Abdulrazzaq, A. Channanath, and T. A. Thanaraj, "Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait," Frontiers in Endocrinology, vol. 10, 2019.

[5] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. Ohlrogge, and B. Malanda, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," Diabetes Research and Clinical Practice, vol. 138, pp. 271–281, 2018.

[6] L. Zhang, X. Shang, S. Sreedharan, X. Yan, J. Liu, S. Keel, J. Wu, W. Peng, and M. He, "Predicting the Development of Type 2 Diabetes in a Large Australian Cohort Using Machine-Learning Techniques: Longitudinal Survey Study," JMIR Medical Informatics, vol. 8, no. 7, 2020.

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

[8] W. R. Farwell, J. M. Gaziano, E. P. Norkus, and H. D. Sesso, "The relationship between total plasma carotenoids and risk factors for chronic disease among middle-aged and older men," British Journal of Nutrition, vol. 100, no. 4, pp. 883–889, 2008.

[9] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[10] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," The Kaohsiung Journal of Medical Sciences, vol. 29, no. 2, pp. 93–99, 2013.

[11] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," Procedia Computer Science, vol. 82, pp. 115–121, 2016.

[12] Pradhan, M. A. "A genetic programming approach for detection of diabetes." International Journal of Computational Engineering Research 2.6 : 91-94 2012.

[13] O. F.y, A. J.e.t, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," International Journal of Computer Trends and Technology, vol. 48, no. 3, pp. 128–138, 2017

[14] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," 2018 15th Learning and Technology Conference (L&T), 2018.

[15] A. Tharwat, "Classification assessment methods," Applied Computing and Informatics, vol. ahead-of-print, no. ahead-of-print, 2020.

[16] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1806–1819, 2017.

About Authors

Abdul Qayyum, Student of ME at Computer Information Engineering, Mehran University of Engineering Technology Jamshoro.

Dr. Shahnawaz Talpur, PhD ,Dept of Computer Systems Engineering Mehran UET Jamshoro.

Dr. Moazzam Jawaid Dept of Computer Systems Engineering Mehran UET Jamshoro.