2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)

Diabetes Prediction and Classification using Machine Learning Algorithms

Yogita Dubey
Department of Electronics and Telecommunication
Yeshwantrao Chavan College of Engineering
Nagpur, India
yogeetakdubey@yahoo.co.in

Pushkar Wankhede
Department of Electronics and Telecommunication
Yeshwantrao Chavan College of Engineering
Nagpur, India
pushkar.wankhede143@gmail.com

Tanvi Borkar
Department of Electronics and Telecommunication
Yeshwantrao Chavan College of Engineering
Nagpur, India
tanvivinayborkar@gmail.com

Amey Borkar
Department of Electronics and Telecommunication
Yeshwantrao Chavan College of Engineering
Nagpur, India
amey.borkar01@gmail.com

Kajal Mitra
Dean, NKP Salve Institute of Medical Sciences and Research Center and Lata Mangeshkar Hospital
Nagpur, India
mitrakajal@gmail.com

***Abstract*—*Diabetes is one of the most grievous diseases in the world which has no remedy to cure it after a particular stage. Over 422 million people in the world are diagnosed with diabetes and many others are at jeopardy. Thus, timely diagnosis and medication is required to inhibit diabetes and its associated health problems. In this paper a framework is proposed for diabetes diseases prediction and classification using Machine Learning (ML) algorithms. The dataset is collected from Shalinitai Meghe Hospital and Research Centre, Nagpur, NKP Salve Institute of Medical Sciences and Research Centre and Mendeley Data. Four different ML algorithms Logistic Regression, Naïve Bayes, Support Vector Machine and Random Forest are applied and evaluated the model with various quantitative measures. The motive of this framework is to diagnose diabetes early and to save money and time of a patient using various machine learning approaches.*

Keywords*—*Diabetes, Prediction, Classification, Machine Learning, Model Evaluation, Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest.

I. Introduction

Because of its incessantly escalating incidence, a large number of families are controlled by Diabetes Mellitus (DM). Diabetes, a metabolic disease, is identified by excessive blood Sugar levels. It is a chronic disease distinguished by hyper-glycemia [1]. Generation of insufficient amount of insulin by the pancreas and ineffective usage of insulin in the body are both pathologic causes for DM. DM is of 2 kinds. The pathogenesis of type 1 diabetes mellitus (T1DM/Type1) is that the pancreas discharges damaged β-cells, avoiding it from reducing blood glucose level in time [2]. Insulin opposition and insulin emission deficiency are the pathogenesis of type 2 diabetes mellitus (T2DM/Type2), which is also known as insulin independent DM [3]. With the event of living conditions, diabetes is progressively ordinary in people's lifestyle. Hence there is still a scope to detect and examine diabetes rapidly and precisely. In medical science, fasting blood glucose, glucose tolerance and arbitrary blood glucose levels are used to detect the diabetes [4][5][6]. If the diagnosis happens before time, it will be much effortless to control it.

Application of machine learning algorithms can be helpful to people to make a preliminary judgment about DM consistent with their daily physical examination data, and it can function as a reference for doctors [7][8][9]. In this paper, the framework is provided for prediction and classification of diabetes using machine learning algorithms. The rest of the paper is organized as follows, section II describes the methodology used for the framework along with data pre-processing, machine learning models and evaluation metrics. Results are described in section III followed by conclusion in section IV.

II. Materials and Methods

The complete framework for diabetes prediction and classification is described in this section. Three different datasets are used to evaluate the ML model. First is from N.K.P. Salve Institute of Medical Science & Research Centre and Lata Mangeshkar Hospital, Nagpur; second is from Shalinitai Meghe Hospital & Research Center, Nagpur; and, third is from Mendeley Data [10]. Patients' files had been taken and data was obtained from them and entered into the database to assemble the diabetes dataset. The data comprises laboratory evaluation and medical evidence. The dataset contains the information about the diabetes with the parameters as: Gender, Age, Body Mass Index [BMI], High Density Lipoprotein [HDL in mmol/L], Low Density Lipoprotein [LDL in mmol/L], Triglycerides [TG], Cholesterol, Haemoglobin A1c [HbA1c], Classification (i.e., Type1, Type2, or No Diabetes). Total 1313 samples are used out of which, 60 are of Type-1, 1150 are of Type-2, and 103 are of No Diabetes. Fig. 1 shows the block diagram for the proposed framework.

978-1-6654-6621-9/21/$31.00 ©2021 IEEE

2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON) | 978-1-6654-6621-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/BECITHCON54710.2021.9893653

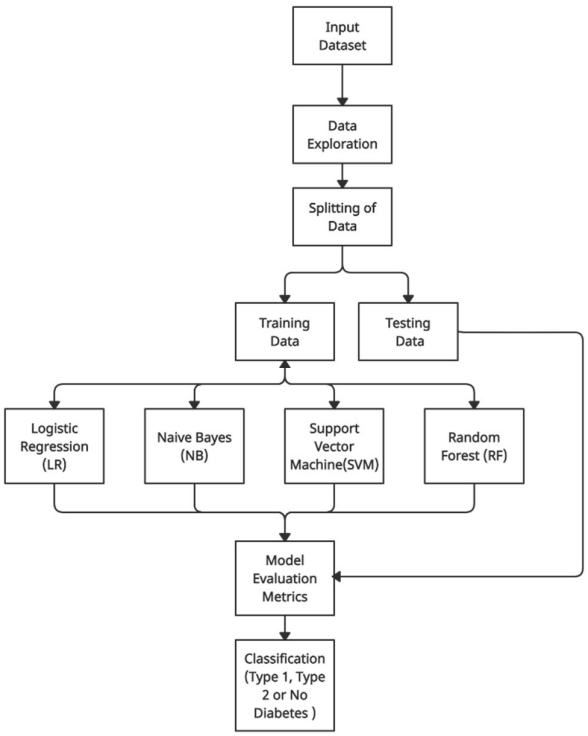

Fig. 1-Proposed Framework for Diabetes Prediction and Classification using various Machine Learning Algorithms

A. Data Pre-Processing

As our data is real world, hence it may contain noise, missing values which cannot be directly given to fit machine learning models. It contains some categorical parameters like gender and class. Therefore, we encode "Male" as 1 and "Female" as 0, "Type 1" classification as 1, "Type 2" classification as 2 and "No diabetes" as 0. Our gender parameter contains 1 NaN value, age contains 0 NaN values, BMI contains 207 NaN values, HDL contains 156 NaN values, LDL contains 160 NaN values, TG contains 157 NaN values, Cholesterol contains 151 NaN values and HbA1c contains 130 NaN values, where NaN is Not a Number. We replaced these NaN values with the mean of their respective parameters to get an optimised dataset.

B. Machine Learning Models

1) Logistic Regression (LR)

Here, we are classifying Diabetes into 3 categories as Type 1, Type 2 and No Diabetes with the help of Logistic Regression Model. Output of a class dependent variable is predicted by the LR model. The output fundamentally be categorical or discrete value [11, 12]. LR can be used to sort the interpretations using discrete kinds of data and can easily detect the most functional variables used for classification. The hypothesis used in LR model is described by Equation (1)

$$h_\theta(x) = g\left(\theta^T x\right) \tag{1}$$

Here, θ is the parameter which is required to be tuned to get the best accuracy for the features x and g is the sigmoid function or logistic function which can be given as-

$$g\left(\theta^T x\right) = \frac{1}{1+e^{-\theta^T x}}$$

which results in the range of 0 to 1.

2) Naïve Bayes (NB)

Naive Bayes algorithm is used here as it performs well in multi-class prediction. This algorithm is based on Bayes theorem of independent assumption. Bayes' theorem is mathematically modelled as:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Where, [P(A|B)] and [P(B|A)] is posterior and likelihood probabilities respectively. [P(A)] and [P(B)] is prior and marginal probability of A and B respectively. Naïve Bayes' Classifier generates a likelihood table by finding the probabilities of given features by converting the dataset into a frequency table. Then calculate the posterior probability which is used for the prediction or classification [13],

3) Support Vector Machine (SVM)

SVM algorithm is one of the most popular supervised learning algorithms and can be used for binary as well as multi-class classification. The SVM model is a representation of various classes in a hyperplane in multidimensional space. The SVM generates the hyperplane in an iterative manner hence error is minimized. The SVM divides the dataset into a number of classes to find a maximum marginal hyperplane (MMH) [14].

4) Random Forest (RF)

A random forest consists of multiple decision trees. The bagging or bootstrap aggregating is used to train 'forest' generated by the random forest algorithm. Bagging is used to improve the accuracy of machine learning algorithms; it is an ensemble meta-algorithm. It shows the result based on the predictions of the decision trees. It predicts the output from various trees by taking the mean of various trees. [15]

C. Evaluation Metrics

To assess the performance of the projected framework, different quantitative indices are used. These are discussed below.

Accuracy is the consummate classification metric use for binary and multiclass classification problems. It gives the proportion of true results among the entire number of cases examined.

$$Accuracy = \frac{|T_P + T_N|}{N}$$

Precision is an appropriate choice of evaluation metric where we might wish to be very sure of our prediction. it's calculated by taking the ratio of TP to the sum of TP and FP. Here, precision displays the figure of patients that are indicated as true (diabetic) from all the positively predicted (true) patients.

$$Precision = \frac{|T_P|}{|T_P + F_P|}$$

Recall is correctly classified proportion of actual positives. Recall is a rational selection of evaluation metric to capture positives. Here, it denotes the figure of patients predicted as true (diabetic) from all the true diabetic. It is also known as Sensitivity of the model, calculated by the ratio of TP to TP and FN.

$$Recall = \frac{|T_P|}{|T_P + F_N|}$$

F1-score fetches the balance among the precision and the recall. Thus, it accounts mutually the false positive and the false negative examinations into consideration. F1-score is determined by the subsequent formula:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Kappa Index is the ratio of difference in observed agreement (accuracy) with an expected agreement (accuracy) to one minus expected agreement. KI index should be near to 1. KI is calculated using:

$$KI = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - ExpectedAccuracy}$$

Log-loss indicates the similarity between the prediction probability and corresponding actual/true value. It is defined for only two or more labels. Higher the predicted probability differs from actual higher is the log loss value. Less value of log loss is the measure of good classification. Log loss is calculated using

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i)$$

III. Results and Discussion

This section describes the result obtained using the proposed framework on the dataset using four different ML algorithms.

Table 1: Overall Accuracy, Precision, Recall, F1-score and Kappa Index obtained using various algorithms on Diabetes dataset

Model	Accuracy	Precision	Recall	F1-Score	Kappa Index
LR	0.90	0.85	0.90	0.88	0.41
NB	0.69	0.84	0.70	0.75	0.16
SVM	0.92	0.88	0.92	0.90	0.65
RF	0.99	0.99	0.99	0.99	0.97

Table 1 summarizes the overall accuracy, Precision, Recall and F1-score and Kappa Index for each of the algorithms on the dataset. It can be seen that maximum accuracy of 99% is achieved by a Random Forest (RF). Support Vector Machine (SVM), Logistic Regression (LR) and Naive Bayes (NB) reported accuracy of 99%, 90% and 69% respectively. Maximum Precision of 99% is reported by RF. RF, LR, NB, and SVM algorithms reported recall of 99%, 90%, 70% and 92% respectively. And Finally, Maximum F1-score of 99% is reported by RF. SVM, LR and NB algorithms reported precision of 88%, 85% and 84% respectively. The maximum Kappa Index (KI) reported by RF is 97%. SVM, LR and NB reported KI as 65%, 41% and 16% respectively.

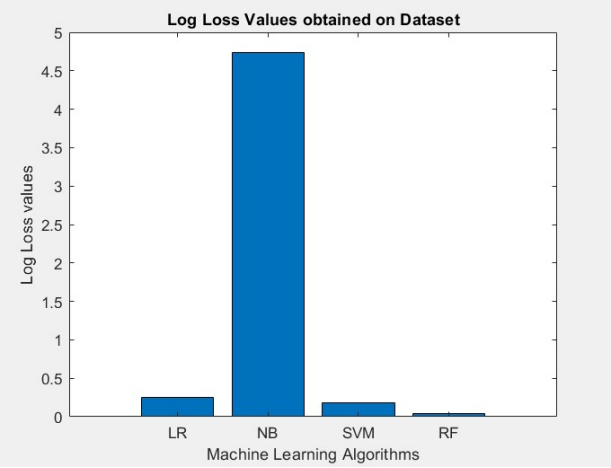

Fig.2 The graphical representation of log loss values obtained using all four machine learning algorithms

Fig 2 shows the values of log loss obtained by fitting various machine learning models. As mentioned in earlier, less the value of log loss hence better is the classification hence it is illustrated from following results. Best classification is given by RF with highest accuracy hence low value of log loss that is 0.03. SVM reported log loss of 0.17 and LR reported log loss value near 0.25 and log loss value obtained using NB algorithm is more than 4.

IV. Conclusion

Machine Learning has the immense competency to transmute the diabetes risk prognostication with the aid of highly developed computational methods along with convenience of a large amount of genetic diabetes risk dataset. Early diagnosis and treatment of diabetes is the only remedy to cure it. With the help of the framework a machine learning approach has been proposed to predict diabetes type as early-stage diabetes is dangerous disease that may also lead to premature death. To demonstrate the performance of the proposed framework, various evaluation metrics are used. The results of our method show better performance in terms of efficiency and accuracy. We have applied many Machine Learning Algorithms on diabetes dataset and the performance of those algorithms have been analyzed. The accuracy of Logistic Regression is 90%, Naive Bayes is 69% (Lowest Accuracy), SVM is 92% and that of RF is 99% (Highest accuracy).

V. REFERENCES

[1] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang "Type 2 diabetes mellitus prediction model based on data mining"

[2] Mayo Clinic Q and A: Childhood diabetes, March 31, 2021, 06:00 p.m. CDT

[3] Science Saturday: Could regenerative medicine provide a new approach to diabetes care?, Nov. 28, 2020, 12:00 p.m. CDT

[4] Iancu, I., Mota, M., and Iancu, E. (2008). "Method for the analysing of blood glucose dynamics in diabetes mellitus patients," in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883

[5] Cox, M. E., and Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. Clin. Diabetes 27, 132–138. doi: 10.2337/diaclin.27.4.132

[6] American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. doi: 10.2337/dc12-s064

[7] Lee, B. J., and Kim, J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform. 20, 39–46. doi:10.1109/JBHI.2015.2396520

[8] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., and Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. PLoS One 12:e0179805. doi: 10.1371/journal.pone.0179805

[9] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15, 104–116. doi: 10.1016/j.csbj.2016.12.005

[10] Rashid, Ahlam (2020), "Diabetes Dataset", Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1

[11] Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, "An Introduction to Logistic Regression", Indiana University-Bloomington, September 2002

[12] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[13] Pouria Kaviani, Mrs. Sunita Dhotre, "Short Survey on Naive Bayes Algorithm", International Journal of Advance Engineering and Research, Department of Computer Engineering, Bharati Vidyapeeth University, College of Engineering, Pune, November - 2017.

[14] Thandar M., Usanavasin S. 2015 Measuring Opinion Credibility in Twitter. In: Unger H., Meesad P., Boonkrong S. (eds) Recent Advances in Information and Communication Technology 2015. Advances in Intelligent Systems and Computing, vol 361. Springer, Cham.

[15] Leo Breiman, "Random Forests", Statistics Department University of Califor-nia Berkeley, CA 94720, January 2001.