DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES

* A complete guide for getting best model

Md Sazid Ahmed Tonmoy
Computer Science and Engineering
North South University
Dhaka,Bangladesh
sazid.tonmoy@northsouth.edu

Sun Debnath
Computer Science and Engineering
North South University
Dhaka,Bangladesh
sun.debnath@northsouth.edu

Showvik Saha
Computer Science and Engineering
North South University
Dhaka,Bangladesh
showvik.saha@northsouth.edu

***Abstract*—Among all the diseases nowadays, diabetes has become one the most critical and affects any age group from child to old age.[1] The leading causes of diabetes include obesity, living lifestyle, exercise, family diabetes history, not maintaining a proper diet, and many more.[2]The primary risk factor for diabetes has heart disease, stroke, kidney disease, and, most commonly seen, loss of eye vision.[3] To identify a patient who is diabetic needs to run after the hospital, but we have developed a website using machine learning techniques and some proper algorithms that would determine whether the outcome for the patient is diabetic positive or not. Other existing models have less accuracy rate, whereas we have implemented five supervised algorithms techniques which include KNN, Decision Tree, Logistic Regression, SVM, and Random Forest for the best accuracy possible. Besides this, we have also done data preprocessing and data balancing which shows Random Forest has the best accuracy outcome.[4]**

***Index Terms*—Machine Learning , Support vector machine, Artificial Neural Network, Decision Tree, Naive Bayes, Data Mining**

I. Introduction

Diabetes is becoming deadly day by day, and tension in the world health care crisis. IDF Diabetes Atlas Tenth edition 2021 says that around 537 million adults aged 20 to 79 have diabetes. By the year 2030, diabetes will rise to 643 million, and by 2045 the number might be up to 783 million.[5] 1 in 2 people among 240 million adults is currently living with diabetes and not being diagnosed. Clinically diabetes is of two types type 1 type 2. Among these two 10% of people have type 1 diabetes and the rest 90% of type 2. Type 1 one occurs for absolute deficiency of insulin and is mostly diagnosed in childhood. Type 2 deficiency of insulin production occurs due to a loss of inefficiency of insulin.[6] Common symptoms of type 1 and 2 diabetes are excessive thirst and dry mouth, frequent urination, lack of energy, blurred vision, and many more.[7] Gestational is another form of diabetes that occurs during pregnancy time, which is temporary. Clinically it is the diagnosis by measuring blood glucose levels.[8] If fasting plasma glucose goes over 7.0mmol/l and after two hours if it goes over 11.1mmol/l, it is considered a person with diabetes.[9] Diabetes is appreciations to either the exocrine gland not manufacturing plentiful hypoglycemic agent not responding properly to the hypoglycemic agent created. Various information mining algorithms presents different decision support systems for assisting health specialists. The effectiveness of the decision support system is recognized by its accuracy. Therefore, the objective is to build a decision support system to predict and diagnose a certain disease with extreme amount of precision. The AI consist of ML which is its subfield that resolves the real world difficulties by "providing learning capability to workstation without supplementary program writing.

II. Organization

The organization of this paper is followed by contribution, related work, Proposed System, result, and conclusion. The contribution describes the machine learning technique, exploratory data analysis, data balancing, and website application that we have used to develop the project. In the section of related work, we have compared our and others' work. The methodology describes in detail exploratory data analysis, data preprocessing, machine learning algorithms, and deployment. The result part will have our project related to all the results. And lastly the conclusion.

III. Contribution

Contribution to this project:

*1) **Machine learning**:* We take a supervised data set and apply five machine learning models KNN, Decision Tree, Logistic Regression, SVM, and Random Forest. Where KNN, Decision Tree, and Random Forest show the best result.

*2) **Exploratory Data Analysis(EDA)**:* After importing libraries and loading the dataset, we Check for missing values, Visualize the missing values, Replace them with missing values, and apply various types of EDA. Also, we preprocess our dataset.

*3) **Unbalanced Data**:* For testing or teaching a machine learning model, we need a data set for testing it. In our project, we used a supervised data set with an unbalanced data set, and we balanced it. For this, we change variables. Also, we use random numbers to make this a balanced data set.

*4) **A Web Application**:* We created a web application so that anyone can predict if they are diabetic or not. We use streamlit for interface desing. Also, the website has login and sign-up system. So, sqlite3 library helps us to manage a local data server.

IV. Literature Review

Veena Vijayan V. And Anjali C has discussed, the diabetes disease produced by rise of sugar level in the plasma. Various computerized information systems were outlined utilizing classifiers for anticipating and diagnosing diabetes using decision tree, SVM, Naive Bayes and ANN algorithms [8].
P. Suresh Kumar and V. Umatejaswi has presented the algorithms like Decision Tree, SVM, Naive Bayes for identifying diabetes using data mining techniques [9].
Ridam Pal, Dr. Jayanta Poray, and Mainak Sen discussed Diabetic Retinopathy (DR), which is one of the most common causes of vision loss in diabetic individuals. They examined the performance of a number of machine learning algorithms and verified their accuracy for a specific data set [10].
Dr. M. Renuka Devi and J. Maria Shyla have studied the use of Naive Bayes, Random Forest, Decision Tree, and J48 algorithms to predict diabetes utilizing various mining talents [11].
Rahul Joshi and Minyechil Alehegn explained machine learning approaches that are used to estimate datasets at an early stage in order to save lives. Using the Naive Bayes and KNN algorithms [12].
The results of methods built in order to improve diagnostic reliability have been discussed by Zhilbert Tafa and Nerxhivane Pervetica [13].
Mr. Mahale Kishor M. and Prof. Dhomse Kanchan B. have presented the research of Machine Learning Algorithms such as Support Vector Machine, Nave Bayes, Decision Tree, and PCA for Special Disease Prediction utilizing Principal Component Analysis [14].

V. Proposed System

Block diagram shows the idea how the model is being built upon using classification algorithms. The base classification algorithms are: Decision tree, Support Vector Machine for accuracy authentication.

A. Dataset Collection

Global dataset:
The training phase is completed. The dataset contains seven sixty eight instances and nine features. The dataset features are:
• Total number of times pregnant
• Glucose/sugar level
• Diastolic Blood Pressure
• Body Mass Index (BMI)
• Skin fold thickness in mm
• Insulin value in 2 hour
• Hereditary factor- Pedigree function

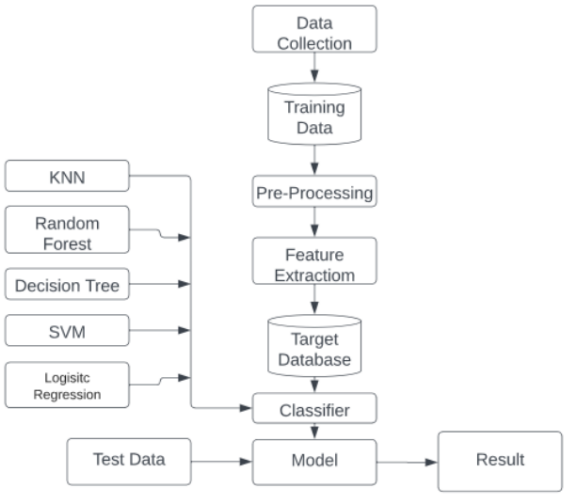


Fig. 1. Block diagram of diabetes prediction system.

• Age of patient in years
Percentage split option is provided for training and testing. Out of 768 instances 75% is used for training and 25% is used for testing.

B. Training Data and Test Data

The training data set in Machine Learning is used to train the model for carrying out abundant actions. Detailed features are fetched from the training set to train the model. These structures are therefore combined into the prototype. In sentiment analysis, single words or sequences of consecutive words are taken from the tweets. Therefore, if the training set is labelled correctly, then the model will be able to acquire something from the features. So for testing the model such type of data is used to check whether it is responding correctly or not.

C. Pre-processing

Pre-processing refers to the transformations applied to our data before providing the data to the algorithm. Data Pre-processing technique is used to convert the raw data into an understandable data set. In other words, whenever the information is gathered from various sources it is collected in raw format that isn't possible for the analysis. Fig 2 Shown below data preprocessing.

D. Feature Extraction

Feature Extraction is used to transform the input information as the outcome of features. Attribute square measures are characteristic of input designs that facilitates in differentiating between the classes of input designs. In the algorithm if the input data is too huge for processing it will be suspected to be redundant as the repeat occurrence of images which are represented as pixels, which are changed into a condense set

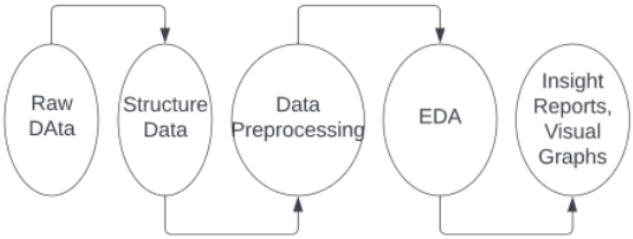


Fig. 2. Working steps diagram of diabetes prediction system.

of attribute. Using the extracted feature instead of the complete initial data the chosen task can be achieved.

E. Target Database

The target database is the database to which the new changes are moved. For example, you install the certified Upgrade Source database, referred to as demo. Then you produce a duplicate copy of your production database. You then copy the changed definitions from the Demo database into the Copy of Production. Here the Demo database is your source and the target is Copy of Production.

F. Feature Scaling

Feature scaling is a strategy for putting the data's independent characteristics into a set range. It is used to handle significantly changing magnitudes, values, or units during data pre-processing.

*1) **Robust Scaler**:* In this project, we have used a robust scaler. By using this, the median is removed, and the data is scaled according to the quantile range (defaults to IQR: Interquartile Range). The IQR is the distance between the first and third quartiles (25th and 3rd quantiles) (75th quantile). By computing the necessary statistics on the samples in the training set, each feature is individually centered and scaled. The median and interquartile range are then saved for further usage using the transform technique on new data. A basic need for many machine learning estimators is dataset standardization. This is usually accomplished by eliminating the mean and scaling to unit variance. Outliers, on the other hand, might have a negative impact on the sample mean / variance. The median and interquartile range are frequently superior in these situations.

*2) **Standard Scaler**:* StandardScaler helps when there is much deviation in input data or even if there is measurement in different units. It mainly sets the data to unit variance and cuts off the mean. But outliers create an influence while calculating mean and standard deviation, shrinking the characteristic value. We have used this technique so that the observed value for the mean would be zero, and one would be the standard deviation.

G. Machine Learning Algorithms Used:

*1) **Decision Tree**:* Decision tree is a basic classification method and also a supervised learning method.Decision tree used when the response variable is categorical. Decision tree has a tree-like structure based model which describes the classification process based on input features.Input variables are any types like graph, text, discrete, continuous etc

*2) **Random Forest**:* It is a type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is greater than compared to other models. This method can easily handle large datasets.

*3) **KNN**:* KNN is a supervised ML algorithm which helps to solve both the classification and regression problems. KNN is a lazy prediction technique and assumes that similar things are near to each other.KNN algorithm records all the records and classifies them according to their similarity measure.

*4) **Logistic Regression**:* The method of modeling the likelihood of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary result, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be used to model situations with more than two discrete outcomes. Logistic regression is a handy analytical tool for determining if a fresh sample fits best into a category in classification tasks.

*5) **SVM**:* Another simple approach that any machine learning expert should know about is the support vector machine. Many people prefer the support vector machine because it delivers great accuracy while using minimal computing resources. SVM stands for Support Vector Machine and may be used for both regression and classification. However, it is extensively utilized in categorization goals.

H. Machine learning Matrix:

*1) **Confusion Matrix**:* It gives us gives us a matrix as output and describes the complete performance of the model.

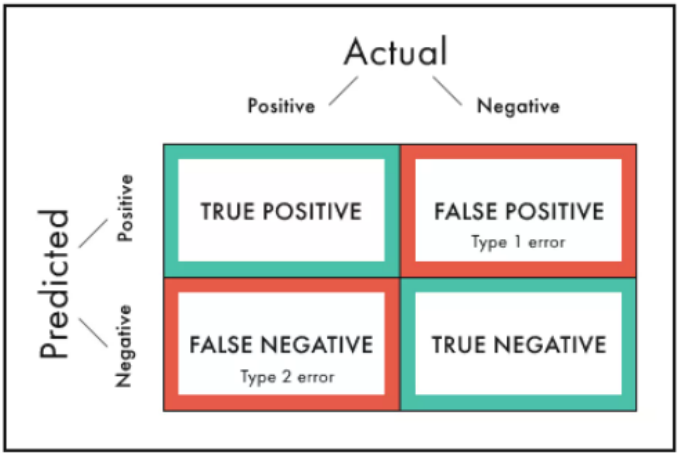


Fig. 3. Confusion matrix

Where, TP: True Positive
FP: False Positive

FN: False Negative
TN: True Negative

2) ***Precision****:* The precision can be defined as the number of TP upon the number of TP '+' number of FP. False positives are cases where the model is incorrectly tagged as positive that are actually negative.

$$Precision = TP/(TP + FP) \tag{1}$$

3) ***Recall****:* It is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as-

$$Recall = TP/(TP + FN) \tag{2}$$

4) ***F1 score****:* It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

$$F1score = 2 * (Recall * Precision)/(Recall + Precision) \tag{3}$$

I. Result Analysis

After taking the input dataset the model will predict the data by applying the ML algorithms and provide the best result in the form of comparison between to predict the best accuracy to treat diabetes.

VI. Implementation and Results

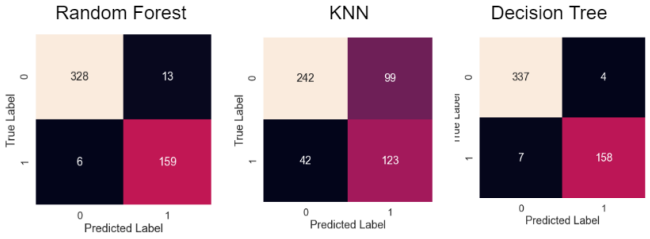
In this experimental study, five machine learning algorithms were used. These algorithms are KNN, SVM, LR, DT and RF.All these five algorithms were applied on the same dataset.Predicting accuracy is the main evaluation parameter that we used in this work.Accuracy is the overall success rate of the algorithm.

	Model	Training Accuracy	Mean Validation Accuracy	Test Accuracy
0	KNN	100.000	83.264	96.838
1	Decision Tree	100.000	82.617	97.826
2	Logisitc Regression	75.423	75.035	73.518
3	SVM	86.169	78.816	81.423
4	Random Forest	100.000	85.294	96.245

From these five models we took the three best models according to the output. We will see their confusion matrix, and classification report to compare them.
These can also be shown in Confusion Matrix which gives output and describes the complete performance of the model.

Now we are showing classification report to get enough information about those model, and their result. Also here is also our error.

- Mean Absolute Error: 0.023715415019762844



- Mean Squared Error: 0.023715415019762844
- Root Mean Squared Error: 0.15399810070180361

	precision	recall	f1-score	support
0	0.85	0.71	0.78	341
1	0.56	0.75	0.64	165
accuracy			0.72	506
macro avg	0.70	0.73	0.71	506
weighted avg	0.76	0.72	0.73	506

Fig. 4. KNN Classification Report

	precision	recall	f1-score	support
0	0.98	0.96	0.97	341
1	0.93	0.96	0.95	165
accuracy			0.96	506
macro avg	0.96	0.96	0.96	506
weighted avg	0.97	0.96	0.96	506

Fig. 5. Random Forest Classification Report

	precision	recall	f1-score	support
0	0.98	0.98	0.98	341
1	0.96	0.96	0.96	165
accuracy			0.98	506
macro avg	0.97	0.97	0.97	506
weighted avg	0.98	0.98	0.98	506

Fig. 6. Decision Tree Classification Report

VII. Conclusion

With the help of machine learning predicting diabetes at home and early stages have become the key to treatment. In this project, we have used various classification of machine learning techniques to predict diabetes as accurately as possible. We have used KNN, Decision Tree, Logistic Regression, SVM, and Random Forest as classification. Among these, we worked with Random Forest, KNN, and Decision Tree for their best accuracy values.We use Random forest for web development because it has one of the highest accuracy, and it is good for classification problem. Anyone can use this diabetes prediction website at any place.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, “Title of paper if known,” unpublished.
[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
[8] Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, “A MachineLearning Approach” ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) — 10- 12 December 2015 — Trivandrum.
[9] P. Suresh Kumar and V. Umatejaswi, “Diagnosing Diabetes using Data Mining Techniques”, International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.
[10] Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, , “Application of Machine Learning Algorithms on Diabetic Retinopathy”, 2017 2nd IEEE International Conference On Recent Trends In Electronics Information Communication Technology, May 19-20, 2017, India.
[11] Dr. M. Renuka Devi and J. Maria Shyla, “Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications. http://www.ripublication.com
[12] Rahul Joshi and Minyechil Alehegn, “Analysis and prediction of diabetes diseases using machine learning algorithm”: Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 — Oct -2017
[13] Zhilbert Tafa and Nerxhivan Pervetica, “An Intelligent System for Diabetes Prediction”, 4th Mediterranean Conference on Embedded Computing MECO – 2015 Budva, Montenegro.
[14] Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis”. International Conference on Global Trends in Signal Processing, Information Computing and Communication 2016.