Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications

Dr. Trilok Nath Pandey
Computer Science and Engineering (of Affiliation)
Vellore Institute of Technology *(of Affiliation)*
Chennai, India
triloknath.pandey@vit.in

Abhishek N.N.
Computer Science and Engineering (of Affiliation)
Vellore Institute of Technology *(of Affiliation)*
Chennai, India
abhishek.nn@vitstudent.ac.in

Mayank Gupta
Computer Science and Engineering (of Affiliation)
Vellore Institute of Technology *(of Affiliation)*
Chennai, India
mayank.gupta2020@vitstudent.ac.in

Fidal Mathew
Computer Science and Engineering (of Affiliation)
Vellore Institute of Technology *(of Affiliation)*
Chennai, India
fidal.mathew2020@vitstudent.ac.in

Ayush Kapri
Computer Science and Engineering (of Affiliation)
Vellore Institute of Technology *(of Affiliation)*
Chennai, India
ayush.kapri2020@vitstudent.ac.in

***Abstract*— One of the most important health issues in both industrialized and developing nations is diabetes mellitus. Consequently, the According to the International Diabetes Federation, 425 million people worldwide have diabetes. Within 20 years, this number is projected to increase to 380 million. Due to its significance, a classifier design for the early diagnosis of diabetes that is both affordable and effective is now necessary. Testing data mining methods to determine their predictive accuracy in the classification of diabetes data has become a standard at the UCI machine learning lab using the Pima Indian diabetic database. The machine learning technique is focused on categorizing the diabetic illness from a large medical dataset into type 1 and type 2. The goal of this project is to create a model that can predict a patient's chance of developing diabetes with the highest degree of accuracy. As a result, this experiment uses Decision Tree, Naive Bayes, Random Forest, Kernel SVM and Logistic Regression five machine learning classification methods, to identify diabetes at an early stage. Confusion Matrix, Precision, Accuracy, F-Measure, and Recall are just a few of the metrics used to assess how well the three algorithms perform. Correctly and wrongly labelled examples are used to gauge accuracy. According to the results, Logistic Regression surpasses other algorithms with a highest accuracy of 79.6%. These findings are properly and methodically validated using Receiver Operating Characteristic (ROC) curves.**

Keywords*— *Diabetes, Logistic Regression, Kernel SVM, Naïve Bayes, Decision Tree, Random Forest Accuracy Machine Learning

I. INTRODUCTION

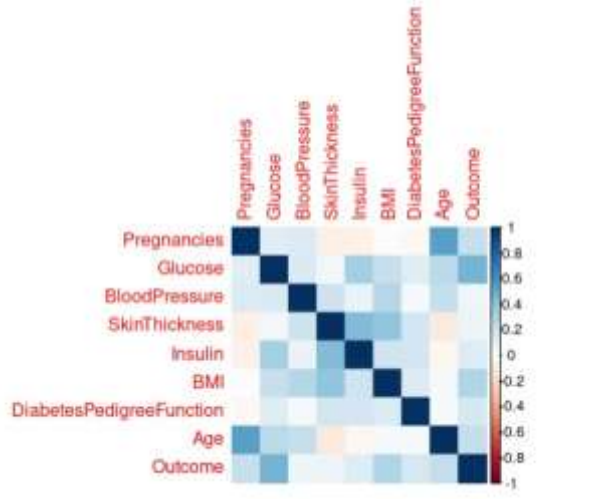
In the medical industry, classification algorithms are frequently used to categorise data into different groups in accordance with specified constraints as opposed to using a single classifier. Diabetes is a condition that impairs the body's ability to produce the hormone insulin, which causes improper carbohydrate metabolism and raises blood glucose levels. A person with diabetes typically experiences elevated blood sugar. Increased hunger, increased thirst, and frequent urination are a few signs and symptoms of high blood sugar. Diabetes has a lot of side effects if it is not addressed. Diabetes-related ketoacidosis and nonketotic hyperosmolar be managed are two serious consequences. Diabetes is influenced by a number of variables, including height, weight, hereditary factors, and insulin, but the main component that is taken into considerationis sugar concerntration among all factors. The early identification is the only remedy to stay away from the complications. Diabetes is examined as a vital serious health matter during which the measure of sugar substance cannot parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. DATASET USED

A. *PIDD-Pima Indians Diabetes Dataset*

The proposed methodology is evaluated on Diabetes Dataset namely (PIDD), which is taken from UCI Repository. This dataset comprises of medical detail of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class ’0’ treated as tested negative for diabetes and value of another class ’1’ is treated as tested positive for diabetes. Dataset description is defined by Table-4 and the Table-5 represents Attributes descriptions.

B. *Correlation Matrix*

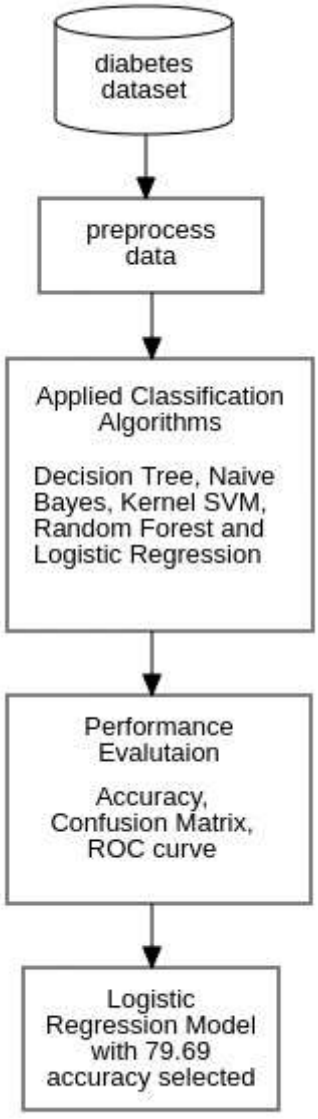


Data cleaning is **the process of correcting or eliminating data that is erroneous, corrupted, poorly formatted, duplicate, or incomplete within a dataset**. We have

removed some unecessary features of our data by careful observations and experimentation. We implemented correlation heatmaps. They help us to understand which variables are related to each other and the strength of this relationship. By observation, we notice **Age** and **Pregnancies** features are very much correlated. To achieve a good machine learning model, we removed **Pregnancies** feature as **Age was more correlated to Outcome**. We also removed **Skin Thickness** from all the models as it has very slight correlation with the outcome. Having features which do not contribute to the final result makes the model more complex and inefficient.

III. METHODOLOGY USED

Learning (determining) good values for all of the weights and the bias from labelled examples is all that is required to train a model. In this experiment, we will using different machine learning algorithms to predict whether the person is subjected to type-1 or type-2 diabetes.

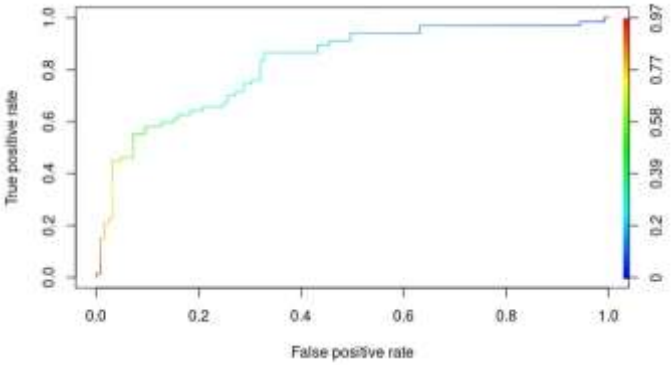


A. Logistic Regression

The Logistic regression helps to classify the concern person will get diabetes or not. Since we are using the logistic regression we have to mention that, family binomial. We are using all the attributes we have in the dataset. The Confusion Matrix of Logistic Regression is as follows:

Confusion Matrix for Logistic regression

	A	***B***
Test Positive	116	9
Test Negative	30	37



B. Naive Bayes Classifier

Naive Bayes is a classification technique based on the idea that all features are independent and unconnected to one another. It specifies that the status of one feature in a class does not impact the status of another. Because it is based on conditional probability, it is regarded as a powerful method used for classification. It works well with data that has imbalancing issues and missing values. The Bayes Theorem is used by Naive Bayes, a machine learning classifier. The posterior probability P(C|X) can be determined using the Bayes theorem from P(C), P(X), and P(X|C). Therefore,

$$P(X|C) = P(X|C)\, P(C)/P\,(X).$$

Where,

P(C|X) is the posterior probability of the target class.

P(X|C) denotes the probability of the predictor class.

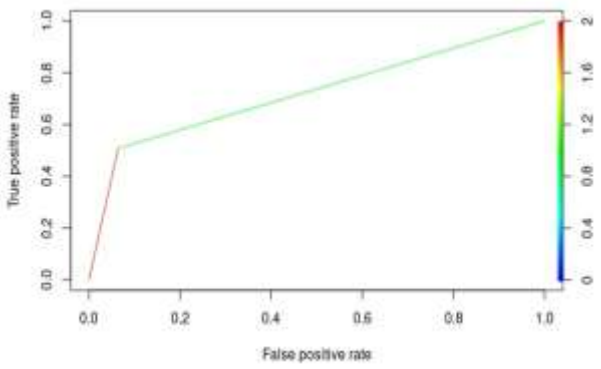
P(C) is the probability that class C is true.

P(X) denotes the predictor's prior probability.

The Naive Bayes algorithm's performance was tested using confusion matrix:

Confusion Matrix for Naïve bayes

	A	***B***
Test Positive	117	8
Test Negative	33	34

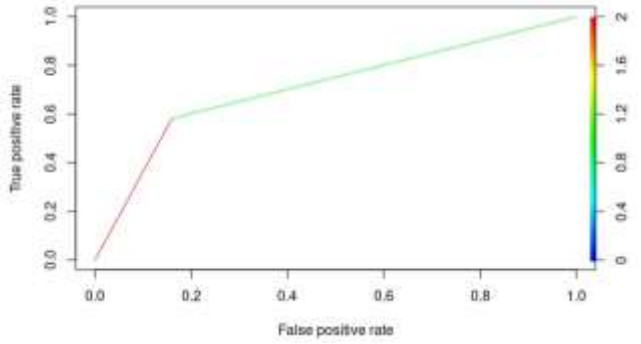


C. *Decision Tree Classifier*

A supervised machine learning algorithm used to tackle categorization problems is Decision Tree. The prediction of target class using decision rule drawn from prior data is one of the benefits of employing Decision Tree in this research endeavour. It predicts and classifies using nodes and internodes. Root nodes categorise instances based on various characteristics. The root node may have two or more branches, whereas the leaf node represents classification. At each level, the Decision Tree selects a node based on the maximum information gain among all qualities. The results of his evaluation of the Decision Tree technique using the Confusion Matrix are as follows:

Confusion Matrix for Decision tree

	A	***B***
Test Positive	105	20
Test Negative	28	39

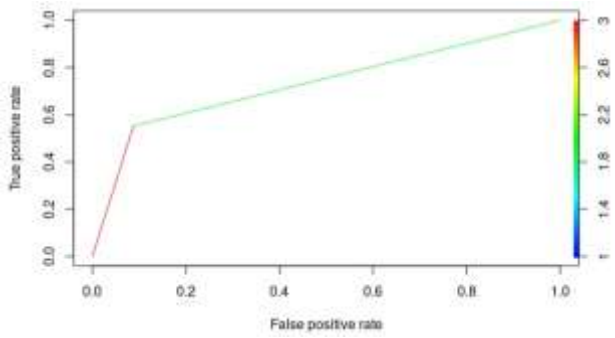


D. *Kernel SVM*

SVM is a supervised machine learning model that is commonly used in classification. A support vector machine's goal, given a two-class training sample, is to determine the optimal highest-margin separation hyperplane between the two classes. For greater generalisation, the hyperplane should not be located closer to data points from the other class. A hyperplane that is far from the data points in each category should be chosen. The support vectors are the spots closest to the classifier's margin. The estimated performance of the SVM algorithm for diabetes prediction using the Confusion Matrix is as follows:

Confusion Matrix for Kernel svm

	A	***B***
Test Positive	114	11
Test Negative	30	37

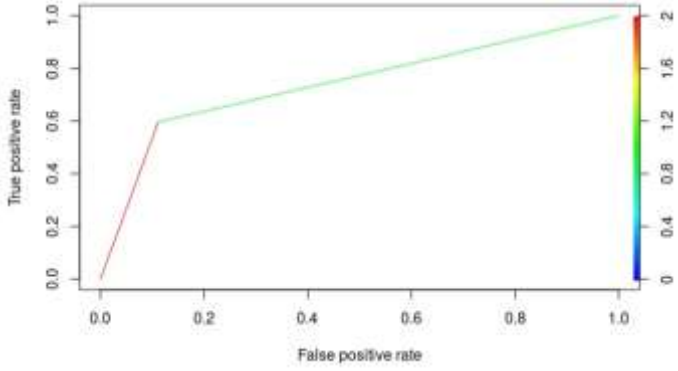


E. *Random Forest*

The method of supervised learning includes Random Forest. It can be used to solve ML problems involving both classification and regression. It is based on the concept of ensemble learning, which is a method of combining several classifiers to address tough difficulties and improve model performance. Following is an evaluation of the SVM algorithm's performance in predicting diabetes using the Confusion Matrix:

Confusion Matrix for Random forest

	A	***B***
Test Positive	111	14
Test Negative	27	40



IV. ACCURACY MEASURES

In this study, the algorithms Naive Bayes, SVM, and Decision Tree are employed. Tenfold internal cross-validation is used throughout experiments. For the classification of this study, the metrics Accuracy, F-Measure, Recall, Precision, and ROC (Receiver Operating Curve) are employed. Below are the accuracy measures defined in Table-6:

A. *Accuracy*

The accuracy of a machine learning classification method can be used to determine how frequently it properly identifies a data point. Accuracy is defined as the fraction of correctly predicted data points among all data points.

B. Precision

Precision is defined as the proportion of correctly categorised positive samples (True Positive) to the total number of positively classified samples (either correctly or incorrectly).

C. F-Measure

The F-measure is calculated using the harmonic mean of precision and recall, with each given the same weight. It enables the evaluation of a model's precision and recall with a single score, which is valuable for discussing model performance and comparing models.

D. Recall

The recall is calculated by dividing the proportion of Positive samples that were correctly recognised as Positive by the total number of Positive samples. The recall metric measures how successfully the model can detect positive samples. The greater the number of positive samples detected, the greater the recall.

E. ROC

A receiver operating characteristic curve, also known as a ROC curve, is a graphical representation of how the diagnostic capacity of a binary classifier system changes as the discrimination threshold is changed.

TABLE I. ACCURACY MEASURES

SN	***Measures***	***Formula***
1	Accuracy	A=(TP+TN) / (Total no of samples)
2	Precision	P = TP / (TP+ FP)
3	Recall	R =TP / (TP+FN)
4	FMeasure	F=2*(P*R) / (P+R)
5	ROC	Used to compare the usefulness of tests.

TABLE II. COMPARATIVE ACCURACY GAIN USING FEATURE SELECTION

SN	***Measures***	***OLD Accuracy***	***Accuracy with Feature Selection***
1	Logistic Regression	78.65	79.69
2	Naïve Bayes	77.08	78.65
3	Decision Tree	74.48	75.00
4	Kernel SVM	77.60	78.65
5	Random Forest	77.60	78.65

TABLE III. COMPARATIVE PERFORMANCE OF CLASSIFICATION ALGORITHMS ON VARIOUS MEASURES

SN	**Performance Evaluation**				
	Machine Learning Models	***Accuracy***	***F Measure***	***Recall***	***Precision***
1	Logistic Regression	79.69	85.60	79.45	92.80
2	Naïve Bayes	78.65	85.09	78.00	93.60
3	Decision Tree	75.00	81.39	78.94	84.00
4	Kernel SVM	78.65	84.75	78.72	90.98
5	Random Forest	78.65	83.65	80.43	88.80

ACKNOWLEDGMENT

This is an acknowledgement to VIT Chennai as a whole to motivate and encourage us to develop a curious mind that never stops learning. This research has only been possible with the inordinate and meticulous efforts VIT has put in while shaping our journey of becoming more diligent and capable than I already was.

I extend my gratitude to the professor Dr. Trilok Nath Pandey, Associate Professor B.Tech Computer Science and Engineering , SCOPE, VIT Chennai who took our CSE3505 course. I truly appreciate him and his time he spent helping us do this research, pushing us to be good researcher. I also extend my gratitude to my parents who constantly supported me throughout the journey.

REFERENCES

[1] 1. A. Mir and S. N. Dhage "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare" Proceedings - 2018 4th International Conference on Computing Communication Control and Automation ICCUBEA6 2018 Jul. 2018.

[2] 2. D. Dutta D. Paul and P. Ghosh "Analysing Feature Importances for Diabetes Prediction using Machine Learning" 2018 IEEE 9th Annual Information Technology Electronics and Mobile Communication Conference IEMCON 2018 pp. 924-928 Jan. 2019.

[3] 3. Deepti Sisodiaa and Dilip Singh Sisodiab Prediction of Diabetes using Classification Algorithms 2018.

[4] 4. M. A. Sarwar N. Kamal W. Hamid and M. A. Shah "Prediction of diabetes using machine learning algorithms in healthcare" ICAC 2018 – 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing Sep 2018.

[5] 5. Neha Prernatigga Shruti Garg et al. Prediction of Type 2 Diabetes using Machine earning Classification Methods 2019.

[6] 6. Priyanka Indoria Yogeshkumar Rathore et al. Detection and Prediction of Diabetes Using Machine Learning Techniques 2018.

[7] 7. M. Alehegn and R. Joshi "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach" Int. Res. J. Eng. Technol. Aug. 2017.

[8] 8. R. M. Khalil and A. Al-Jumaily "Machine learning based prediction of depression amongtype 2 diabetic patients" Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering ISKE 2017 vol. 2018-January pp. 1-5 Jul 2017.

[9] 9. Saiteja Prasad Chatrati Gahangir Hossain Ayush Goyal Anupama Bhan Sayantan Bhattacharya Devottam Gaurav et al. Journal of King Saud University - Computer and Information Science 2020.

[10] 10. S. P. Chatrati et al. "Smart home health monitoring system for predicting type 2 diabetes and hypertension" J. King Saud Univ. - Comput. Inf Sci. Jan. 2020.

[11] 11. S. F. Sayyad S. Bhingardive P. Ghatnekar K. More and P. Rajendran "Resume Extractor and Candidate Recruitment System using Online Test and SMTP" IJIRST-International.

[12] 12. J. Innov Res. Sci. Technol. vol. 5 2019.

[13] 13. S. K. Dey A. Hossain and M. M. 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 2018 Jan. 2019.

[14] 14. N Nai-Arun and R Moungmai "Comparison of Classifiers for the Risk of Diabetes Prediction" Procedia Computer Science 2015.

[15] 15. M Komi J Li Y Zhai and Z Xianguo "Application of data mining methods in diabetes prediction" 2nd International Conference on Image Vision and Computing ICIVC 2017.

[16] 16. N El_Jerjawi and S Abu-Naser "Diabetes Prediction Using Artificial Neural Network" International Journal of Advanced Science and Technology 2018.

[17] 17. D Sisodia and D Sisodia "Prediction of Diabetes using Classification Algorithms" Procedia Computer Science 2018.

[18] 18. M Sarwar N Kamal W Hamid and M Shah "Prediction of diabetes using machine learning algorithms in healthcare" ICAC 2018 – 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing 2018.

[19] 19. A Mujumdar and V Vaidehi "Diabetes Prediction using Machine Learning Algorithms" Procedia Computer Science 2019.

[20] 20. D Dutta D Paul and P Ghosh "Analysing Feature Importances for Diabetes Prediction using Machine Learning" 2018 IEEE 9th Annual Information Technology Electronics and Mobile Communication Conference 2019.

[21] 21. T Mahboob Alam M Iqbal Y Ali A Wahab S Ijaz T Imtiaz Baig et al. "A model for early prediction of diabetes" Informatics in Medicine Unlocked 2019.

[22] 22. P Sonar and Malini K Jaya "Diabetes prediction using different machine learning approaches" Proceedings of the 3rd International Conference on Computing Methodologies and Communication 2019.

.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.