DOI: http://dx.doi.org/10.26483/ijarcs.v11i5.6654

Volume 11, No. 5, September-October 2020

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

DIABETES PREDICTION AND VALIDATION MODEL USING ML CLASSIFICATION ALGORITHMS

Pradyut Nath
Dept. of Computer Science & Engineering
Meghnad Saha Institute of Technology
Kolkata, India

Sumagna Dey*
Dept. of Computer Science & Engineering
Meghnad Saha Institute of Technology
Kolkata, India

Indrajit Das
Dept. of Information Technology
Meghnad Saha Institute of Technology
Kolkata, India

Subhrapratim Nath
Dept. of Computer Science & Engineering
Meghnad Saha Institute of Technology
Kolkata, India

Dyuti Mohapatra
Dept. of Computer Science & Engineering
Meghnad Saha Institute of Technology
Kolkata, India

Abstract: Diabetes is now a global wide concern, which can critically impact and disrupt the normal lifestyle and the everyday activities of any individual. Due to the lack of insulin and high glucose content in the body, anyone can get diagnosed with diabetes. Apart from all the medical factors, there are few additional non-medical factors in an individual's daily life like hypertension, heredity, daily standard activity, smoking habits, body mass index etc. that might play a part in triggering diabetes. Several medical studies reveal that for women sometimes pregnancy frequencies or any kind of heart issues can also trigger diabetes. The paper aims to predict the most critical factor that contributes in triggering diabetes in any individual by using classification and predictive analysis algorithms. Five well known machine learning classification algorithms are used where a filtering scheme based on 75% threshold accuracy rate is employed followed by verification using AUROC metric aiming low error rate and high prediction accuracy. Additionally, the model used Ensemble learning to make predictions and validates the proposed scheme against PIMA Indian Diabetes dataset.

Keywords: logistic regression; random forest algorithm; support vector machine; naïve bayes; KNN, AUROC; ensemble learning

I. INTRODUCTION

The country's well being framework burns through billions of dollars attempting to treat, oversee and forestall a variety of avoidable conditions that alone keep on developing in predominance. Almost 66% of reported death every year is attributed to chronic condition. Generally, 86% of the $2.9 trillion spent on medicinal services and healthcare in 2013 was identified with chronic diseases [1]. India has records of having almost around 62 million individuals identified to have Diabetes consistently every year. Most of the contributions of works, conducted by researchers worldwide, is attributed in accurately predicting the most critical and important factors that contribute to the occurrence of diabetes in any individual.

Swain et al. in [2] proposed on the forecast and grouping of Diabetes Mellitus utilizing (ANN) and ANFIS. The model was trained with data that was collected from 100 individuals and results showed that ANFIS (90.32%) gave better accuracy with less error rate as compared to ANN (71%).

W. Xu et al. in [3] showed a type-2diabetes prediction model using Random Forest (RF) which analyzed contribution of common factors (like age, waist etc.) on occurrence of diabetes. The RF algorithm bagged the best accuracy (84.13%) over other Naive Bayes (80.50%) algorithm, ID3 algorithm (72.94%) and AdaBoost algorithm (81.63). To check the trustworthiness of the model, the k-fold cross approval was used where k value was 10. It was found that high sensitivity was achieved for all the employed algorithms. Sensitivity of 91.17%, 92.11%, 99.05%, and 100% respectively was obtained for the aforesaid algorithms; however the specificity value was pretty low for all the scenarios. Overall the RF model could effectively predict the risk of diabetes when sufficient data was present.

L. Griva et al. in [4] showed a deep study about the real prediction capabilities of two different models: Wiener type and Autoregressive Exogenous Model (ARX) in the field of Diabetes Mellitus Type-1. To compare these two models RMSE and the Clarke Grid Error Analysis (p-CGA) was used. It was seen that the 3 gatherings of information, for the various models, the data sources are picked are adequately eager to make reliable with the estimation of the parameters. It was seen from results that the ARX model can achieve very good results in case the patient presents his glucose within the safety range in most cases.

J. He. et al. in [5] gave a prediction on glucose concentration in blood, a primary factor leading to diabetes using Canonical correlation or Canonical variants analysis. The historical blood glucose data of patients and the future blood glucose data were modeled by canonical correlation analysis. Blood correlation equation was obtained and blood glucose prediction was conducted on that basis. The genuine estimation information of Type-1 diabetics was utilized to check the forecast impact. The contribution of the test chooses the historical information of 20 min for every patient and the anticipated horizon was 5, 10, 15, and 20 min separately. The normal estimations of the related root mean square errors were: 6.096, 12.022, 17.384, 21.713 mg/dl. The consequences of this examination were contrasted with past expectation techniques, thereby demonstrating that the standard connection investigation strategy has high potential in forecasting blood glucose content and accomplished high-accuracy expectation.

In this proposed work, five well-known ML Classification algorithms namely Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Naïve Bayes (NB) and K-Nearest Neighbor (KNN) algorithms are used for detection of diabetes and a comparative analysis is conducted based on their respective prediction accuracy. A filtering scheme is employed to filter out the algorithms below threshold accuracy rate of 75% .The obtained results are verified in a systematic manner by using favored classification metrics like AUC and ROC curves. Ensemble learning schemes have also been used for prediction purposes. In the paper, the accuracy of the proposed model in predicting diabetes is validated against PIMA Indian Diabetes.

The remaining paper is divided into as follows; Section 2 gives an overview of some classification ML algorithms used in our paper. Section 3 puts forward the proposed methodology that has been pursued in this work. Next part (Section 4) describes the experimental results. Finally rest concludes the paper.

II. CLASSIFICATION OF MACHINE LEARNING ALGORITHMS

There are several classification techniques to perform analysis on a dataset to improve results more accurately and precisely. The classification techniques used by this paper are outlined below briefly.

A. Logistic Regression

Logistic Regression (LR) [6] is applied when the reliant variable is in binary. Like any contemporary regression algorithm, it is also a predictive analysis. The data and one dependent variable's relationship with the list of independent variables is described in this regression. The dependent variable can only contain two values: TRUE/Success (diabetes tested positive) and False/Failure (diabetes tested negative).

LR produces the coefficients (and its standard errors and criticalness levels) of a function to define a log transformation of the probability of occurrence of the target variable.

$$\log(p) = b_0 + b_1 X_1 + b_2 X_2 + \cdots\cdots + b_k X_k + E_0 \quad (1)$$

Here p is the probability of essence of the characteristic target variable. The log transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probabilityofpresenceofcharacteristics}{probabilityofabsenceofcharacteristics} \quad (2)$$

And

$$\log(p) = \ln\left(\frac{p}{1-p}\right) \quad (3)$$

B. Support Vector Machine(SVM) [7]

It is an algorithm that classifies data points into categories. First training data is taken. Data as points are mapped in space, arranged into classes with the greatest margin possible. New testing data are then mapped into the space as points and depending on the side of the gap they get mapped, their categories are predicted. The main objective is to design a hyperplane in this N – dimensional space (with N features) that arranges the mapped points in classes/categories distinctly, such that the interval between the data points and hyperplane is maximized. The loss function used here is hinge loss which is used to increase the margin gap. The cost becomes 0 if actual and predicted values are of the same sign. If not, find the loss value, for which there is a regularization parameter. Regularization parameter objective is to balance loss value and margin maximization. After combining regularization parameter, the cost function is defined as follows:

$$c(x, y, f(x)) = \{ \begin{matrix} 0, if\ 1 \leq y * f(x) \\ 1 - y * f(x), else \end{matrix} \quad (4)$$

C. Naïve Bayes algorithm

It is based on Bayes' theorem assuming that the feature of every pair is independent. Naive Bayes classifier [8] has multiple practical applications such as grouping news and email spam detection. Here, a hypothesis's probability value is computed using Bayes Theorem given the prior knowledge.

Bayes Theorem defines that:

$$P(X \mid Y) = \frac{P(Y \mid X) * P(X)}{P(Y)} \quad (5)$$

In Gaussian Naive Bayes, each feature (assumed to be Gaussian distributed) associates a continuous value. It is also known as Normal distribution. It is a bell-shaped curve when plotted; the point of symmetry is the mean of the feature values. These features are assumed to be Gaussian distributed having probability distribution:

$$P(X_i \mid Y) = \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma^2 y}\right) \quad (6)$$

D. K- Nearest Neighbor algorithm (KNN)

This is a kind of lazy learning, as it doesn't endeavor to develop a general interior model, but stores a reference of the training data. The training data are vectors in multidimensional space, each of these vectors are labeled. The unlabeled vector is first plotted then depending on the larger part vote from k closest training samples, from target point, the class is determined. The classification is computed for each point. The assumption in KNN algorithm [10] is that things with similar properties exist very close to each other. In other words, things with similar properties are near to each other. This algorithm takes the idea of shortest distance between points using techniques like Sum of squared distance (SOS) etc.

III. PROPOSED METHODOLOGY

The proposed model aims to predict the most critical factor that contributes in triggering diabetes in any individual as shown in Fig.1. Data preprocessing is a vital stage where several attributes which trigger diabetes are gathered and processed. The processed data is classified by five classification algorithms and their individual ROC accuracy is computed. The filtering is done based on the threshold value of 75%. Finally prediction of diabetes is done by Ensemble Learning algorithm and validation is done with PIMA Dataset.

A. Data Processing

Data processing is the transformation of data into the desired structure. Most of the data handling is finished by utilizing machines automatically. The yield or "handled" information can be acquired in various structures like a picture, diagram, table, vector document, sound, graphs or some other wanted organization relying upon the product or strategy for information preparation utilized.

Data Collection.

It is the way toward social occasion and procedure of get-together and estimating data on focused variables in an established framework which at that point empowers one to respond to significant inquiries and assess results. In this paper, the data collection is done from various surveys, over different sources (i.e. internet, external source etc.). Based on various researches on the web and medical fields, many non-medical risk factors have been analyzed too, which can affect any individual to be diagnosed with diabetes. Though those datasets mostly contain medical-based factors (i.e. pregnancies, glucose, blood Pressure, skin thickness, BMI, age). Yet the main objective is to predict chances of diabetes evaluated from non-medical risk factors.

Data Preprocessing.

In any ML procedure, information preprocessing is that progression where the information gets changed or encoded, to carry it to such an express, that now the machine can easily parse it. Genuine information is frequently inadequate, conflicting, or potentially ailing in specific practices or slants and is probably going to contain numerous erroneous information. Data pre-processing is a demonstrated strategy for settling such issues. It is a data mining strategy that includes changing raw information into an understandable configuration.

B. Classification Algorithm

In this paper, five classification techniques (LR, SVM, RF, KNN and NB) are used to design prediction models for diabetes.

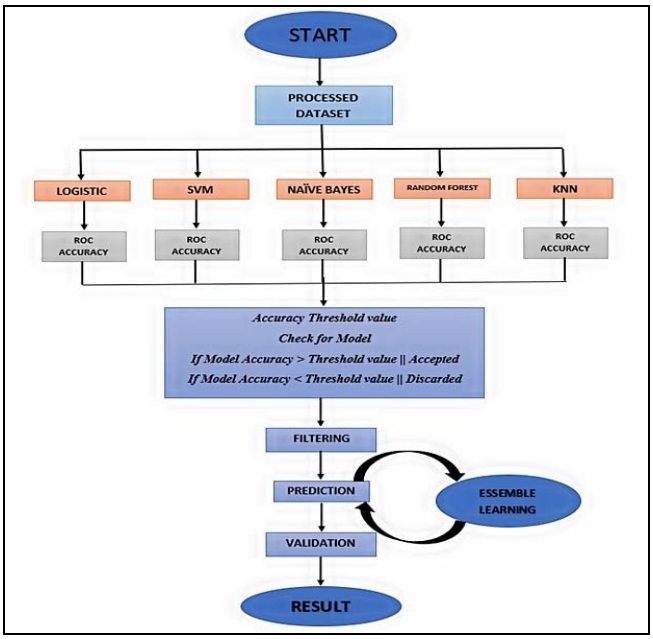

Figure 1. Proposed Methodology

Algorithm.

For evaluation of the classification accuracy of the proposed methodology, the used dataset is partitioned in the ratio 7.5:2.5 to form a Training Data set (on which machine will learn) and test Data Set (on which final classification will be executed).

Algorithm
Input: Dataset Table
Output: Confusion Matrix (Accuracy)
1: Begin
2: read(dataset)
3: split(data (7.5:2.5))
4: training(data)
5: testing(data)
6: model (objective function (4))
7: summary(model)
8: End

C. ROC and Confusion Matrix

The popular classification metrics like ROC Curve and Confusion Matrix are computed with parameters as follows:

- True positives (TP): In this case prediction is true (diabetes tested positive) and they indeed have it.
- True negatives (TN): Prediction is False, also the individual tested negative for diabetes.
- False positives (FP): In this case prediction is true and they indeed have not it.

- False negatives (FN): Prediction is False, but an individual tested positive for diabetes. (Also known as a "Type II error.")
-

D. *Filtering*

Based on threshold accuracy (75%) value of classification algorithm, below threshold algorithms are filtered out. In this paper, contrast with the other four KNN models gives less accurate results i.e. 66%. For that reason since the KNN model outcome is especially poor, filtering out KNN is done and the rest four algorithms are passed on to the prediction stage where Ensemble Learning is used.

E. *Ensemble Learning*

It is the process where multiple models are joined and applied together to solve a computational problem and are strategically generated. It is primarily directed on improving the (classifier, function approximation model, prediction model, etc.) the model's performance, or to minimize the likelihood of a poor unfortunate selection. Voting and averaging are two of the easiest ensemble methods. Voting is useful in classification and for regression average or mean method calculation is used. Each model makes a forecast for each test example and the majority vote is considered the final prediction. If no prediction crosses above half of the votes, it implies that the ensemble method cannot make a stable prediction for that instance.

Ensemble learning Algorithm

Input: Data weight $\{W_n\}$

Output: Final Prediction

1. Begin
2. Iterate for m = 1 to M then
3. fit classifier $y_m(x)$ by minimizing weight error function $J_m(x)$

$$J_m = \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x_n) \neq t_n]$$

4. Calculate

$$\in_m = \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x_n) \neq t_n] / \sum_{n=1}^{N} w_n^{(m)}$$

5. Calculate

$$\alpha_m = \log\log(\frac{1-\in_m}{\in_m})$$

6. Update data weights:

$$W_n^{(m+1)} = W_n^{(m)} \exp\{\alpha_m 1[y_m(x_n) \neq t_n]\}$$

7. End for
8. Find predictions :

$$Y_M(x) = sign(\sum_{m=1}^{M} \alpha_m y_m(x))$$

9. Exit

IV. EXPERIMENTAL RESULT

In this paper, for better and accurate prediction of results, a comparative study is conducted among all the five ML models to rule out the discrepancies. Fig.2 shows the comparative analysis based on standard ML classification metrics like sensitivity, accuracy and specificity computation of the five adopted ML algorithms.

Figure 2.

Figure 3. Accuracy= TPC+TNC / (TPC + FNC+ FPC + TNC) (7)

Figure 4. Sensitivity= true positive/true positive+false negative (8)

Figure 5. Specificity= true negative/false positive+true negative (9)

From Fig.2, it is pretty evident that RF with accuracy of 90.09% is best while KNN model gives least accuracy results (66%) which is lower than threshold accuracy value of 75% among all the other classification algorithms. So, KNN model results can be neglected over the other four model results.

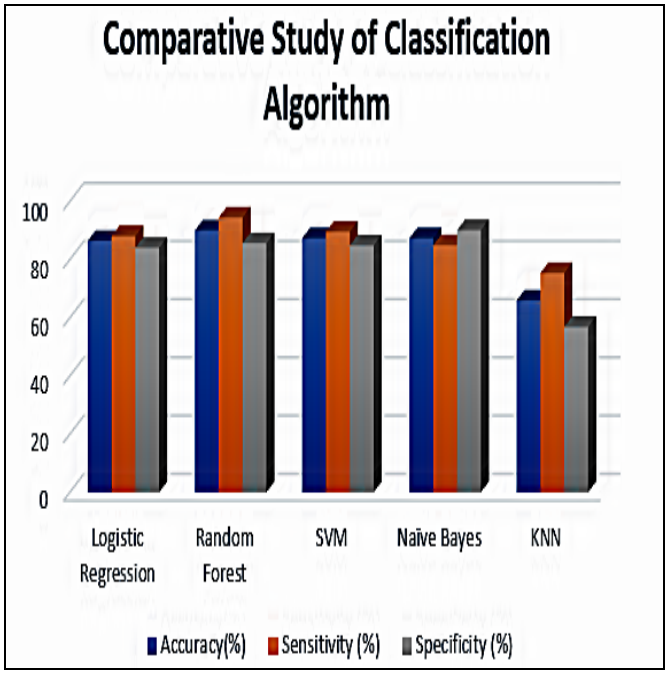

Figure 6. Comparative analysis of classification algorithm

Here, the AUC (Area under the Curve) and ROC (Receiver Operating Characteristics) curves are depicted in Fig.4. More the portion under the curve better is the model performance. AUROC is around the range of 80-90%, which means the model has less error rate and its accuracy of prediction is pretty high. In Fig.3, the ROC curve has been plotted based on predictions from the four models.

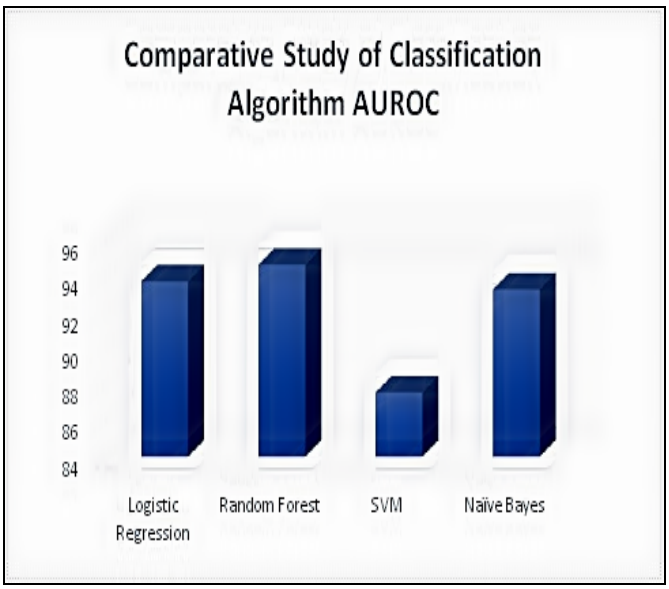

Figure 7. Comparative study of classification algorithm AUROC

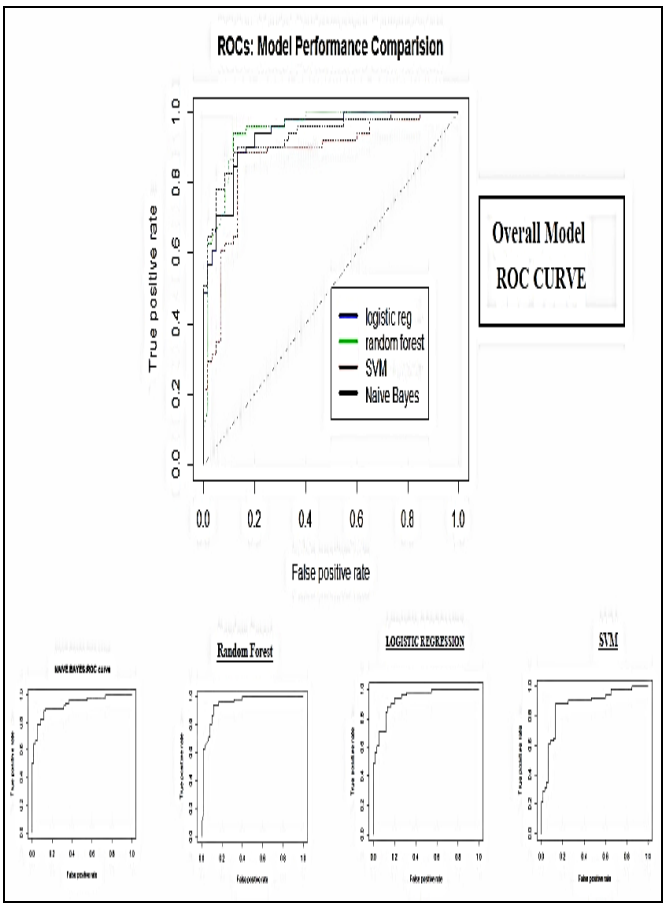

Figure 8. ROC performance comparisom

The model is validated with respect to PIMA Dataset where LR has highest accuracy (80.59%) followed by RF (79.41%) and when validated using the experimental dataset, RF has the highest accuracy (90.09%) and sensitivity (94.55%). This finally proves that in comparison to other four ML classification algorithms that are employed in this model, the Random Forest algorithm gives high accuracy in both the Indian diabetic dataset.

V. CONCLUSION

Detection of diabetes in its early stage is a principal problem in our real world. This work proposes a systematic and powerful model employing machine learning algorithms to design a scheme for predicting diabetes based on medical and non medical attributes with high accuracy. The results reveal that RF algorithm gives best accuracy i.e. 90.09% whereas KNN gives lowest accuracy of 66% with experimental dataset. Verification metric AUROC is around a range of 80-90%, which implies that the model suffers from low error rate and offers high prediction accuracy. On Validation, the model tested against PIMA Indian diabetic dataset gives the highest accuracy of 80.59 % for LR comparable to RF with 79.41%. This proves that the proposed methodology gives comparable results in terms of diabetes prediction accuracy. This will allow that individual to adopt relevant precautionary measures so that chances of suffering from diabetes can be relatively lowered or eliminated. In the near future, with the advent of Deep Learning and higher computational power the diabetic prediction can be predicted with much higher accuracy and precision.

VI. REFERENCES

[1] P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Vol 7, Issue 6, June 2017.

[2] A.Swain, S. N . Mohanty, A.C . Das "Comparative Risk Analysis on Prediction of Diabetes Mellitus using machine learning approach", International Conference on Electrical , Electronics and Optimization Techniques (ICEEOT) – 2016.

[3] W. Xu, J. Zhang, Q. Zhang, X. Wei,"Risk Prediction of type II diabetes based on random forest model", 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio – Informatics (AEEICB17), 2017.

[4] L. O. Griva, M. S Basualdo, "Evaluating clinical accuracy of models for predicting glycemic behavior for diabetes care", Argentine Conference on Automatic Control (AADECA), 2018.

[5] J. He, T. He, Y. Wang, "Blood Glucose Concentration Prediction based on Canonical Correlation Analysis", 38th Chinese Control Conference, July, 2019.

[6] C-Y. J Peng, K.L Lee, G.M. Ingersoll, " An introduction to logistic regression analysis and reporting", The International of Education Research, Vol.96, Issue. 1, 2002.

[7] N. Cristianini and J Shawe-Taylor, 2000 "An introduction to support vector machines: and other kernel-based learning methods",Cambridge university press.

[8] P.Kaviani, S. Dhotre, " Short survey on Naïve Bayes Algorithm", International Journal of Advance Research in Computer Science and Management · November 2017.

[9] G. Biau, " Analysis of a Random forests model", Journal of Machine Learning Research 13 (2012) 1063-1095.

[10] Y-L. Cai, D. Ji, D-F. Cai, " A KNN research paperclassification method based on shared nearest neighbor", Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010.