ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

A Comparison of Machine Learning Algorithms for Diabetes Prediction using Pima Indian Dataset

Athira Rajendran [1]

Department of Computer Science
Sree Narayana College Cherthala
University of Kerala
Thiruvananthapuram 695034,India

151athira@gmail.com

Vishnupriya Mohan [2]

Department of Computer Science

Cochin University of Science and Technology
Kochi, Kerala 682022, India

Vishnupriyamohan992@gmail.com

***Abstract*—Diabetes is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. It is a very common and serious disease in many American Indian tribes, Indians, and many other populations in the world. Several well- known risk factors such as parental diabetes, genetic markers, obesity, and diet are considered as the main risk factors for diabetes mellitus, while the precise nature of the gene or genes remains unknown.**

***Index Terms*—Machine learning, Diabetes patients, Pima Indian Diabetes, Classification algorithms**

I. INTRODUCTION

Diabetes Mellitus also called as "Sugar diabetes "is a condition that occurs when the body cannot use glucose normally. The main source of energy in the body's cells is glucose and its level is controlled by a hormone called insulin. Type1 diabetes is a condition in which our immune system destroys insulin making cells in our pancreas called beta cells. This condition is usually diagnosed in children and young people, so it's also called as juvenile diabetes. Type1 diabetes symptoms are extreme thirst, increased hunger, dry mouth, frequent urination, fatigue, upset stomach and vomiting etc. Type2 diabetes mellitus is a chronic disease. It is usually called as adult onset diabetes because it used to start always in middle and late-adulthood. It is characterized by high levels of sugar in the blood and type2 diabetes is much more common than type 1 diabetes. Its symptoms are excessive urination, hunger and thirst, increased susceptibility to infections especially yeast or fungal infections, weight loss, dizziness etc. Diabetes is a common medical complication during pregnancy would be an understatement; it is a major medical problem for both over and under – fed pregnant populations. It is a major cause of prenatal morbidity & mortality and a significant contributor to bad obstetric history (BOH).50% of GDM patients develop type 2 Diabetes in next 20 years, so the long term complications too cannot be ignored1 Incidence is 1% - 14%2 and varies according to ethnicity, selection criteria and diagnostic test Asians data suggests a local incidence of 5-8%. 90% of them are of Gestational onset and Type 1 diabetes occurs in 7.5%. To study the reason that leading to diabetes, a cluster of dataset about Pima Indian Diabetes was collected. It is consisted of 8 predict variables and 1 response variable. The variables are

PRG, PLASMA, BP, THICK, INSULIN, BODY, PEDIGREE and AGE. After randomly selecting 700 observations from 768 patients, 9 variables were taken to fit a generalized linear model to predict the probability that individual females have diabetes. Then, using stepwise selection provided subgroups of characteristics with higher risk of diabetes. Diabetes Prediction deals with the problem statement that correctly classifies and predicts whether a female has diabetes or not. The research people are Pima Indian females.

II. LITARATURE REVIEW

The PID database availed from UCI Machine Learning Repository, consists of two categories namely tested positive and tested negative. It has 8 features as : number of times pregnant, plasma glucose concentration at 2-hours in an oral glucose tolerance test, diastolic blood pressure (mmHg), triceps skin fold thickness (mm), 2-Hourserum insulin (mu U/ml), body mass index(weight in kg/(height in m)2), diabetes pedigree function and Age (years). A Research Paper given by Sudajai Lowanichchai, SaisuneeJabjone, Tidanut Puthasimma, Informatics Program Faculty of Science and Technology Nakhon Ratchsima Rajabhat University it proposed the appli-cation Information technology of knowledge-based DSS for analysis diabetes of elder using decision tree.

The result showed that the Random Tree model has the highest accuracy in the classification is 99.60 percent when compared with the medical diagnosis that the error MAE is 0.004 And RMSE is 0.0447. The NBTree model has lowest accuracy in the classification is 70.60 percent when compared with the medical diagnosis that the error MAE is 0.3327 and RMSE is 0.454 [1]. In another Research paper presented by Yang Guo, GuohuaBai, Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The dataset used was the Pima Indian diabetes dataset. Preprocessing was used to

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

improve the quality of data. Classifier was applied to the modified dataset to construct the Naïve Bayes model. Finally weka was used to do simulation, and the accuracy of the resulting model was 72.3%. [2]. In a Research paper presented by Ashwinkumar U.M and Dr. Anandakumar K.R. Reva Institute of Technology and Management, Bangalore S J B Institute of Technology, Bangalore. This Paper has proposed a novel learning algorithm i+ Learning as well as i+ LRA, which apparently achieves the highest classification accuracy overID3 algorithm. Literature Review on Diabetes, by National Public health: Women tend to be hardest hit by diabetes with 9.6million women having diabetes. This represents 8.8% of the adult population of women 18 years of age and older in 2003 and a two fold increase from 1995 (4.7%). By 2050, the projected numbers of all persons with diabetes will have increased from 17 million to 29 million.

III. METHODOLOGY

A. *About Dataset*

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- 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 m)^2)
- Diabetes Pedigree Function: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

TABLE I
PATIENT DETAILS (FEMALE) AGED AT LEAST 21 AND ABOVE

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

B. *Data Acquisition*

When encountered with a data set, first we should analyze and "get to know" the data set. This step is necessary to familiarize with the data, to gain some understanding about the potential features and to see if data cleaning is needed. First we will import the necessary libraries and import our data set to the Jupyter notebook. We can observe the mentioned columns in the data set. Visualization of data is an imperative aspect of data science. It helps to understand data and also to explain the data to another person. Python has several interesting visualization libraries such as Matplotlib, etc. We can observe that the data set contain 768 rows and 9 columns. 'Outcome' is the column which we are going to predict, which says if the patient is diabetic or not. 1 means the person is diabetic and 0 means person is not. We can identify that out of the 768 persons, 500 are labeled as (non-diabetic) and 268 as 1 (diabetic).

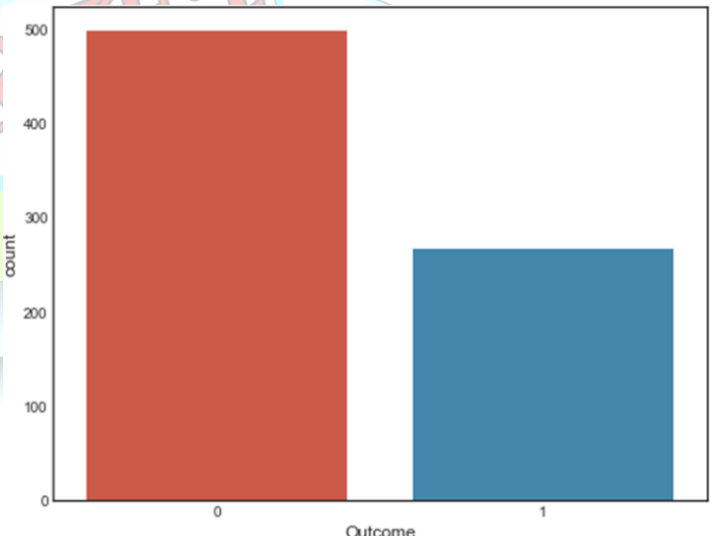

Fig.2 Predicted outcome of diabetes

C. *Pre-Processing*

1) *Missing Values:* One relevant problem in data quality is the presence of missing data. Missing data may have

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

different sources such as death of patients, equipment malfunctions, and refusal of respondents to answer certain questions, and so on.

2) *Data Cleaning:* Next phase of the machine learning work flow is the data cleaning Considered to be one of the crucial steps of the work flow, because it can make or break the model. There is a saying in machine learning "Better data beats fancier algorithms", which suggests better data gives you better resulting models.

3) *Outlier Detection and Treatment:* Presence of outlier has considerable effect on the accuracy of prediction model. Outliers can be detected in the PID dataset using box plot. The attribute serum insulin has the large number of outliers therefore the corresponding data 36 cases are eliminated from data set leaving with 498 cases for the modeling. In this data set 700 cases have label tested negative, while 157 cases have label tested positive.

4) *Model Selection:* Model selection or algorithm selection phase is the most exciting and the heart of machine learning. It is the phase where we select the model which performs best for the data set at hand .First we will be calculating the "Classification Accuracy (Testing Accuracy)" of a given set of classification models with their default parameters to determine which model performs better with the diabetes data set. We imported the necessary libraries to the notebook. We import 7 classifiers namely K-Nearest Neighbors, Support Vector Classifier, Logistic Regression, Gaussian Naive Bayes, Random Forest and Gradient Boost to be contenders for the best classifier. Train/Test Split, This method split the data set into two portions: a training set and a testing set. The training set is used to train the model. And the testing set is used to test the model, and evaluate the accuracy. K-Fold Cross Validation, This method splits the data set into K equal partitions ("folds"), then use 1 fold as the testing set and the union of the other folds as the training set. Then the model is tested for accuracy. The process will follow the above steps K times, using different fold as the testing set each time. The average testing accuracy of the process is the testing accuracy.

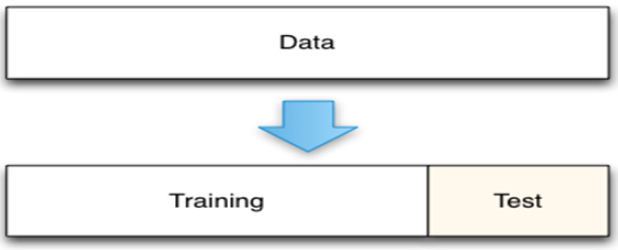

Fig.2. Steps in Random Forest and Gradient Boost classifier

D. Visualization

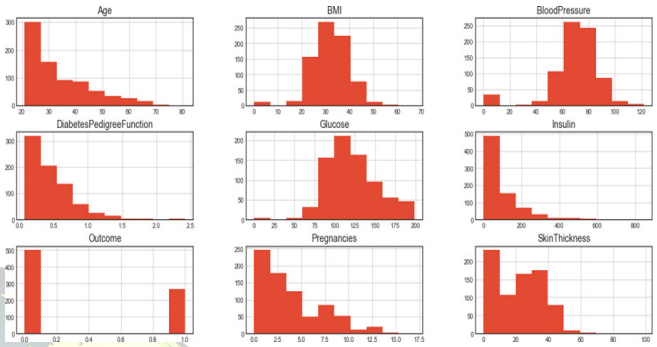

Fig.3 Histogram for the attributes which indicates the distribution

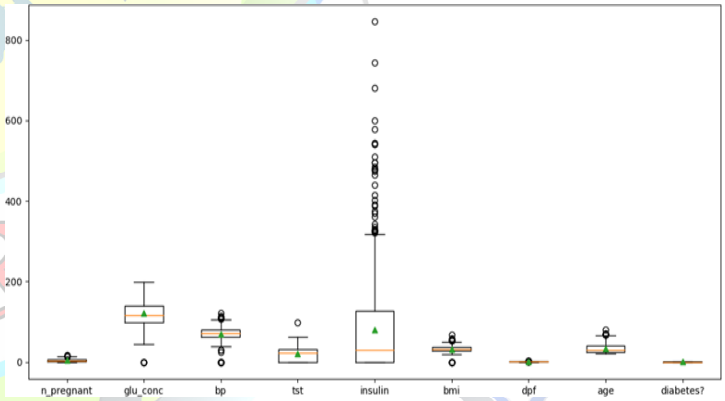

Fig.4 Box plot without normalization

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

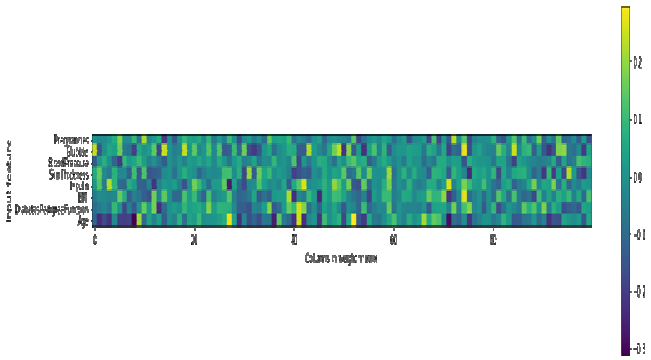

Fig.6 Attributes columns in weight matrix

Properties	Value
Number of Samples	768
Number of attributes	8
Number of classes	2
Type of attributes	Numeric
Type of Class attributes	Binomial

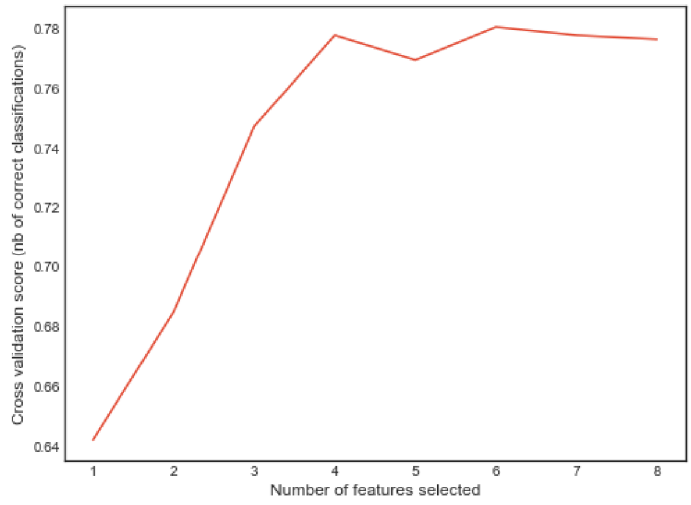

Fig.5 Gradient Boost CV v/s No of features

IV. RESULTS AND DISCUSSIONS

A. *Dataset Description*

The dataset chosen for this work is Pima Indian Diabetes dataset because it has been widely studied and also because it is considered a difficult set. The source of the database is National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients in this dataset are females of at least 21 years old of Pima Indian heritage. In the Pima Indian Diabetes data set, an instance represents one report of diabetes test. Each instance is characterized by 8 attributes, and each instance is classified as either positive or negative. There are 768 instances in the data set in total. The class is distributed as follows: Class value 1 is interpreted as "tested positive for diabetes" and class value 2 is interpreted as “tested negative for diabetes”. In other words, the class label represents if the person has not diabetes (tested negative) or the person has diabetes (tested positive).

TABLE II
REPRESENTS THE DESCRIPTION OF DATASET USED

There are 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’. There are eight clinical findings which are listed as number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mmHg),triceps skin fold thickness (mm), two hour serum insulin (mu U/ml), body mass index, diabetes pedigree function, and age (years).

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

B. Applying Classification Algorithm Without Feature Selection

Applying various classification algorithms such as Support Vector Machine, logistic Regression, Gradient Boosting and Random Forest classifier tree on the original Pima Indian Diabetes Patient Datasets

- K-NN Classifier
  Accuracy of K-NN classifier on training set: 0.79
  Accuracy of K-NN classifier on test set: 0.78

- Logistic Regression
  Training set score: 0.781
  Test set score: 0.771

- Decision Tree Classifier
  Accuracy on training set: 1.000
  Accuracy on test set: 0.714

- Random Forest Classifier
  Accuracy on training set: 1.000
  Accuracy on test set: 0.786

- Gradient Boosting Classifier
  Accuracy on training set: 0.917
  Accuracy on test set: 0.792

- Support Vector Machine Classifier
  Accuracy on training set: 1.00
  Accuracy on test set: 0.65

C. Applying Feature Selection & Classification Algorithm After Feature Selection

Attribute or feature selection is done with the help of greedy stepwise approach. The whole datasets of diabetes patients is comprised of all relevant or irrelevant attributes. By the use of feature selection, a subset (data) of diabetes patient from whole diabetes patient datasets will be obtained which comprises only significant attributes. Applying feature selection or attribute selection using Greedy Stepwise Technique on 9 attributes.

- K-NN Classifier
  Accuracy of K-NN classifier on training set: 0.79
  Accuracy of K-NN classifier on test set: 0.78

- Logistic Regression
  Training set accuracy: 0.785
  Test set accuracy: 0.766

- Decision Tree Classifier
  Accuracy on training set: 1.000
  Accuracy on test set: 0.714

- Random Forest Classifier
  Accuracy on training set: 0.800
  Accuracy on test set: 0.755

- Gradient Boosting Classifier
  Accuracy on training set: 0.804
  Accuracy on test set: 0.781

- Support Vector Machine Classifier
  Accuracy on training set: 0.790
  Accuracy on test set: 0.797

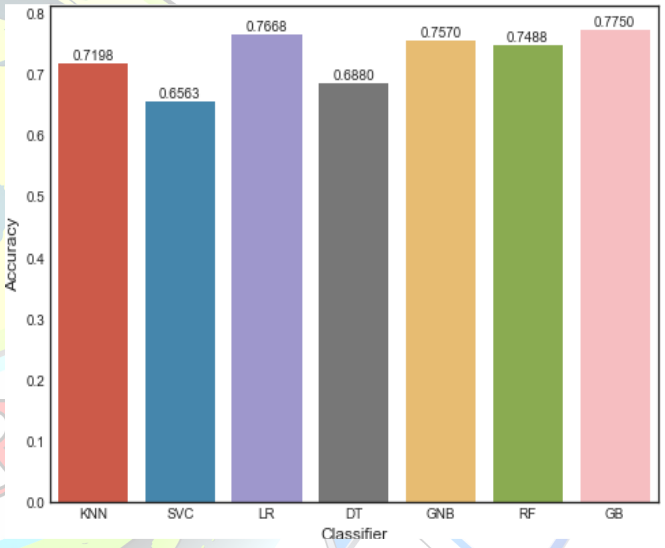

Fig.7 Comparison of classifiers

We can see the Logistic Regression, Gaussian Naive Bayes, Random Forest and Gradient Boosting have performed better than the rest. From the base level we can observe that the Logistic Regression performs better than the other algorithms.

V. CONCLUSION

The classification performance depends on the quality of the data. A data pre- processing is done properly then accuracy of the classifier may get increased. Data is to be used for classification is selected by clustering algorithm where cases which are correctly grouped are only considered for classification. In this report, initially, the dataset, 700 total observations, was randomly selected. Then, according to the picked dataset, the probability that individual females have diabetes was predicted. We finally

find a score of 77 % using Gradient Boosting Classifier and parameters optimization. Based on the feature importance, Glucose is the most important factor in determining the onset of diabetes followed by BMI, Age and Other factors such as Diabetic Pedigree Function, Pregnancies, Blood Pressure, Skin Thickness and Insulin also contributes to the prediction.

ACKNOWLEGDMENT

We would also like to show our gratitude to Dr. B Kannan, Cochin University of Science and Technology for sharing their pearls of wisdom with us during the course of this research, and we thank "anonymous" reviewers for their so-called insights. We are also immensely grateful to Ms. Keerthy A.S, Cochin University of Science and Technology for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons. We thank Dr. Bindu N, Associate Professor Sree Narayana College Cherthala for valuable comments that greatly improved the manuscript. We thank our colleagues from Sree Narayana College Cherthala who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

REFERENCES

[1] A. G. Karaganda, Punya V, M.A. Jayaram and A.S. Manjunath, "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5", *International Journal Of Computer Applications* ISSN: 0975 – 8887, Vol.45, May 2012.

[2] B. M. Patil, R.C. Joshi, Durga Toshniwal, "Hybrid prediction model for Type-2 diabetic patients", *Expert Systems with Applications*, Vol.37 ISSN: 8102–8108, 2010.

[3] Gustavo E. A. P. A. Batista and Maria Carolina Monard, University of Sao Paulo, "A Study of k- Nearest Neighbour as an Imputation Method" *Soft computing systems-designs, management and applications,* HIS 2002.