This work is licensed under a Creative Commons Attribution 4.0 International License Early-Stage Diabetes Risk Prediction: A Comparative Analysis of Classification Algorithms

Article *in* IARJSET · February 2021

DOI: 10.17148/IARJSET.2021.8228

CITATIONS
8

READS
512

2 authors:

Apratim Sadhu
Chandigarh University
7 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Abhimanyu Jadli
Chandigarh University
2 PUBLICATIONS **22** CITATIONS
SEE PROFILE

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

Early-Stage Diabetes Risk Prediction: A Comparative Analysis of Classification Algorithms

Apratim Sadhu[1], Abhimanyu Jadli[2]

Second Year B.E, Department of CSE, Chandigarh University, Mohali, India[1,2]

Abstract: Diabetes is a metabolic disease that results in high blood sugar. The hormone imbalance is one of the main reasons for this metabolic disorder. The specific hormone affected is insulin, the one which regulates sugar in the blood. The disease causes the patient's body either to not make sufficient insulin or can't efficiently and effectively use the insulin made. The same disease also becomes the reason for the death of 1.6 million people every year. Despite our medical development and natural endurance the cases of diabetes have risen in recent decades. We are in the age of information, we have a surplus amount of data to feed our data-hungry machine learning algorithms. The medical data of diabetic patients show a similar pattern which makes it possible to predict diabetes in an early stage. Thus, contributing to fighting back against the disease and for goodwill. The paper presented seven machine learning classifiers that have been implemented on the early-stage risk prediction diabetes dataset and three different evaluation metrics i.e. classification accuracy, F-score and ROC value are used to evaluate the performance of the algorithms on the validation set. The results presented brings out clear results in favour of the Random Forest on the average-sized dataset.

Keywords: Decision Tree Classifier Random Forest Classifier, Support Vector Machines, Multi-Layer Perceptron, K-Nearest Neighbours Classifier, Naïve Bayes Classifier, Logistic Regression, Binary Classification, Diabetes.

I. INTRODUCTION

Diabetes is a chronic disease that is caused due to the inability of the production of insulin by the body. The lack of insulin in the body leads to raised blood glucose level. According to studies conducted by the International Diabetes Foundation(IDF) [1], approximately 463 million adults have diabetes and by 2045 this number will rise to 700 million. Type 2 diabetes accounts for 90% of the total number of diabetes cases, while the remaining 10% of cases are mainly due to Type 1 diabetes mellitus and gestational diabetes. With the increasing proportion, 374 million people are at risk of developing type 2 diabetes. This disease has caused 4.2 million deaths around the globe. According to WHO estimates [2], diabetes mellitus is the ninth leading cause of deaths in the world. Diabetes has severe health impacts such as kidney failure, increased risk of heart attack and strokes. Even if we try the ostrich approach towards this problem basically the ignorance but the result could be lethal. Diabetes if left untreated can result in damaging nerves, kidneys, eyes and other organs. There are different types of insulin mainly Type 1 and Type 2.

Although diabetes mellitus is a fatal disease if not cured in time, early diagnosis can help in reducing the risk. Various medical diagnosis technique is already deployed for early diagnosis. The early risk prediction can be achieved using machine learning techniques. Recent researches have shown promising results in the risk prediction of diabetes mellitus. A range of machine learning techniques and classification techniques such as decision tree, random forests support vector machines, naïve Bayes and artificial neural network works better in the risk prediction. This is due to the computation capability and the ability to manage the data of these algorithms.

Classification accuracy evaluation metrics can be used to find the optimal accuracy of the classification and select the algorithm that performs best. Although only this metric is not enough to fully and properly select the best technique. Other metrics such as ROC value, F-score and computation time should also be taken into account to find the optimal result.

Therefore, the objective of this research is to do a comparative study of few of the classification algorithms namely k-nearest neighbours, Logistic Regression, decision tree, random forest, SVM, naïve Bayes and artificial neural network. These classification algorithms have been implemented on the early-stage diabetes risk prediction dataset[3] to classify the instances into positive and negative classes. Their performances are going to be assessed using various evaluation

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

metrics namely classification accuracy, F-score, ROC value and computation time. The findings of this paper will help future researchers to refer to generate a baseline algorithm for optimal classification of diabetes mellitus.

A. Related Works

Several types of research have been conducted on the application of ML techniques in early risk prediction of Diabetes to increase accuracy.

Pradeep et al.[4] discussed the prediction accuracy of J48, KNN, Random Forest and SVM and compared them on the diabetes dataset. The author concluded that the J48 algorithm provides an accuracy of 73.82% which is better than others before pre-processing the data. After pre-processing, KNN and RF provided better accuracy.

Xue-Hui Men et al.[5] compared J48, Logistic Regression and KNN algorithms on the diabetes dataset. The classification accuracy of J48 came out to be the highest with 78.27% accuracy.

Nongyao and Rungruttikarn[6] created a web application based on the prediction accuracy for diabetes prediction. Before that, they compared decision tree, neural network, logistic regression, naïve Bayes and random forest algorithms along with bagging and boosting for predication. They concluded that random forest performed best based both on accuracy and ROC score with an accuracy of 85.558% and ROC value of 0.912.

Saravananatha n and velmurugan[7] in their research analysed J48, CART, SVM and KNN on the medical dataset. They compared them based on accuracy, specificity, sensitivity, precision and error rate. They concluded that J48 algorithms performed best with an accuracy of 67.15% followed by SVM(65.04%), CART(62.28%) and KNN(53.39%).

Thirumal et al.[8] have analysed naïve Bayes, SVM KNN and C4.5 algorithms for diabetes prediction. They have concluded thatC4.5 have shown better accuracy that others with an accuracy of 78.2552%.
Anuja and Chitra[9] in their research have used five algorithms, namely SVM, Random forest, decision tree, MLP and logistic Regression along with four k-fold cross-validations (k=2,4,5,10). They have concluded that the highest accuracy of 78.7% is achieved by MLP with 4-fold cross-validation. They have shown that MLP performed the best among other algorithms.

Pradeep & Dr Naveen [10] used the J48 decision tree to classify the diabetes dataset. They have mentioned that the J48 algorithm is noted for its accuracy after proper feature selection.

Krati et al.[11] have implemented the KNN algorithm in their study on two datasets. They have obtained an accuracy of 70% on the data test1 and 57% accuracy on data test2.

Prajwala[12] in their research have R language to implement random forest(RF) and decision tree(DT) on the diabetes dataset and have concluded that RF provided better accuracy than DT but along with that execution time of RF is more than DT.

Lin[13] have analysed SVM, ANN and naïve Bayes classifiers in their research for diabetes prediction. They have conducted a weight-adjusted based study. The majority of voting was applied in this study. Their work concludes that a combination of the classifiers provided better accuracy than any single one.

Amit and Pragati [14] have conducted the study of C4.5, RF, MLP and Bayes net on Pima Indian Dataset. They have carried out feature selection on the dataset. Based on their study, they have concluded that feature selection has improved the performance of diabetes mellitus prediction.

Sajida et al.[15] discusses the role of Adaboost and Bagging ensemble machine learning methods [18] usingJ48 decision tree because of the basis for classifying the DM and patients as diabetic or non-diabetic, based on diabetes risk factors. Results achieved after the experiment proves that Adaboost machine learning ensemble technique outperforms well comparatively bagging also as a J48 decision tree.

Munaza Ramzan[16] carried out a study to predict diabetes using naïve Bayes, random forest and J48 techniques. The algorithms were applied using a 10 fold cross-validation method. They concluded that random forest provided the best accuracy than naïve Bayes and J48 techniques.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

Tao et al.[17] have applied naïve Bayes, RF, KNN, SVM, J48 and logistic regression for diabetes prediction and calculated the accuracy, sensitivity, specificity precision and AUC for the mentioned algorithms. They have used a 4 fold cross-validation method and concluded that Logistic regression is the best in terms of accuracy(99%).
Loannis et al.[18] carried out a study to predict diabetes using naïve Bayes, random forest KNN, SVM, decision tree and logistic regression techniques. The algorithms were applied using a 10 fold cross-validation method. In the study, they concluded that SVM provided the best accuracy than the rest with an accuracy of 84%.

Messan et al.[19] carried out a study to predict diabetes using ANN, SVM, GMM, ELM and logistic regression techniques. In the study, they concluded that ANN provided the best accuracy among the rest.

The above-related works have used various classification algorithms for the prediction of diabetes mellitus and have been improved for better performance. Along with classification accuracy, precision, ROC value, specificity and sensitivity have also been calculated by some of the researchers. A comparative study of all the algorithm and along with that the comparison based on accuracy, F-score, ROC and execution time should help in selecting the optimal algorithm for a better prediction of diabetes mellitus. This paper thus presented such a comparative study of various supervised learning algorithms(KNN, SVM, NB, RF, DT, MLP, LR) and the evaluation metrics mentioned above.

II. METHODOLOGY

A. The Dataset

The dataset that we have used is taken from the UCI repository[3].It contains 520 instances and 16 attributes with a few missing values which have been pre-processed by ignoring the tuples with incomplete values. The dataset is summarised in Table 1.

TABLE I. ATTRIBUTE DESCRIPTION

Attributes	**Description**
Age	20 years - 65 years
Sex	1.Male, 2.Female
Polyuria	1.Male, 2.Female
Polydipsia	1.Yes, 2.No.
Sudden Weight loss	1.Yes, 2.No.
Weakness	1.Yes, 2.No.
Polyphagia	1.Yes, 2.No.
Genital thrush	1.Yes, 2.No.
Visual blurring	1.Yes, 2.No.
Itching	1.Yes, 2.No.
Irritability	1.Yes, 2.No.
Delayed Healing	1.Yes, 2.No.
Partial Paresis	1.Yes, 2.No.
Muscle Stiffness	1.Yes, 2.No.
Alopecia	1.Yes, 2.No.
Obesity	1.Yes, 2.No.
Class	1.Positive, 2.Negative

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

After pre-processing the dataset a total of 520 instances remained. Out of these 520 instances, 320 are positive and 200 are negative values The two class variables(positive or negative) are used to find whether a patient has a risk of diabetes or not.

B. Experimental Procedure

The experimental procedure is carried out in the steps described below:

The dataset is partitioned into training and test set in a ratio of 80 : 20 respectively using 10-fold cross-validation set and the seven classification algorithms mentioned earlier is applied to classify the dataset into positive and negative classes. Four evaluation metrics: classification accuracy, F-score, ROC value and computation time is calculated to compare the performance of the specified algorithms. This is done to find the best classification algorithm.

C. Algorithms

1) Support Vector Machines :

Support Vector Machines(SVM) was first introduced by Vapnik[20]. SVM works by selecting critical samples from all classes known as support vectors and separating the classes by generating a function that divides them as broadly as possible using these support vectors. Therefore, it can be said that a mapping between an input vector to a high dimensionality space is made using SVM that aims to find the most suitable hyperplane that divides the data set into classes [21]. This linear classifier aims to maximize the distance between the decision hyperplane and the nearest data point, which is called the marginal distance, by finding the best-suited hyperplane [22].

This paper upon comparing various kernel, used Radial Basis function(RBF) kernel to classify the data, also known as the Gaussian kernel. When training an SVM with the Radial Basis Function (RBF) kernel, two hyperparameters must be considered: C and gamma[23]. The hyperparameter C, common to all SVM kernels, trades off misclassification of training examples against the simplicity of the decision surface. A low value of C smoothens the decision surface, while a high value of C aims at classifying all training examples correctly. The extent of influence of a single training example is defined by gamma. The larger the value of gamma is, the closer other examples must be to be affected.

The distance between data points is measured by the Gaussian kernel:

$$K_{rbf}(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2) \quad (1)$$

Here, x_i and x_j are data points, $\| xi - xj \|$ denotes Euclidean distance.

The choice of kernel functions is dependent on the respective data and specific domain problem. The various values of C and gamma is checked over the validation set. The appropriate value of C for best accuracy on the validation set is selected based on its accuracy. The best validation score is obtained for gamma=0.1 and C=1.7 .

2) Multi-layer perceptron :

Multi-layer perceptron is often referred to as Artificial Neural Networks. The entire model goes through optimization until the best yield is obtained. The MLP works on the principle of activation function and optimizing the weights.

$$f(x)= \sigma(b + W^TX) \quad (2)$$

The optimization takes place using various techniques like Gradient Descent, Newton Method, Quasi-Newton method, and many others. The purpose of the MLP gets decided by the use of the activation function. The sigmoid function also known as logistic is mostly used for binary classification. We have employed this method for classification as it compares probabilistic scores and a threshold mostly 0.5 differentiate between the concerned classes.

$$f(x)= \frac{1}{1+e^{-x}} \quad (3)$$

The dataset had to go around 200 iterations to get trained with an adaptive learning rate, initialized at 0.01.There were several experiments carried out for setting the number of hidden layers and it worked best when set to 100.

3) Decision Tree :

Decision Tree classifier is a supervised and very powerful machine learning algorithm for classification. It involves taking decisions based on prior data. In Decision Tree classifier, we have certain attributes that form various nodes of the tree. The algorithm, in every stage, chooses a node by evaluating the highest information gain among all the attributes[24].

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

The priority in nodes in the decision tree is set using Gini or Entropy, it is a score given to the best classifier among the set of attributes. The decision tree is ultimately a set of series of questionnaires which helps you classify. The efficiency of the algorithm increases on tuning hyperparameters like max depth, the depth of a tree, a criterion which is to be set either as 'Gini' or 'Entropy'. For the given dataset, the best accuracy is obtained for a max-depth of 7 and 'Gini' criterion.

4) Random Forests:

A random forest is essentially an ensemble of a number of decision trees. The logic that sticks with random forest is to combine different sets of values from training sets to form decision trees thus reducing the chances of overfitting and misclassification by averaging the results of various decision trees[25].

In the model described in the paper, we have used max-depth of 13 and 100 estimators to classify the data points. On increasing the max-depth, we experienced a decrease in the classification accuracy.

5) Naive Bayes:

The naive Bayes classifier is one of the most popularly used probabilistic classifiers. It implements Bayes Theorem[26] and discards the order and rules making the independent assumptions among the features. Hence deriving its naive nature. There are various types of Naive Bayes Algorithms, multinomial Naive Bayes, Gaussian Naive Bayes, Bernoulli Naive Bayes, and more.

We have used a Gaussian Naive Bayes classifier for our model. Gaussian Naive Bayes is used for high-dimensional data. Naive Bayes algorithms are fast to train and predict data points.

6) KNN classifier :

K-nearest neighbour classifier is one of the most simple and non-probabilistic machine learning algorithms. The training dataset is stored and the prediction involves looking for the closest data point from the training set.

$$\text{Euclidean} = \left(\sum_{i=1}^{k} |x_i - y_i|^p\right)^{\frac{1}{p}} \quad (4)$$

The primary use of the equation stated above is to calculate the distance between two data points x_i and y_i where k is the number of dimensions which is determined based on the dataset. We set p=2, Euclidean Distance formula. In the model mentioned in the paper, we have set 3 neighbours at which a good accuracy is achieved and increasing the neighbours although indicates better classification but a chance of misclassification significantly rises thus the model faces overfitting.

7) Logistic Regression:

Logistic Regression is a binary classification algorithm that follows the equation :

$$f(x) = \frac{1}{1+e^{-x}} \quad (5)$$

The range of the sigmoid function lies between (0,1). Logistic regression applies an L2 regularisation by default. The value of C plays an important role to train the model in order to minimize the misclassification and enhance the accuracy of the model. A low value of C smoothens the decision surface, while a high value of C aims at classifying all training examples correctly. The best value of the accuracy is achieved at C=1.7. There are not many hyperparameters to fine-tune the algorithm. The results can be favourable on large datasets. The increase of iterations didn't show any impact on the accuracy of the model.

III.RESULTS AND OBSERVATIONS

In this section, the classification accuracy, F-score and ROC value of the seven algorithms used for the diabetes dataset[3] is presented. All the evaluation metrics are summarised in table II.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

TABLE II. COMPARISON OF ACCURACY, F-SCORE, ROC VALUES OF 7 MODELS.

Model	Accuracy	F-Score	ROC
K-Nearest Neighbors	92.5436%	0.9337	0.9597
Logistic Regression	92.5436%	0.9359	0.9779
Support Vector Machines	94.4715%	0.9529	0.9866
Naive-Bayes	90.6388%	0.9206	0.9537
Decision Tree	94.2276%	0.9503	0.9445
Random Forests	98.0778%	0.9790	0.9979
Multi-Layer Perceptron	95.4413%	0.9507	0.9911

Table II displays the performance comparison of seven machine learning algorithms. The best classification accuracy is achieved by random forests with an accuracy of 98.0778% on the validation set. The next best results are shown by MLP followed by SVM(RBF kernel) with an accuracy of 95.44% and 94.47% respectively. The best performance in terms of F-score is achieved by random forests with a score of 0.979 followed by SVM, MLP and decision tree with scores of 0.9529,0.9507 and 0.9503 respectively. According to figure 1, Random forest shows the best performance in terms of classification accuracy, F-score and ROCS values among the seven algorithms presented in the paper.

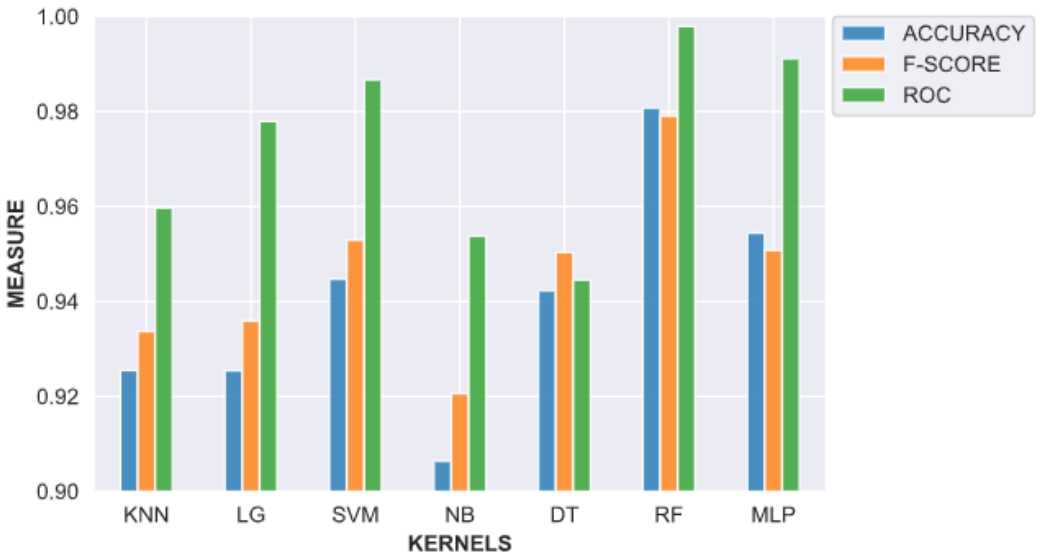

Fig 1. Performance Comparison of SEVEN CLASSIFICATION algorithms.

The comparison of ROC values of the seven algorithms presented in the paper shows that the Random Forests classifier performs best with a ROC value of 0.9979. The ROC value of Multilayer Perceptron(MLP) is not far behind that of random forests' with a value of 0.9911. The ROC value of support Vector machines(SVM) classifier is 0.9866. The ROC curve is shown in Fig 2.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

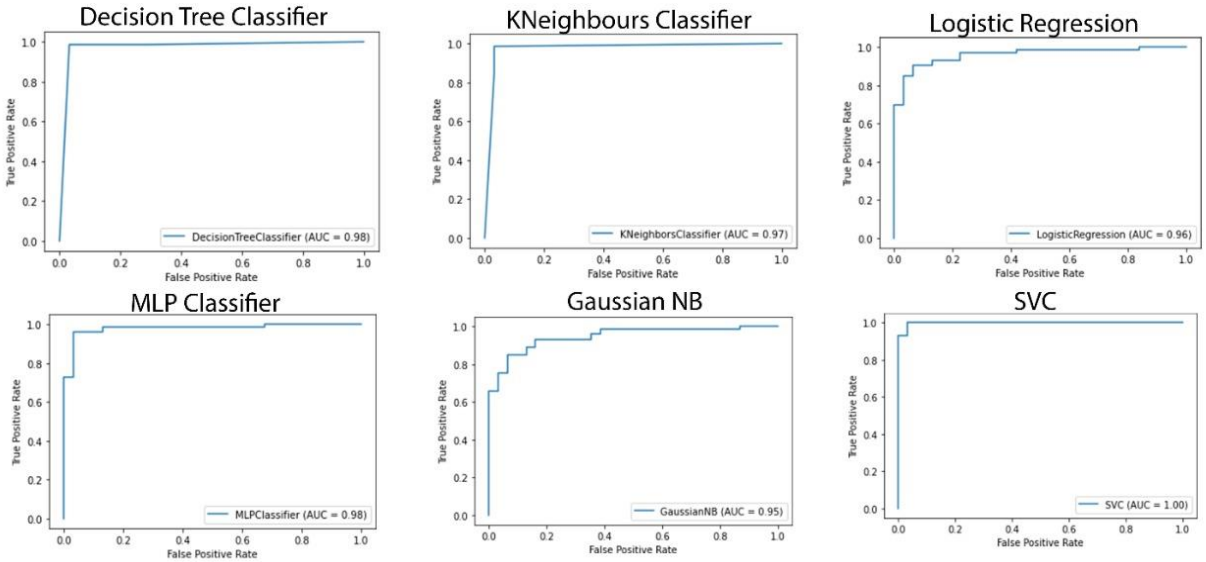

Fig. 2. ROC Curves of six classifiers

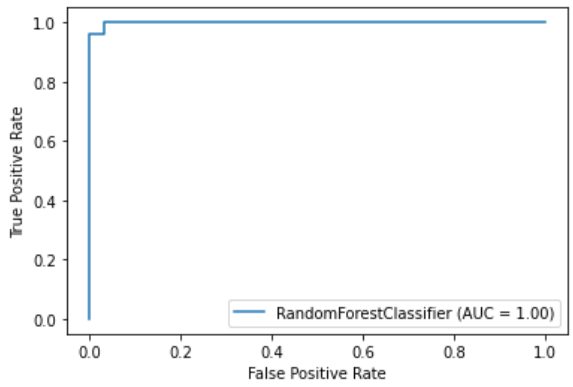

Fig. 3. ROC curve of Random Forests Classifier

From the performance comparison presented, it is clear that the top three classifiers are Random Forests, MLP and SVM with Random Forest being the best classifier for the given diabetes dataset with the best classification accuracy, F-score and ROC value on the validation set. Moreover, the classifiers performed well without overfitting in the test and train set. The rest five classifiers have also performed well-having accuracy more than 90% and F-score and Roc value of more than 0.90.

IV. CONCLUSION

The involvement of technology in the medical sector was one of the most significant milestones which we have yet achieved. Machine learning models can be used to predict various serious diseases like diabetes in humans at an early and curable stage.

In this paper we experimented with a diabetes dataset with different classification algorithms. Seven classification algorithms have been implemented on the validation sat of the used dataset. The results drawn from training several machine learning models clearly indicate that Random Forest Classifier proved to be the best model among the models used in the paper for the concerned dataset with an accuracy score of 98.0778%, ROC score of 0.9979 and F-score of 0.9790. Top three classifiers for the dataset are Random Forests classifier, Multi-layer perceptron and Support Vector Machine. Although, rest of the algorithm showed an accuracy of more than 90% and a F-score and ROC value of more than 0.9, the random forest classifier stands out with the maximum score in all the three evaluation metrices. Hence, as per the results obtained we can firmly believe that Random Forest Classifier is one of the most effective algorithms

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

against binary-based classification datasets. For Multi-layer Perceptron to work with highest accuracy it needs to be fed more training datapoints. This is one of the most valid reason for its underperforming in the concerned dataset.

In future more data must be collected from across the world for a more precise and accurate classification of the disease. Future study will concentrate on finding more factors that have the potential to cause diabetes and to include those potential factors in the dataset for a better classification. This can help in the enhancement and automation of diagnosis of the disease. Future studies on the disease and application of various data mining and ML algorithm can help in better early prediction of diabetes.

REFERENCES

[1] https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
[3] Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at an early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.
[4] Kandhasamy, J pradeep & Balamurali, Saminathan. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Computer Science. 47. 45-51. 10.1016/j.procs.2015.03.182.
[5] Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013 Feb;29(2):93-9. doi: 10.1016/j.kjms.2012.08.016. Epub 2012 Oct 16. PMID: 23347811.
[6] Nongyao Nai-arun, Rungruttikarn Moungmai, Comparison of Classifiers for the Risk of Diabetes Prediction, Procedia Computer Science, Volume 69, 2015, Pages 132-142, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.10.014. (https://www.sciencedirect.com/science/article/pii/S1877050915031786)
[7] Kavakiotis, Ioannis & Tsave, Olga & Salifoglou, Athanasios & Maglaveras, N. & Vlahavas, I. & Chouvarda, Ioanna. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal. 15. 10.1016/j.csbj.2016.12.005.
[8] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017). Application of data mining methods in diabetes prediction. *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, 1006-1010.
[9] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.
[10]Pradeep, K.R., & Naveen, N. (2016). Predictive analysis of diabetes using J48 algorithm of classification techniques. *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 347-352.
[11]Saxena, K., Khan, Z., & Singh, S. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm.
[12]T R, Prajwala. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. IJARCCE. 196-199. 10.17148/IJARCCE.2015.4142.
[13]Li, L. (2014, November). Diagnosis of diabetes using a weight-adjusted voting approach. In *2014 IEEE International Conference on Bioinformatics and Bioengineering* (pp. 320-324). IEEE.
[14]Kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *International Journal of Engineering and Applied Sciences*, *2*(5), 257905.
[15]Perveen, Sajida & Shahbaz, Muhammad & Guergachi, Aziz & Keshavjee, Karim. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science. 82. 115-121. 10.1016/j.procs.2016.04.016.
[16]Ramzan, M. (2016). Comparing and evaluating the performance of WEKA classifiers on critical diseases. *2016 1st India International Conference on Information Processing (IICIP)*, 1-4.
[17]Saravananathan, K & T, Velmurugan. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i43/93874.
[18]Zheng, Tao & Xie, Wei & Xu, Liling & He, Xiaoying & Zhang, Ya & You, Mingrong & Yang, Guixin & Chen, You. (2016). A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records. International Journal of Medical Informatics. 97. 10.1016/j.ijmedinf.2016.09.014.
[19]Thirumal, P. C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. *ARPN Journal of Engineering and Applied Science*, *10*(1), 8-13.
[20]Vapnik V. Statistical Learning Theory. John Wiley; 1998.
[21]G. Williams, “Descriptive and Predictive Analytics", Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R, pp. 193- 203, 2011
[22]Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, Volume 13, 2015, Pages 8-17, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2014.11.005. (https://www.sciencedirect.com/science/article/pii/S2001037014000464)
[23]Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
[24]Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774
[25]Mller, Andreas C., and Sarah Guido. *Introduction to Machine Learning with Python: a Guide for Data Scientists*. O’Reilly, 2017.
[26] Ray, S., 2017. 6 Easy Steps to Learn Naive Bayes Algorithms (with code in Python).

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8228

BIOGRAPHY

Abhimanyu Jadli is currently pursuing Bachelors In Engineering in Computer Science from Chandigarh University, Mohali, India. His area of specialization is the undergraduate degree in Artificial Intelligence and Machine Learning. He is the author of No Directions, a published science fiction novel.

Apratim Sadhu is currently pursuing Bachelors in Engineering in Computer Science from Chandigarh University, Mohali, India. His area of specialization in the under-graduate degree is Artificial Intelligence and Machine Learning. He is a rank holder in 19th National Children Science Congress. He has written 1 research article on Machine Learning.