An Analysis On Comparitive Study Of Diabetic Prediction Using Machine Learning

[1]Abhinav Sengar
[1]Student
[1]Jiwaji University, Computer and Science somas

Abstract - **Nowadays, diabetes has become a typical disease to the mankind from young to the old persons. The expansion of the diabetic patients is increasing day-by-day because of various causes like toxic or chemical contents mix with the food, obesity, bad diet, change in lifestyles, eating habit etc. Therefore, diagnosing the diabetes is extremely essential to save human lives. The information analytics can be defined as a process of examining and identifying the hidden patterns from great amount of information to draw conclusions. In health care, this analytical process is carried out using machine learning algorithms which analyzes medical data in order to use the machine learning models for medical diagnoses. In this paper, we have evaluated comparative study of different machine learning classifications algorithms for predicting diabetes more accurately. In this research work, we are analyzing the accuracy of different algorithms like: Linear Discriminant Analysis, Random Forest Classifier, Decision Tree Classifier, MLP Classifier in order to identify best classifier.**

keywords - **Diabetes Prediction, Machine Learning Techniques: Data pre-processing, Linear Discriminant Analysis, Random Forest Classifier, Decision Tree Classifier, MLP Classifier.**

I. Introduction

Diabetes is a sickness that happens when your blood glucose, additionally called blood sugar, is excessively high. Blood glucose is your essential wellspring of imperativeness and begins from the sustenance you eat. Glucose into your cells is used for imperativeness which is empowered by Insulin. A portion of the time your body doesn't make enough—or any—insulin or doesn't use insulin well. Glucose at that point remains in your blood and doesn't arrive at your cells.

After some time, having a lot of glucose in your blood can mess wellbeing up. Notwithstanding the way that diabetes has no fix, you can figure out how to manage your diabetes and stay sound.

Here and there individuals call diabetes "a dash of sugar" or "marginal diabetes." These terms propose that somebody doesn't generally have diabetes or has a less genuine case, yet every instance of diabetes is not a ignorable case.

In this paper work, we compare different ML classifier algorithm like: Linear Discriminant Analysis, Random Forest Classifier, Decision Tree Classifier, and MLP Classifier in order to identify best result.

II. Literature Review

Diabetes is a proper ailment for information mining innovation because of various reasons. In each age period this infection is normal. It charges a lot of financially and its impact is developing rapidly. The body of a diabetic individual doesn't deliver or proficiently use insulin, the hormone that "opens" the cells of the body, permitting glucose to show up and fuel them. A diabetic individual has danger of having different sicknesses as vein hurt, visual impairment, coronary illness, nerve harm and kidney infection [14]. Diabetes is by and large of 2 sorts: type 1(insulin ward diabetes) and type 2(non-insulin-subordinate diabetes). Diabetes is an infection wherein the blood glucose levels get increment which is because of the imperfections in discharge of insulin, or its activity, or both. Diabetes is a delayed clinical sickness. In diabetes, the cells of an individual produce inadequate measure of insulin or damaged insulin or may unfit to utilize insulin appropriately and productively that further prompts hyperglycemia and type-2 diabetes. In type 1 diabetes there is total absence of insulin, typically optional to a ruinous procedure troubling the insulin-delivering beta cells in the pancreas. There is overabundance decay of beta cells that improves procedure of raised blood sugars in type 2 diabetes. In current time it is one of the significant general medical issues. The International Diabetes Federation has asserted that by and by 246 million individuals are experiencing diabetes worldwide and this number is relied upon to increment up to 380 million by 2025 [16].

So cure of diabetic will be more accurate and efficient when we know the disease more accurately at the early stage of it. For that we are using following dataset for comparison:

Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age of diabetic patient.

III. Data Collection, Techniques and Features

Data collection

We have used dataset for our this paper from [4]. Total 800 no. of instances we have in our dataset [4]. We have divided the dataset in 40:60 ratios for training and testing purpose. The input to these algorithms is previous diabetic patient data which include the Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age of diabetic patient and output. The output feature having values 1 or 0 representing diabetic and non-diabetic respectively.

Features

Following features are used in our dataset for diabetic prediction:
(1) Pregnancies,
(2) Glucose,
(3) Blood Pressure,
(4) Skin Thickness,
(5) Insulin,
(6) BMI,
(7) Diabetes Pedigrees Function and
(8) Age

IV. TECHNIQUES USED FOR COMPARING, TRAINING AND TESTING DATASET

We are comparing four machine learning algorithms in our dissertation:

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a dimensionality decrease procedure. As the name suggests dimensionality decrease methods techniques the quantity of measurements (for example factors) in a dataset while holding however much data as could reasonably be expected.

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[10][11] Random decision forests correct for decision trees' habit of overfitting to their training set.[9]

Decision Tree Classifier

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

MLP Classifier

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to refer to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold activation); see § Terminology. Multilayer perceptrons are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.[12]

V. FINDING AND RESULTS

Here we compare four machine learning classification modals. In this section, we present a thorough evaluation of these models trained with the diabetic data from [4]. We divided the dataset in the ratio of 40:60 in which the first set of dataset shows training dataset and rest is for testing purpose.

Features used in machine learning models.

In the above figure you can see the following features we are going to use for training and testing purpose.
On the basis of these diabetic features, model predicts the result that patient is diabetic or not. Prediction results are represented in the form of 1 and 0, where 1 represent diabetic and 0 represents non-diabetic.

Features:

- Pregnancies,
- Glucose,
- Blood Pressure,
- Skin Thickness,
- Insulin,
- BMI,
- Diabetes Pedigrees Function and
- Age

Machine learning models used for comparisons.

Here we are using above listed features in following Machine learning models.
Models: Linear Discriminant Analysis, Random Forest Classifier, Decision Tree Classifier, MLP Classifier.

Different Measures for identifying accuracy of Machine learning modals

1. *Training data Accuracy*

For training all the four models, We have used 40 % of our dataset and deduced the following result for all the machine learning modals.

Table 5.1 – Train data accuracy for all models.

Model	Train data Accuracy (%)
Linear Discriminant Analysis	0.7717
Random Forest Classifier	0.9935
Decision Tree Classifier	0.7609
MLP Classifier	0.6457

Here, we can identifies that Random Forest gives the best accuracy among all.

2. *Test data Accuracy*

For testing all the four models we used 60 % of our dataset. And gets the following result for all the machine learning modals.

Table 5.2 – Test data accuracy for all models.

Model	Test data Accuracy(%)
Linear Discriminant Analysis	0.7825
Random Forest Classifier	0.7662
Decision Tree Classifier	0.7013
MLP Classifier	0.6623

Here, we can identify that Random Forest gives the best accuracy among all.

3. *Precision*

Precision (also referred to as positive prognosticative value) is the fraction of relevant objects among the retrieved objects. Here is the precision result for all the models.

Table 5.3 – Precision result for all models.

Model	Precision
Linear Discriminant Analysis	0.7209
Random Forest Classifier	0.7037
Decision Tree Classifier	0.5528
MLP Classifier	1.0000

4. *Recall*

Recall (also called sensitivity) is that the fraction of the full quantity of relevant instances that were truly retrieved. Here is the recall result for all the models.

Table 5.3 – Recall result for all models.

Model	Recall
Linear Discriminant Analysis	0.5904
Random Forest Classifier	0.5428
Decision Tree Classifier	0.6476
MLP Classifier	0.0095

Graphical representation of comparison of all algorithms

By execute the above classification algorithms; the Random Forest Classifier model gives highest accuracy as comparative to other models as shown in below table and figure.

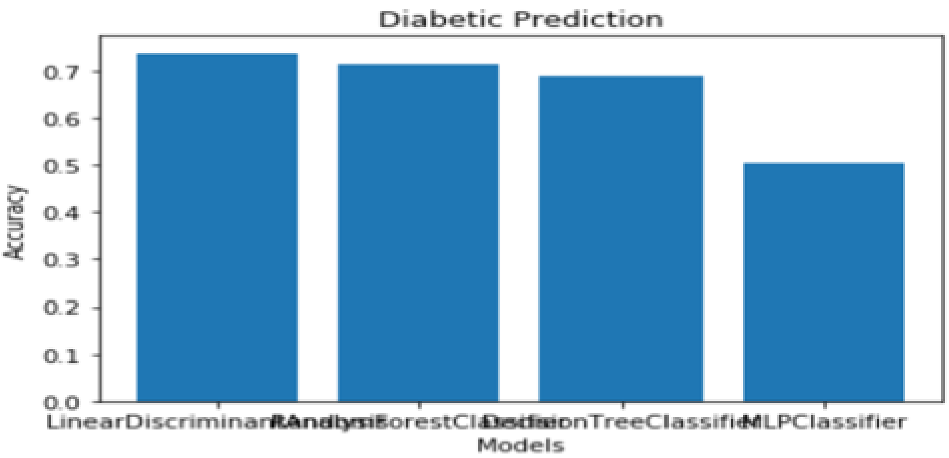

Figure- 5.1 Graphical representation of comparison of all models.

VI. CONCLUSION

AI and Machine learning has the extraordinary capacity to upset the diabetes hazard expectation with the assistance of cutting edge computational techniques and accessibility of enormous measure of epidemiological and hereditary diabetes dataset.
Identification of Diabetes in its beginning times is the key for treatment.

This research work has compared existing machine learning algorithms for diabetic prediction. These machine learning techniques can also facilitate researchers to provide solutions and assists in order to make better decisions for diabetic prediction. Our evaluation result shows that these machine learning models helps in diabetes prediction.

The results of our research work identified that Random forest classifier provides relatively better accuracy in comparison of other models (Decision Tree, Linear discriminant analysis, MLP classifier) for diabetic prediction.in the training dataset. This data will improve the performance of our prediction models.

REFERENCES

[1] G. Kaur, “Improved J48 Classification Algorithm for the Prediction of Diabetes.”, 2014

[2] A. Mortona, “An analysis of supervised Machine Learning Techniques for Predicting short-run In-Hospital Length of keep among Diabetic Patients.”, 2010

[3] Iqbal H. Sarker, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus.” 2019

[4] V. Karthikeyni, “Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction of Diabetes Disease.” 2013

[5] K. Sojanya, “MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices”, 2013

[6] M. Fatima, “Survey of Machine Learning Algorithms for Disease Diagnostic”, 2017

[7] Huang, Feixiang; Wang, Shengyong; Chan, ChienChung, "Predicting disease by using data mining based on healthcare information system," Granular Computing (GrC), 2012 IEEE International Conference on , vol., no., pp.191,194, 11-13 Aug. 2012

[8] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology And Secured Transactions, 2012 International Conferece For, pp. 471-472. IEEE, 2012.

[8] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. “The Elements of Statistical Learning (2nd Ed.).” 2008

[9] Ho, Tin Kam, “Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal”, 2016.

[10] Ho TK, "The Random Subspace Method for Constructing Decision Forests" (PDF), 1998

[11] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction.” 2009.