2020 IEEE INTERNATIONAL CONFERENCE ON
ADVANCES AND DEVELOPMENTS IN ELECTRICAL AND ELECTRONICS ENGINEERING (ICADEE 2020)

Comparison of Different Machine Learning Models for diabetes detection

Rahul Katarya
Department of computer science
Delhi technological University
New Delhi, India
rahulkatarya@dtu.ac.in

Sajal Jain
Department of computer science
Delhi technological University
New Delhi, India
sj31296@gmail.com

2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE) | 978-1-7281-9251-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICADEE51157.2020.9368899

***Abstract*—Diabetes metilus which is commonly known as diabetes is a major metabolic disorder which has a severe effect on a human being. Diabetes results in high blood sugar. In a human body, there is a hormone which is secreted by the pancreas called insulin which helps to move the glucose from the blood to the cells which are used for energy later. In diabetes, one's body doesn't produce insulin inadequate amount or is impotent to use insulin effectively. When diabetes is not treated properly the danger of heart attack, retinopathy or vision loss, skin conditions and some other disorders increases. There are more than a million people currently who are suffering from this disease. The detection of diabetes in early stages can help one to take appropriate measures. The rapid increase in the number of people suffering from diabetes is gaining everyone's attention. The subset of artificial intelligence is Machine learning(ML) in which the system learns from the experience without doing any explicit programming. In this research, we have applied the machine learning technique for the detection of patterns and risk factors in Pima Indian diabetes dataset using python data manipulation tool. For the categorization of the patient into diabetic or non-diabetic, we have applied six machine learning algorithms specifically support vector machine(SVM), k-nearest neighbour (KNN), Gradient boosting, Decision tree, Random forest and logistic regression.**

Keywords—: Diabetes Mellitus; Big Data Analytics; Healthcare; Machine Learning

I. INTRODUCTION

The ongoing changes in the field of biotechnology and the generation of data in massive amount take the area of computational biology in the area of big data. The source of data in the healthcare sector is magnetic resonance imagery(MRI), mass spectrometry etc. these technologies produce a large amount of data but do not provide any analysis or interpretation of that data. In today's world, knowledge discovery in the bionic data is very important and necessary[1]. The main goal is to Analyze the growth in the bionic data and develop a model to increase the results for the simple queries in the medical field**.** The extent to which a technique identifies patterns and help us to create a model from the data will define its effectiveness and accuracy[2]**.** The availability of the data in a large number of results in the strengthening of data-oriented research in the field of biology**.** In this type of complex field**,** the most important application of this is in early detection of a life-threatening disease for a human being**.** Diabetes Mellitus (DM) is one of those diseases. Diabetes which is also known as Diabetes metilus is a chronic disease and millions of people all around the world are suffering from it. The major symptom of diabetes is the high glucose level for a long duration of time. Too much urination and increase in hunger are the identification of high glucose. If diabetes is not diagnosed in the early phase it can lead to various health-related problems or even can lead to one's early death. This will eventually lead to problems like eye complications, skin conditions etc. In a human body, there is a hormone which is secreted by the pancreas called insulin which helps to move the glucose from the blood to the cells which are used for energy later [3]. Diabetes is caused when one's body is unable to make insulin sufficiently or when the cells and tissues in the body are unable to make use of it.DM is categorized in the following

If a patient **is** suffering from type-1 diabetes the body of patient doesn't produce insulin in adequate amount and we have to externally inject insulin in a patient's body.

When the cells in a patient's body are unable to make use of the insulin produced then we know he is suffering type-2 diabetes which is also referred to as Non-Insulin-Dependent Diabetes Mellitus (NIDDM).

Another type of diabetes is Gestational diabetes which is spotted in a woman's body when she is pregnant and there is an increase in her blood sugar level where diabetes is not detected earlier[4].

An analysis which consists of machine learning algorithms together with data mining techniques and statistical methods Is known as predictive analysis. If we apply this analysis on healthcare data meaningful results can be obtained and important decisions can be taken based on that. This analysis is carried out by utilizing machine learning and regression techniques[5]. Machine learning is an Artificial intelligence which helps us to develop systems that can make the predictions based on previously acquired knowledge. As time is passing by the use of machine learning approaches is also increasing in every field. Machine learning will help to increase automation and thus reducing human efforts in every field. The detection of diabetes is done using various lab tests like oral glucose tolerance test (OGTT), Urine test, RPG test etc. This paper mainly focuses on developing a diabetes prediction model by utilizing different machine learning algorithms[6].

II. METHODOLOGY

In the first step, we collect the contextualized Pima Indian diabetes dataset. To understand the sources from which our data is collected the exploratory data analysis is carried. In the next step, we have to preprocess our data that is we have to clean our dataset in which we have to remove the duplicate, missing or unusual values present in our dataset. In the next step, we have to select the models for the training of our data and to fit the model. Then, the models will be compared based

978-1-7281-9251-2/20/$31.00 ©2020 IEEE

on different performance metrics like accuracy, f1-score, Recall etc. The proposed methodology is shown in the figure which shows the steps followed in the implementation[7].

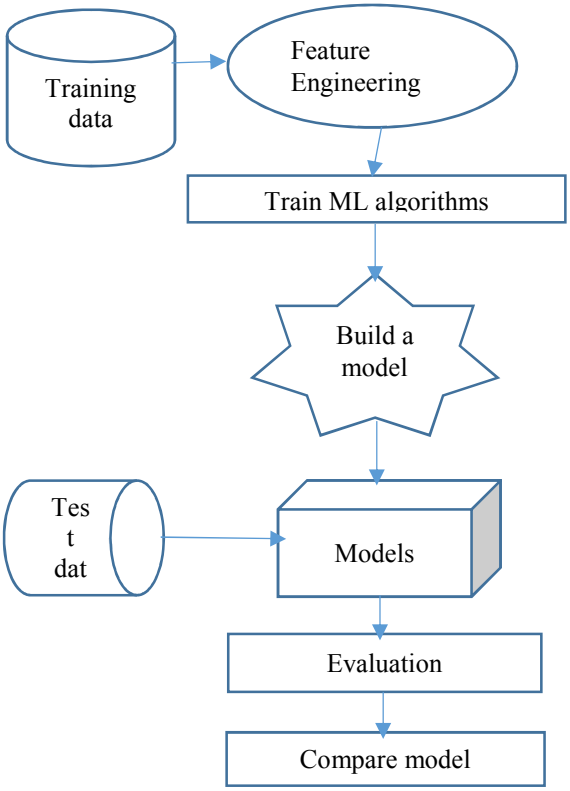

Fig 1 Proposed framework

III. Brief Description Machine Learning Techniques

K-NN (K-nearest neighbour)

K-nearest neighbour is the simplest supervised learning algorithm which is used for both classification and regression problems. It is a supervised learning algorithm which means the data should contain both input and output parameters based on which the model will be trained. The K-NN algorithm for a given value of k will find the k nearest data points. Then the class will be assigned to the data point based on the class of the largest group of data points having the same class. For the detection of the K nearest neighbour, it uses similarity metric or Euclidean distance. The formula for Euclidean distance is

$$d(x,y) = \sum_{j=1}^{k} \sqrt{(x_j - y_j)^2} \qquad (4)$$

After this, the class which have the highest probability is assigned to that data point. The probability can be represented as

$$P(y = j|X = x) = \frac{1}{k} \sum_{y \in A} I(y^i = j) \qquad (5)$$

For the regression problems, the methodology is the same but instead of classes of neighbours, it uses the target values. One of the biggest problem in KNN is choosing a suitable k. If k is smaller than the decision boundary will be more irregular and on the other hand, the higher value of k will result in smoother decision boundary[8].

Naïve Bayes

The main base of the Naïve Bayes algorithm is the Bayes probability theorem[9]. The major advantage of naïve Bayes over other is that it is less complex and requires less RAM and CPU as compared to others. Let A and B are some events, P(A) and P(B) are their prior probabilities and p(B/A) is the probability of event B when event A has already occurred. So, for calculating posterior probability Bayes theorem can be represented as[10]

$$p\left(\frac{A}{B}\right) = \frac{P(\frac{B}{A})P(A)}{p(B)} \qquad (1)$$

Logistic Regression

Logistic regression is used in the classification problems. The logistic regression name come from the function it uses which is logistic function or sigmoid function. The sigmoid function takes an input value and maps it into a value that ranges from 0 to 1. The sigmoid function is represented as follows

$$\text{Sig(value)} = \frac{1}{1+e^{-value}} \qquad (2)$$

In logistic regression, the data items are classified based on whether the predicted probability is above a threshold value or not. In logistic regression, the decision boundary can be both linear or nonlinear, which depends on the distribution of the data points[11]. The logistic regression can be represented mathematically as follows

$$\ln \frac{p(y=1)}{1-p(y=1)} = w_0 + w_1 x_1 + \ldots + w_n x_n \qquad (3)$$

Decision Tree

Decision tree algorithm comes under supervised machine learning algorithms which are used for both classification and regression problems. A decision tree is a supervised machine learning algorithm so that means the data set should be labelled. Indecision tree algorithm the classification is done based on a set of rules. In a decision tree, a node will represent a feature, the branch will represent a rule and a leaf node will represent the outcome. It can be represented in a tree-like structure which provides higher stability and accuracy.
Indecision tree Algorithm following steps is taken in the first step a tree will be constructed which will have input features as its nodes. In the next step, it will select a feature from the input features for predicting the output which gives the highest information gain. Now use the above steps for the creation of subtrees by making use of the features which are not used earlier[2].

Random Forest

Random forest is an ensemble machine learning model. It is used for both classification and regression problems. It is an ensemble model which means it uses many machine learning algorithms to increase its performance as compared to other machine learning algorithms. Random forest randomly picks up a subset from the training data set and generate various

decision trees. It will predict the class of test class objects using the decision trees[12].

Support Vector Machine

There is another supervised machine learning algorithm support vector machine(SVM). In both classification and regression problems, this algorithm can be used. In support vector machine the data points are grouped and represented in space. In SVM p-1 planes are used to separate the data set which is considered as a collection of p vectors. These planes are known as hyperplanes. The data points should not fall very near to the hyperplane. It will detect the best hyperplane among all based on the gap between the classes. The maximum margin hyperplane can be defined as the hyperplane which creates the maximum gap between the classes. If there are n data points they can be represented as

$(\overrightarrow{x_1}. y_1) \dots \dots \dots (\overrightarrow{x_n}. y_n)$

Where x is the real vector and y is used to represent the class (0 or 1)

The maximum margin hyperplane can be defined as $\vec{w}.\vec{x} - b$

Where $\vec{w}$ is the normal vector and $\frac{b}{[\![\vec{w}]\!]}$ is the offset of hyperplane along $\vec{w}$ [13].

IV. DESCRIPTION ABOUT THE DATA SET

The features, data types and its statistics are represented in the tables I and II. Based on the data set attributes, the classifier will predict whether a patient is diabetic or not. In this dataset, all individuals are females who are at least 21 years old. This problem is a two-class classification problem in which classifier has to classify whether a patient is diabetic or not. In this dataset, the class value 1 and 0 will be interpreted as "diabetic" or "non-diabetic". The dataset contains 768 records in which 500 are diabetic and other 268 are non-diabetic. By generating the profile report of the dataset it is noticed that there are no null values present but there are some unusual values present in it such as 0 value in blood pressure etc. This dataset contains total 9 attributes out of which 8 are independent and 1 is dependent. Table 1 and 2 show the descriptions of the attributes

Table 1. Attribute Description

s.no	Attribute	Type	Data type	Zeroes	missing
1	Pregnancy	Discrete	int	111	0
2	Glucose Concentration	Discrete	int	5	0
3	Blood Pressure(mm Hg)	Discrete	int	122	0
4	Skin Thickness(mm)	Discrete	int	227	0
5	Insulin(U/ml)	Discrete	int	374	0
6	BMI (kg/m)	Continuous	int	11	0
7	Diabetes Pedigree Function	Continuous	int	0	0
8	Age	Continuous	int	0	0
9	Outcome Class	Continuous	bool	500	0

Table 2. Dataset statistics

Data-set	PIDD
Number of Variables	9
Number of observations	768
Missing Cells	0
Missing Cells(%)	0.0%
Duplicate Rows	0
Duplicate Rows(%)	0.0%
Total Size in Memory	54.1KIB
Average record size in memory	72.2B

V. RESULTS AND DISCUSSIONS

The detection of diabetes in early phases can increase one's life. Several models have been developed using supervised machine learning algorithms These models are developed Using python. The dataset is split into two parts i.e. training and testing. We have used 70% of the data for training the model and the rest 30% for testing it. To classify the patients into diabetic or non-diabetic five different supervised machine learning algorithms are used namely K-nearest neighbour, DT, NB, SVM, logistic regression and random forest. All models are compared based on 5 performance metrics which are accuracy, precision, recall, f1-score and Auc (Area under the curve). Accuracy represents how many times our model is correctly predicting whether a patient is diabetic or non-diabetic Recall is the proportion of the diabetic patients that are correctly identified by a model. Specificity is the proportion of correctly identifying the non-diabetic patients by a model. The harmonic mean of a model's recall and precision is known as f1-score. ROC (Receiver operating characteristic) AUC (Area under the curve) curve will tell how efficient a model is in distinguishing between the classes. After applying various machine learning algorithms on the PIMA diabetes dataset we get the following results as shown in table 3

Table 3. Comparison of classification models

s.no	Classifier	Accuracy (%)	precision	Recall	f1-Score	ROC-AUC Score
1	KNN	75	68	70	69	74.06
2	DT	82.7	80	76	78	81.6
3	NB	75	69	67	68	73.6
4	SVM	74	66	72	69	73.7
5	LR	76	68	74	71	75.5
6	RF	84	83	76	80	83

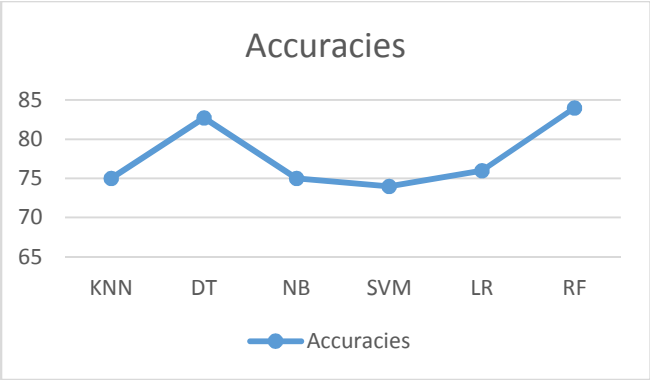

Fig 2. Comparison of accuracies

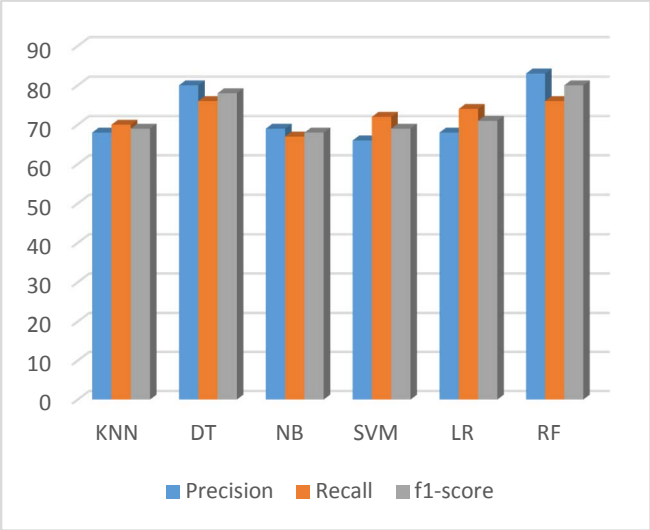

Fig 3 Comparison of precision, recall and f1-score

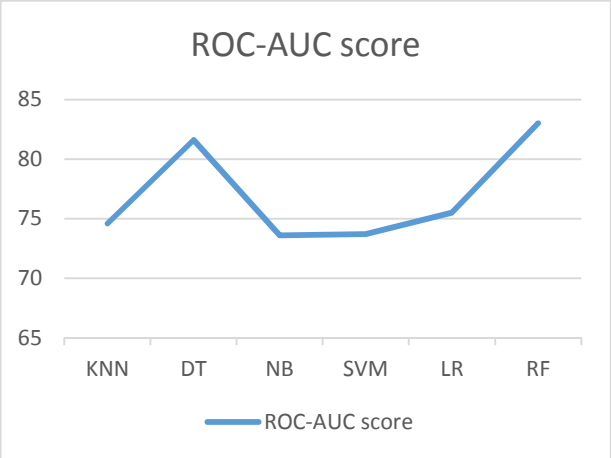

FIG 4. ROC-AUC SCORES

Table 3 shows different performance metrics for all the models. It is seen that Random forest is giving the highest accuracy of 84% then Decision tree with 82.5% accuracy. SVM is giving the least accuracy of 74%. So the accuracy of Random forest and decision tree is greater as compared to others. Other than accuracy f1-score will help to decide the best performing model among all. The model with greater f1 score will perform better. Fig 3 and Fig 4 shows the comparison of models based on precision, recall, f1-score and ROC-AUC curve. Random forest performs better among all other algorithms with f1-score of 80 and ROC-AUC value 83.

VI. CONCLUSION AND FUTURE SCOPE

The detection and prediction of diabetes are one of the most common and important medical problems in today's world. If not diagnosed in the early phase it can lead to other health problems. In this paper, an organized experiment is done by using six different machine learning algorithms which are KNN, Naïve Bayes, Support vector machine, Decision tree, Random forest and logistic regression. These are compared based on five performance metrics accuracy, recall. precision, f1-score and ROC-AUC curve. Our results tell us that Random forest performs better than other with 84% accuracy, precision 83, recall 76, f1-score 86 and ROC-AUC score 83. Further, this work can be improved using other ensemble machine learning methods.

VII. REFERENCES

[1] A. Mujumdar and V. Vaidehi, "ScienceDirect ScienceDirect ScienceDirect ScienceDirect Diabetes Prediction using Machine Learning Aishwarya Mujumdar Diabetes Prediction using Machine Learning Aishwarya Mujumdar Aishwarya," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019.

[2] A. Z. Woldaregay *et al.*, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artif. Intell. Med.*, vol. 98, no. April 2018, pp. 109–134, 2019.

[3] D. R. Nair *et al.*, "Trend in the clinical profile of type 2 diabetes in India - Study from a diabetes care centre in South India," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 6, pp. 1851–1857, 2020.

[4] M. M. Islam, M. J. Rahman, D. Chandra Roy, and M. Maniruzzaman, "Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 3, pp. 217–219, 2020.

[5] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017.

[6] O. Ben-Assuli, T. Heart, N. Shlomo, and R. Klempfner, "Bringing big data analytics closer to practice: A methodological explanation and demonstration of classification algorithms," *Heal. Policy Technol.*, vol. 8, no. 1, pp. 7–13, 2019.

[7] D. Jashwanth Reddy *et al.*, "Predictive machine learning model for early detection and analysis of diabetes," *Mater. Today Proc.*, no. xxxx, 2020.

[8] J. Huang, Y. Wei, J. Yi, and M. Liu, "An improved knn based on class contribution and feature weighting," *Proc. - 10th Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2018*, vol. 2018-Janua, pp. 313–316, 2018.

[9] H. Zhang, L. Jiang, and L. Yu, "Attribute and instance weighted naive Bayes," *Pattern Recognit.*, vol. 111, 2021.

[10] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 706–716, 2020.

[11] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," *Proc. IEEE 7th Int. Conf.*

Comput. Sci. Netw. Technol. ICCSNT 2019, pp. 135–139, 2019.

[12] S. Liu, H. Li, Y. Zhang, B. Zou, and J. Zhao, "Random forest-based track initiation method," *J. Eng.*, vol. 2019, no. 19, pp. 6175–6179, 2019.

[13] P. Naraei, A. Abhari, and A. Sadeghian, "Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data," *FTC 2016 - Proc. Futur. Technol. Conf.*, no. December, pp. 848–852, 2017.