Performance Analysis of Diabetes Prediction by using different Machine Learning Algorithms

Aakash Singh

Student, Electrical and Electronic Engineering, IMS Engineering College, Ghaziabad, UP, India, 201009

Abstract: ***Nowadays, daily life people are living bad lifestyle more having habit to consume more fast food and lack of health awareness. Fast food usually contains more sugar, fats, and too much oil which is the main cause of the increase of diabetes patient nowadays. Diabetes contributes to heart disease, kidney damage, nerve damage, eye damage, hearing impairment. To detect diabetes patient need to test their various bloods sample and are require visiting their nearby diagnostic center to get their report after consultation. These tests are too expensive and also takes too much time during test while diagnosis of large number of peoples. Machine learning algorithms help to predict whether a person is diabetic or non-diabetic. This paper's approaches to detect diabetes risk of patient using medical data with the help Logistic Regression, Decision Tree, Linear SVM, Random Forest, Gradient Boosting, XG boosting with better accuracy. We are using 20% of dataset for testing purpose and 80% for training purpose. The analysis result shows that gradient boosting classifier had achieved highest accuracy than other classifiers which 79 % accuracy***

Keywords: Machine Learning, logistic regression, Support vector machine, decision tree, K-Nearest Neighbours, Classification Report, Random Forest Classifier, gradient boosting classifier, XGB classifier.

1. Introduction

Diabetes is a complex and severe disease that can develop at any time during a person's life. It is a disorder of metabolism where the body has troubled using glucose for energy. People's daily diet contains huge amount of fat and sugar. Due to these factors' diabetes risk have increased among people worldwide. Due to which they had to visit there nearby health center to test their various bloods sample which are too much expensive and lot of time-consuming every year. Machine Learning Algorithms are of two types supervised and unsupervised. In supervised learning algorithms, the output for the given data is known, whereas, in unsupervised learning algorithms, the output for given input is unknown. Supervised learning is also known as classification. There are different support systems, and their effectiveness is recognized by their accuracy.

As we know that different size and kind of data are suitable for different machine learning algorithms. In this paper different machine learning Algorithms are applied on diabetes dataset. Our main objective of study is to provide enough understanding to reader about how health care industry can utilize big datasets for a better decision-making or disease prediction and also to evaluate performance of machine learning algorithms in predictive analytic for diabetes disease.

Types of Diabetes

Type 1 diabetes: disease which occurs due to failure of pancreas to supply enough hypoglycemic agent. This type of diabetes is found UN children below twenty years old. In this disease patient required to follow workout and fit regime which is recommended by doctors for type1 diabetes patient.

Type 2 diabetes: It is one of the most common types of diabetes. People having type 2 diabetes, their body does not make or use insulin well. People may suffer from type 2 diabetes at any age, even during childhood. Mostly this type of diabetes occurs most often in middle-aged and older people.

Gestational diabetes: This type of disease mostly occurs in some women when they are pregnant. Most of the time, this type of diabetes goes away after the baby is born. If you've had gestational diabetes, you have a greater chance of developing type 2 diabetes later in life. Sometimes diabetes diagnosed during pregnancy is actually type 2 diabetes.

2. Literature Review

Comparison of various literature works:

Sr	Paper Title	Author	Objective	Methodology	Conclusion
1	Diabetes Prediction Using Ensemble Perceptron Algorithm	Roxana Mirshahvalad and Nastaran Asadi Zanjani	The aim of this study is to design a more accurate classifier for diabetes diagnosis.	Learning algorithm that ensembles Boosting Algorithm with *Perceptron Algorithm* to improve performance of *Perceptron Algorithm.*	The proposed algorithm is validated on three different NHANES datasets confirming that AUC value improves from 0.72 to 0.75 by the proposed algorithm.
2	Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare	Ayman Mir and Sudhir N. Dhage	This paper aims at building a classifier model using WEKA tool to predict diabetes disease	To predict diabetes disease by employing Naive Bayes, Support, Vector Machine, Random Forest and Simple CART algorithm.	The overall performance of Support Vector machine to predict the diabetes disease is better than Naive Bayes, Random Forest and Simple Cart.

3	Prediction of Diabetes Using Machine Learning Algorithms in Healthcare	Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah	This paper aims to help doctors and practitioners in early prediction of diabetes using machine learning techniques.	Comparison of the different machine learning algorithms SVM, KNN, LR, DT, RF and NB.	SVM and KNN is appropriated for predicting the diabetes disease.
4	Diabetes prediction using different machine learning approaches	Prof. K. JayaMalini, And Priyanka Sonar	The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with better accuracy.	Model development is based on categorization methods as Decision Tree, ANN, Naive Bayes and SVM algorithms.	ANN: Gives good prediction and easy to implement.
5	Classification Of Diabetes Disease Using Support Vector Machine	V. Anuja Kumari and R . Chitra	testing data mining algorithms to see their prediction accuracy in diabetes data classification.	The proposed method uses Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes	The performance parameters such as the classification accuracy, sensitivity, and specificity of the SVM and RBF have found to be high thus making it a good option for the classification process

3. Methodology

It is a procedural framework which contains various step like shown above in the block diagram for finding better accuracy using machine learning Algorithms

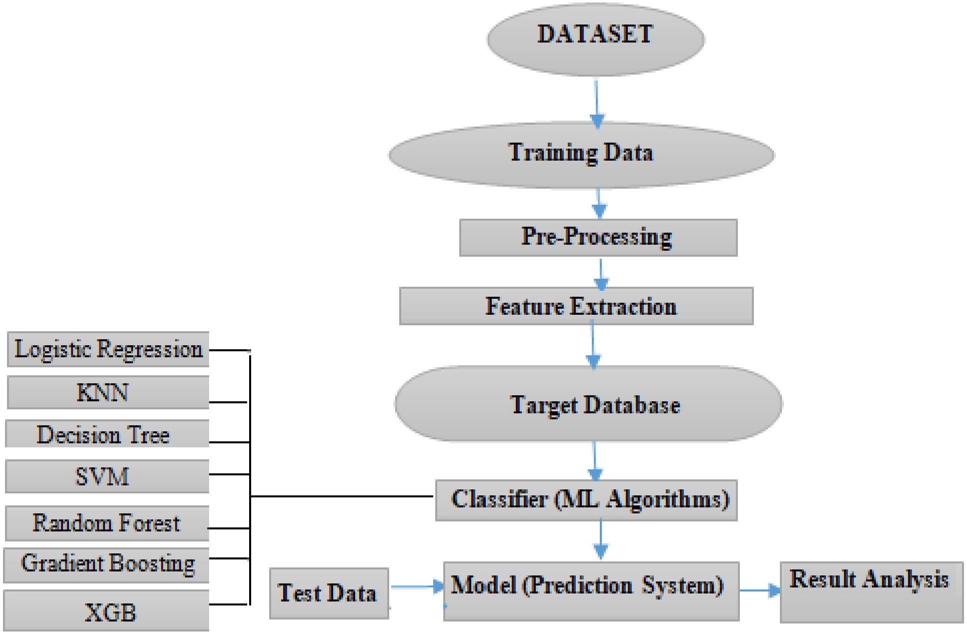


Figure 1.1: Block Diagram of Diabetes Analysis

3.1 Data Collection

This dataset contains 768 instances and 9 features and the datasets features are described in table 1.1 Out of 768 instances present in dataset we had used 20% for testing and 80% for training by using train_test_splitting

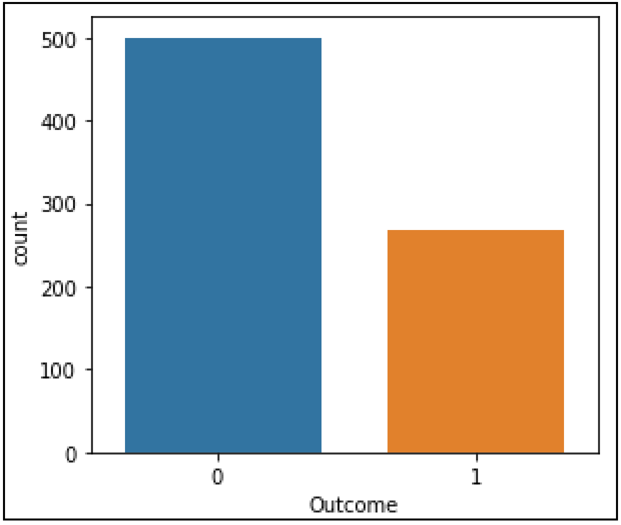
Table 1.1: Diabetes Disease Dataset

Attribute	*Description*
Pregnancies	No of pregnancies
Glucose	Plasma glucose concentration
Blood pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg / (height in square m)
Diabetes pedigree Function	Diabetes pedigree function
Age	Age (years)
Outcome	Diabetic or non-diabetic

Table 1.2: Data Statistics

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
Count	768	768	768	768	768	768	768	768	768
Mean	3.84	120.89	69.10	20.53	79.79	31.99	0.47	33.24	0.34
STD	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76	0.47
MIN	0	0	0	0	0	0	0.07	21	0
MAX	17	199	122	99	846	67.10	2.42	81	1

Analysis of Patient



Here when outcome =0 then diabetic person, outcome=1 then non-diabetic person

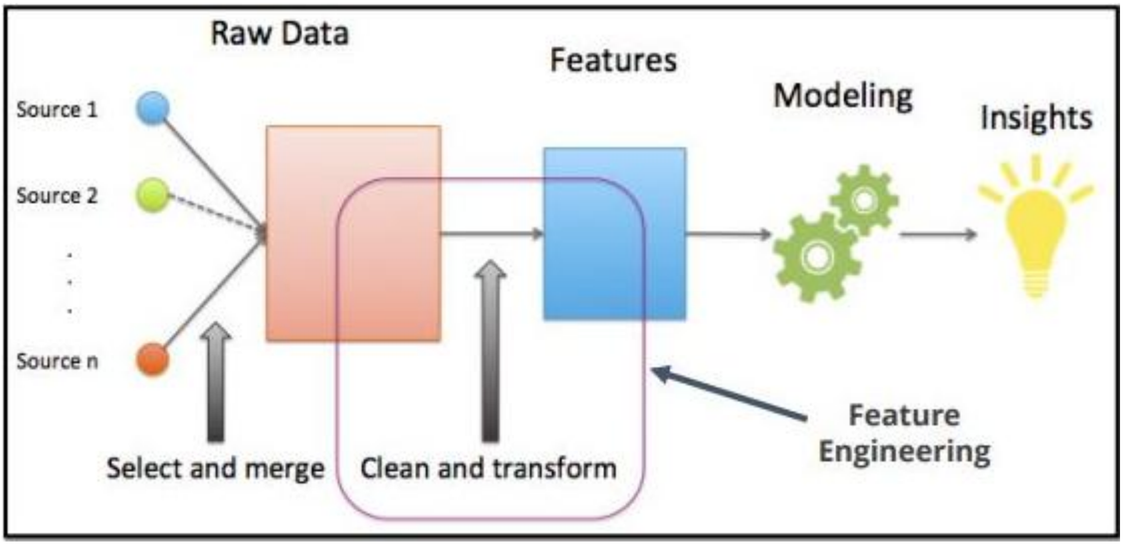
3.2 Training Data and Test Data

Training data: It is an actual dataset that we used to train our model. It is used to train the model for carrying out abundant actions.

Testing data: It is used when our model is completely trained and is generally what is used to evaluate our competing models.

3.3 Preprocessing

Data Preprocessing refers to the technique which is used to convert the raw data into an understandable data set. Which is provided to our data before providing any algorithms. So it is an extremely important that we preprocess our data before feeding it into our model.



3.4 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

3.5 Machine Learning Algorithms Used:

3.5.1 Logistic Regression

Logistic regression is also called logistic model or logit regression. It takes in independent features and returns output as categorical output. The probability of occurrence of an categorical output can also be found by logistic regression model by fitting the features in the logistic curve. The Logistic Regression model can be replaced by the simpler Linear Regression model when the output variable is taken to be continuous. Logistic Regression model was chosen over the other models because of its mathematical clarity and flexibility. This model can have single or multiple predictors.

3.5.2 KNN

K-nearest neighbours (KNN) is more widely used for classification problems in the industry. KNN algorithm fairs across all parameters of considerations. It is most commonly used for its easy of interpretation and low calculation time. KNN algorithm uses ‘feature similarity’ to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. By Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate is Euclidean distance.

3.5.3 Decision Tree

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification.

Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses every node by evaluating the highest information gain among all the attributes.

3.5.4 SVM

SVM is one of the standard sets of supervised machine learning model employed in classification. Given at two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes. For a better generalization hyperplane should not lie closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors. The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyperplane**.**

3.5.5 Random Forest

Random forest is a supervised machine learning algorithm. It can be used for both classification and regression algorithm. It is a most flexible and easy to use algorithm. Random forest comprises trees. It is said that as the number of trees increases it becomes more Robust. Random forest creates decision trees from randomly selected data sample and make prediction from each tree and selects the best solution by means of voting. The random forest is composed of multiple decision trees. By averaging out the impact of several decision trees, random forest tend to improve prediction.

3.5.6 GRADIENT BOOSTING CLASSIFIER:

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive model problems. Gradient boosting is also known as gradient tree boosting, stochastic gradient boosting, and gradient boosting machines, or GBM for short.

It is age neralization of boosting to arbitrary differentiable loss functions. It is an effective machine learning algorithm and is often the main, or one of the main, algorithms used to win machine learning competitions on tabular and similar structured datasets.

3.5.7 XGB CLASSIFIER

XG Boost (extreme Gradient Boosting) is an advanced implementation of the gradient boosting algorithm. XG Boost has proved to be a highly effective ML algorithm, extensively used in machine learning competitions and hackathons.

XG Boost has high predictive power and is almost ten times faster than the other gradient boosting techniques. It also includes a variety of regularization which reduces over fitting and improves overall performance. Hence, it is also known as regularized boosting technique in ensemble learning.

XG Boost allows users to define custom optimization objectives and evaluation criteria adding a whole new dimension to the model. XG Boost has an in-built routine handle values. Here user is required to supply a different value than other observations and pass that as a parameter. XG Boost tries different things as it encounters a missing value one each node and learns which path to take for missing values in the future.

4. Experimental Results

This section describes the experimental results that are obtained after training Logistic Regression, KNN, Decision Tree, Support Vector Machine, Random Forest classifier and Gradient Boosting and XGB classifier on the diabetes patient dataset. The purposes of these experimental results are for performance evaluation of all four classifier and to recommend the best algorithm suited for prediction.

4.1 Tabular form comparison of Models

SR	Model	Mean Absolute Error	R2_Score Error	Accuracy
1	Logistic Regression	0.23	0.01	77%
2	KNN	0.31	-0.32	69%
3	Decision Tree	0.30	-0.26	70%
4	Linear SVM	0.24	-0.01	76%
5	Random Forest	0.24	-0.02	76%
6	Gradient Boosting	0.21	0.09	79%
7	XG Boosting	0.25	-0.04	75%

4.2 Classification report for diabetic patient

SR	Model	Precision	Recall	F1_Score
1	Logistic Regression	0.75	0.75	0.75
2	KNN	0.70	0.70	0.70
3	Decision Tree	0.74	0.74	0.74
4	Linear SVM	0.75	0.75	0.75
5	Random Forest	0.75	0.75	0.75
6	Gradient Boosting	0.79	0.79	0.79
7	XG Boosting	0.76	0.76	0.76

4.3 Classification report for non-diabetic patient:

SR	Model	Precision	Recall	F1_Score
1	Logistic Regression	0.81	0.81	0.81
2	KNN	0.65	0.65	0.65
3	Decision Tree	0.63	0.63	0.63
4	Linear SVM	0.79	0.79	0.79
5	Random Forest	0.81	0.81	0.81
6	Gradient Boosting	0.77	0.77	0.77
7	XG Boosting	0.73	0.73	0.73

5. Result and Discussion

In this paper, we had used seven different machine learning Algorithms Logistic Regression, KNN, Decision Tree, Linear SVM, Random Forest, gradient boosting, XG boosting for diabetes prediction on diabetes Dataset which includes 9 features and 768 instances, and we had used 20% dataset for testing and 80% for training purpose from experimental result obtained, it can be seen that gradient boosting classifier gives highest accuracy for predictive models.

Gradient boosting classifier provides 79% accuracy which is highest as compared to other algorithms used in this paper. Therefore, it can be concluded that gradient boosting classifier is appropriated for predicting the diabetes disease.

Some limitations of this study are the size of dataset and missing attribute values. To build a prediction model for diabetes with 99.99% accuracy, we will need thousands we will need thousands of records with zero missing values. Our future work will focus on integration of other methods into the used model for tuning the parameters of models for better accuracy. Then testing these models with large dataset having minimum or no missing attribute values will reveal more insights, and better prediction accuracy.

References

[1] Ayman Mir, Sudhir N. Dhage" Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare".2018 ieee Sardar Patel Institute of Technology Mumbai, India

[2] Prof. K. JayaMalini, Priyanka Sonar, "*Diabetes prediction using different machine learning approaches ".2019* Mumbai University, Mumbai, India

[3] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, " Prediction of Diabetes Using Machine Learning Algorithms in Healthcare". 2019 School of Electrical Engineering and Computer Science, NUST. Pakistan.

[4] Roxana Mirshahvalad, Nastaran Asadi Zanjani, " Diabetes Prediction Using Ensemble Perceptron Algorithm".2017, Eastern Mediterranean University Famagusta, Cyprus

[5] H. Daume, A Course in Machine Learning, 1st ed., United States: TODO, 2015.

[6] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques.", 2007.

[7] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed., New York: Wiley, 2001.

[8] G. I. Webb, Z. Zheng, "Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques.", IEEE Transactions on Knowledge and Data Engineering. Aug. 2004, 16(8), pp. 980-991.

[9] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms.", Pattern recognition 30, no. 7, Jul. 1997, pp.1145-1159.

[10] H. Altay Guvenir, and M. Kurtcephe, "Ranking Instances by Maximizing the Area under ROC Curve", IEEE Transactions on Knowledge and Data Engineering, vol.25, no.10, pp. 2, October 2013.

[11] Niharika G. Maity, Dr. Sreerupa Das, "Machine Learning for Improved Diagnosis and Prognosis in Healthcare", IEEE 2017

[12] Emrana Kabir Hashi, Md. Shahid Uz Zaman, Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using ClassificationTechniques", International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, IEEE

[13] Md. Golam Rabiul Alam, Rim Haw, Sung Soo Kim, Md. Abul Kalam Azad, Sarder Fakhrul Abedin, Choong Seon Hong, "EM-Psychiatry: An Ambient Intelligent System for Psychiatric Emergency", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12, NO. 6, DECEMBER 2016

[14] Brett Hannan, Xiaoqin Zhang, Kristen Sethares, "iHANDs: Intelligent Health Advising and Decision-Support Agent", 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

Author Profile

Aakash Singh is B.Tech 2nd year student in Department of Electrical and Electronic IMS Engineering College, Ghaziabad, UP, India. His area of Interest is Programming in Python & Machine Learning.