TEST Engineering & Management

January – February 2020
ISSN: 0193-4120 Page No. 15183 - 15187

Optimal Diabetes Predictive analysis using different machine learning models

Gowri S, Srija U, Jabez J, Vimali J S, Senduru Srinivasulu
Sathyabama Institute of Science and Technology

Article Info
Volume 82
Page Number: 15183 - 15187
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 28 February 2020*

Abstract:
Disease prediction is one of the applications where a machine learning model provides excellent results. Diabetes disease has become one of the most common diseases which have more ill consequences. If the disease is predicted at the first stage, then it becomes quite easy to control it. Otherwise, it may lead to problems like cardiovascular disease, Eye damage, Kidney damage, etc. Using machine learning model diabetes can be predicted beforehand and also it let us know the feature that affects diabetes. This paper identifies the model that is efficient to do this application. Advantage of this application is accuracy is higher than the pre-existing

Keywords :Machine learning, Artificial Intelligence, Predictive models.

I. INTRODUCTION

Machine learning is defined as an application of artificial intelligence (AI) which provides the systems to automatically learn from experience without being explicitly programmed. Diabetes is a disease which occurs when your blood sugar is high. Blood glucose is the source of energy comes from the food we consume. Insulin is thee hormone made by the pancreas, which helps glucose from food get into the cell of our body to be used for energy. For diabetic patient, their body doesn't make enough insulin so that glucose from food will stay in blood and doesn't reach cells.

Diabetes currently affects more than 62 million peoples in India, which is more than 7.1% of the population. The average age on onset is 42.5 years. Nearly 98 million people in India may have type2 diabetes by 2030. There are diabetes warning signs and symptoms like excessive thirst and hunger, frequent urination, weight loss or gain, Blurred vision.

Machine learning model is a mathematical representation to predict the result for real world problem [10]. The model learns the patterns from training data. Output of the training process the model is used for the prediction of new input.

Supervised learning is one of the types of machine

learning in which the model learns from labeled data. Classification is one the categories of supervised learning which can be defined as a technique which makes model to learn from the data input and then uses this learning to classify new observation. In this paper, the classification technique like Decision tree, Random forest, KNN, Gradient boosting is used.

Unsupervised learning is also another type of machine learning in which the model learns from unlabeled data. Clustering is one of the categories of unsupervised learning which can be defined as grouping a set of objects in such a way that objects in same group are similar to each other.

II. LITERATURE SURVEY

Diabetes prediction is an enormous challenge for the health systems. The researchers are ongoing with their work continuously to predict the diabetes. [1]AsmaShaheen Khan, Waqas Ahmed proposed a decision support system in diabetic care. The proposed system stores the patients' information and predicts the result. It also gives them the optimal advices according to their condition. [2]Sushant Ramesh, H. Balaji, N.Ch .S.N Iyengar and Ronnie D. Caytiles proposed the optimal predictive analytics of Pima diabetics using deep learning. [3]Tawfik Saeed Zeki, Mohammad V. Malakooti, Yousef Ataeipoor, and TalayehTabibi proposed a system for diabetes. After Data is collected they designed a rule-based expert system and this is tested in ShahidHasheminezhad Teaching Hospital affiliated to Tehran University of Medical Sciences. [4]Rahman Ali, Jamil Hussain, Muhammad Hameed Siddiqi, Maqbool Hussain and Sungyoung Lee proposed a model for prediction and management of Diabetes mellitus. When a new incoming data is evaluated and compared against the knowledge and to classify the type of diabetes of the patient. [5]T.P. Kamble and Dr. S.T. Patil proposed a system where deep learning based approach which is used to detect whether patient is diabetic or not as this model is popular for classification and recognition purpose. This model is also used to detect the patient belongs to Type 1 or Type 2.[6] Margret Anouncia S., Clara Madonna L. J., Jeevitha P., Nandhini R. T proposed a design for a Diabetic Diagnosis system where the experts created a knowledge base from the existing data set using upper and lower approximation when new data is evaluated the it compute its own class.[7] RahmatZolfaghari proposed a three-layer hierarchy multi- classifier. And the min-max normalization is

TEST Engineering & Management

January – February 2020
ISSN: 0193-4120 Page No. 12873 - 12879

applied on data to avoid the difficulties in calculation. [8] K. Sridar, Dr. D. Shanthi used a back propagation and Apriori algorithm for diabetes detection. First, the input is taken from the user and the input is analyzed and used to detect the diabetes. [9] Gaganjot Kaur, Amit Chhabra proposed classification which uses Decision tree algorithm to predict the given new patient is diabetic or not. [10] Ranveet Jyot Singh, Williamjeet Singh proposed a system uses Decision tree, Association rule, Kdd, KNN, SVM to classify the given new patient is diabetic or not.

III. METHODOLOGY

A. Data collection

This involves the collection of medical information from various sources like hospitals, discharge slips of the patient and from kaggle.com..

B. Data Preprocessing

For the collected data set the Preprocessing techniques will be applied so that making itself ready to fed into the model which will be used to predict the patient is diabetic or not. This preprocessing will remove all the unnecessary data and extract the important features from the data. And it also finds the average of the particular attribute so that any loss data is replaced with average. This can achieved by using the inbuilt function called Imputer in python.

C. Cross Validation

For the preprocessed data, cross validation is applied such that data set is divided into two halves. One part is said to be Training data set and other one is said to be Testing data set. This can be done using the predefined function called train_test_split in python and sample in R programming.

D. Model Making and Testing

The model keeps learning from the training data set. And later on, the model is tested on the testing data to verify whether the mode is predicting the right outcome or not.

E. Analysis on real life scenario

The model which is learned as well as tested on the data set is ready to make a decision for the new input which is not present in the data set. This is the part where ML plays a crucial role.

In this stage the function "predict" is used to predict the output for new real time data.

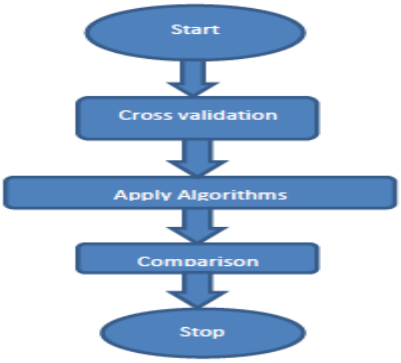

Fig.1. Workflow of optimal diabetes prediction

F. Dataset

The dataset is available on both kaggle.com as well as github.com. The dataset can be downloaded from below link https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/diabetes.csv

The attributes of the dataset are Glucose, BMI, Age, Pregnancies, Diabetes Pedigree Function, Blood Pressure, Skin Thickness and Insulin.

For the Random forest, Decision Tree, Gradient boosting model the plotting is done against the important feature to predict the outcome which looks a like below:

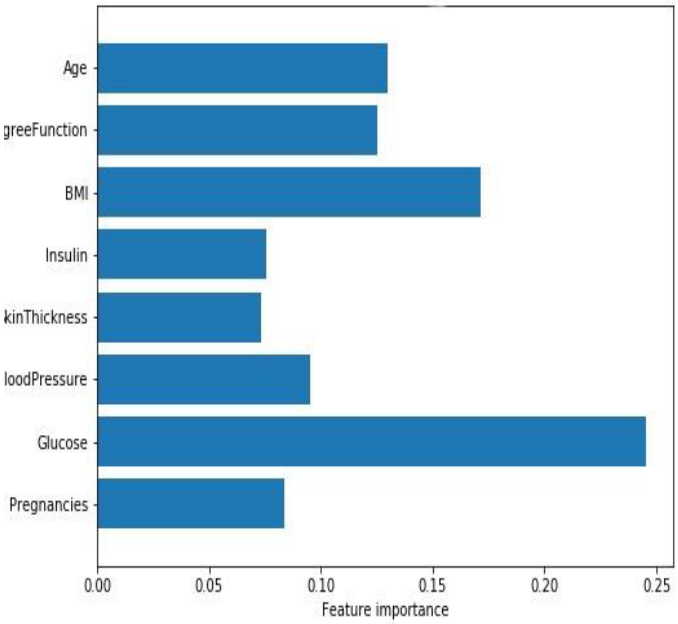

Fig.2. Feature importance for prediction

IV. TYPES OF MODELS

This paper is mainly done for the comparison of various models and declares the most effective one among them. Here, we used 6 models where 5 supervised learning models and only one is unsupervised

A. .KNN [K- Nearest Neighbor]

KNN is a non-parametric, lazy learning algorithm. Here, data points are separated into classes. And for the new input the data point is checked for the nearer ones.

The KNN's steps are:
1—Receive an unclassified data;
2—Measure the distance using Euclidean formulae 3—Gets the K (K is a parameter that you define) smaller distances;
4—Check the list of classes had the shortest distance and count the amount of each class 5—Takes as correct class the class that appeared the most times;
6—Classifies the new data with the class that you took in step 5;

Euclidean distance can be found by followed formulae:

TEST Engineering & Management

January – February 2020
ISSN: 0193-4120 Page No. 15183 - 15187

$$E(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

Where x, y are the coordinates of the data points Euclidean distance will be the square root of summation of data point's square The function which is used for the KNN classifier in Python is KNeighborsClassifier

B. Decision Tree

Decision Tree is the supervised learning model where a tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning.

The Decision Tree's steps are:
1—Place the best attribute of the dataset at the root of the tree.
2—Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3—Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.
To find Entropy of an observation the formulae can be used is as below:

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Where pi is the probability of class label To find an information gain we can use the given formulae:

$$Gain(A) = Info(D) - Info_A(D)$$

Where Info (D) is entropy of an observation InfoA(D) is entropy of the respective attribute
The function which is used to build a decision tree model is DecisionTreeClassifier and in R programming ctree will be used.

C. Random forest

A random forest consists of multiple random decision trees. Two types of randomness's are built into the trees.

1—Randomly select "k" features from total "m" features, Where k << m
2—Among the "k" features, calculate the node "d" using the best split point.
3—Split the node into daughter nodes using the best split.
4—Repeat 1 to 3 steps until "l" number of nodes has been reached.
5—Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.
At each node, the decision tree searches through the features for the value to split on that results in the greatest reduction in Gini impurity.
Where the pi is the probability of the class labels.

$$I_G(n) = 1 - \sum_{i=1}^{J}(p_i)^2$$

In this Model the predefined function Random Forest Classifier is used in Python and in case of R programming language random forest function be used.

D. Gradient Boosting

The gradient boosting algorithm (gbm) can be most easily explained by first introducing the AdaBoost Algorithm. The AdaBoost Algorithm begins by training a decision tree in which each observation is assigned an equal weight.

1—Train a decision tree
2—Calculate the weighted error rate (e) of the decision tree.
3—Calculate this decision tree's weight in the ensemble the weight of this tree = learning rate * log((1—e) / e)
4—Update weights of wrongly classified points
5—Repeat Step 1(until the number of trees we set to train is reached)
6— Make the final prediction The Mean Squared Error(MSE) as loss defined as

$$Loss = MSE = \sum (y_i - y_i^p)^2$$

Where yi is the ith target value Yi p is the ith prediction This is the most efficient model which had been used; the function for this is GradientBoostingClassifier in Python.

E. Logistic Regression

Logistic Regression is used when the dependent variable (target) is categorical

1—get a dataset
2—train a classifier
3—make a prediction using such classifier The logistic regression is a S-shaped curve that can take any real-valued number and map it into a value between 0 and 1. The formulae for the logistic regression will be as follows:

$$1 / (1 + e^{-value})$$

Where e denotes the exponential factor Logistic Regression is the supervised learning model where the inbuilt function called LogisticRegression in python and in the case of R programming glm is used.

F. KMeans clustering

K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroid

1—Initialization
2—Cluster Assignment
3—Move the Centroid

TEST Engineering & Management

January – February 2020
ISSN: 0193-4120 Page No. 12873 - 12879

This algorithm has an objective function to know a squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

Where c is the number of cluster centers ci is the number of data points in ith cluster ||xi-vj|| is the Euclidean distance
These are the 6 models which are used to predict the patient is suffering from diabetes or not. From importing the dataset into spyder notebook, the first step is to split the data into training and testing data in consideration of the supervised learning. But in the case of unsupervised learning the data will not have the class label part. On using the function for the respective model the outcome is the machine learning model which is used further to predict the output for new result. And the pictorial representation is just only to understand the dataset in better way.
The Machine Learning techniques are used to predict the output for the new real time data. This is mainly used to train the model with the input and make the model to take the input and output and give the output as the program for the respective input and output.

V. RESULTS

After analyzing the dataset, the different models are created and the accuracy are compared and tabulated below. Taking these into consideration we can further do model effectively as well as efficiently which means less time and space required as well as the predicted output is well effectiveness

Table- II:Predicted output of different Model

Sl.no	Model	Training accuracy	Testing accuracy
1	KNN	0.79	0.77
2	Logistic Regression	0.78	0.77
3	Decision tree	0.86	0.84
4	Random Forest	0.92	0.93
5	Gradient Boosting	0.94	0.95

From the above table, we can clearly say that the Gradient Boosting is the one which is most accuracy in training as well as testing dataset. Basically, the boosting technique is differing from bagging technique by reducing bias and variance and sequential classifier. By comparing the bagging (e.g. Random Forest) and boosting (e.g. Gradient Boosting) bagging handles over fitting but Boosting itself can over fit.

The k-means clustering has the result of the two clusters which one of the cluster defines the person is suffering from diabetic and the other one said to be the person who is not suffering from diabetic. The diagrammatical representation of the clusters will be the as follows:

.

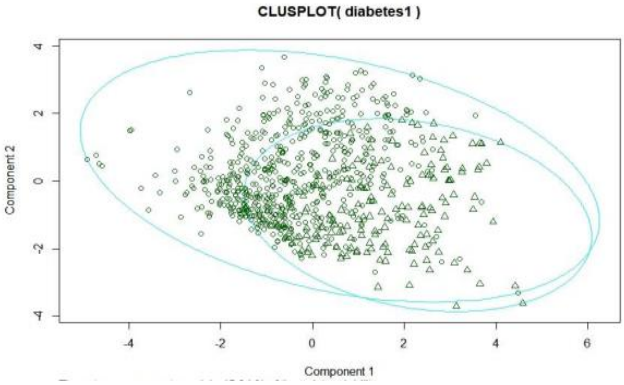

Fig. 1. These two components explain 45.84% of the point variability

The Prediction of the diabetic can be done through the decision tree model where the labels are represented as the form of trees and the leaf node will have the respective class label. According to the information gain of the attribute the attributes are divided to form a tree.

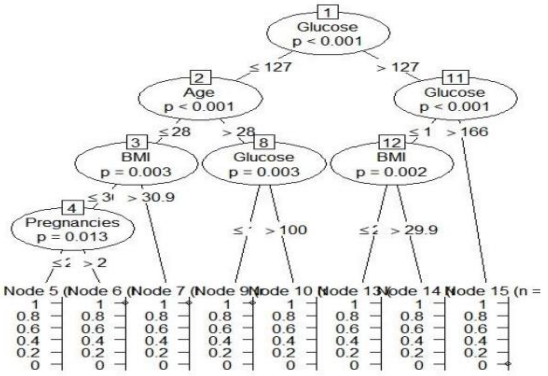

Fig. 2. The pictorial representation of the tree

The KNN training and testing accuracy is compared by changing the n_neighbors.

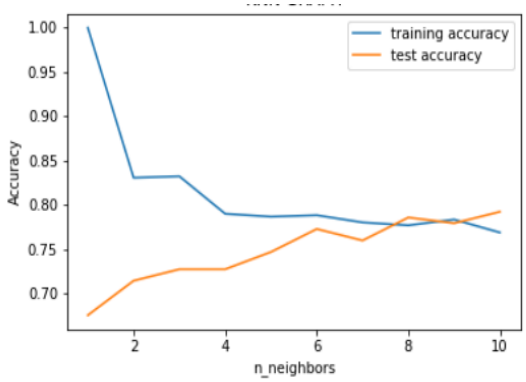

Fig. 3. KNN Graph

The Random Forest is the collection of trees where the error rate will be reduced as the number of trees gets increased. The mechanism for that will be represented in below graph.

January – February 2020
ISSN: 0193-4120 Page No. 15183 - 15187

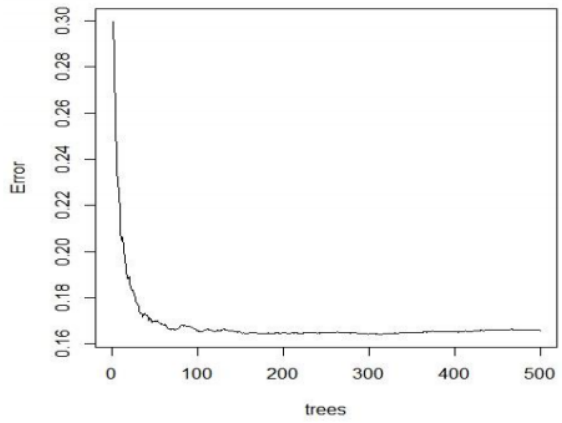

Fig. 4. Random Forest with reduced error rate

REFERENCES

1. S. Gowri, G. S. Anandha Mala (2014), "Strategic Enhancement of Collaborative Framework for Novelty in Retrieval from Digital Textual Data Deploying DPSC and RBWM Algorithms for Forensic Analysis", Journal of Engineering Research, Vol. 3 / No. 4 / 2015, ISSN: 1024-8684.
2. Sivasangari. A and Martin Leo Manickam. J (2014) " A Light Weight Cryptography Analysis for Wireless Based Healthcare Applications " , Journal of Computer Science, , ISSN : 1549-3636 , Vol.10,No. 5, pp.2088-2094.
3. P. Ajitha, Dr. G. Gunasekaran., Semantic Based Fuzzy Inference System(SBFIS) Prediction of Patient Emotion and Prescription using support vector machine" in the Journal of Medical Imaging and Health Informatics, PP 769-773 ISSN: 2156-7018, Vol.6,No.3,June2016.
4. Jabez, J., Gowri, S., Vigneshwari, S., Albert Mayan, J., Srinivasulu, S."Anomaly detection by using CFS subset and neural network with WEKA tools ",Smart Innovation, Systems and Technologies,ISBN : 978-981-13-1746-0/2018/pp. pp 675-682.
5. Bharathi, M., Nirmalrani, V. "Foetuscare: An android app for pregnancy care", Advances in Intelligent Systems and Computing, ISSN: 156-161, Vol.6,No.32016.
6. Vimali, J.S., Gupta, S., Srivastava, P.,"A novel approach for mining temporal pattern database using greedy algorithm "in International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 20171549-3636 , Vol.9,No. 5, pp1088-1094.
7. A.Sivasangari,Suvam Bhowal,R.Subhashini,"Secure Encryption in Wireless Body Sensor Networks", Volume 3, : Emerging Technologies in Data Mining and Information Security, pp.679-686.W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
8. Sushant Ramesh, H. Balaji, N.Ch .S.N Iyengar and Ronnie D. Caytiles https://www.researchgate.net/publication/320468806_Optimal_Predictive_analytics_of_Pima _Diabetics_using_Deep_Learning.
9. Rahman Ali, Jamil Hussain, Muhammad Hameed Siddiqi, Maqbool Hussain and Sungyoung Lee: https://www.sciencedirect.com /science/article/pii/S2352914817301405.
10. Palagati Harish, Dr.R.Subhashini and K.Priya, "Intruder Detection by Extracting Semantic Content from Surveillance Videos" IEEE International Conference on Green Computing, Communication and Electrical Engineering, ICGCCEE 2014.