Volume: 14 Issue: 01 Pages: 5294-5299(2022) ISSN: 0975-0290

On The Analysis of Some Machine Learning Algorithms for the Prediction of Diabetes

Bello A. Bodinga
Department of Computer Science,Usmanu Danfodiyo University, Sokoto, NIGERIA
Email: bello.bodinga@udusok.edu.ng
Mukhtar A. Abdulsalam
Undergraduate Student, Department of Computer Science,Usmanu Danfodiyo University, Sokoto, NIGERIA
Bello A. Buhari
Department of Computer Science,Usmanu Danfodiyo University, Sokoto, NIGERIA
Muzzammil Mansur
Department of Computer Science, Waziri Ummaru Federal Polytechnic. Birnin-Kebbi, NIGERIA

----------ABSTRACT----------

Diabetes or Diabetes Mellitus (DM) is noxious diseases in the world. Diabetes is caused by obesity or high blood glucose level, lack of exercise and so forth. It can be manage if it's detected at early state. Machine learning is the construction of computer system or program that can adapt and learn from their experience. PIMA dataset is used in this research works. The dataset contains some 9 attributes of 768 patients. There are different kinds of machine learning algorithms but in this research works we choose three algorithms which are under supervised learning. The algorithms are Logistic regression, Decision tree and Random forest. Each of these algorithms model were trained and tested. We later use some measure to compare and analyze the performance of the machine learning algorithms. The performance measures used are Accuracy, F-measure, Recall and Precision. Logistic Regression has the highest accuracy score which is 77%, also have the highest precision score 0.77 and have the highest f-measure 0.64. Decision Tree has the highest recall score 0.58.

Keywords**: Diabetes, Machine learning, Logistic Regression, Decision tree, Random forest**

Date of Submission: Apr 19, 2022 Date of Acceptance: Jun 04, 2022

I. INTRODUCTION

Diabetes is deadly diseases. Diabetes is caused by obesity or high blood glucose level, living lifestyle, lack of exercise and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes.

Machine Learning is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning are capable of prediction, however it's tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

Machine learning involves computers discovering how they can perform tasks without explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create needed algorithm. In practice, it can turn out to be more effective to help the machine develop its own algorithms, rather than having human programmers specify every needed step.

As a scientific endeavor, machine learning grew out of the quest for artificial intelligence. In the early days of AI as an academic disciple, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods as well as what was then termed "neural network"; these were mostly perceptron and other models that were later found to reinvention of the generalized linear model of statistics. Probabilistic reasoning was also employed, especially in automated medical diagnosis.

II. STATEMENT OF PROBLEM

Diabetes is a problem killing many people throughout the world. With the advancement in technologies, human life is prospering. Therefore why not use the technologies for the betterment of healthy lifestyle. The deep learning

Volume: 14 Issue: 01 Pages: 5294-5299(2022) ISSN: 0975-0290

technologies and various machine learning algorithms are used for much type of prediction facilities. Often used by business giants for profits and sales. Here we are presented with the question of how we can use these technologies for the betterment of mankind. The various algorithms we have used and learned with time are to be challenged for prediction of something whose specialization only resides in the hands of experts. The machine has to be trained with the mind of doctors in order to learn the complexity of various features of bio mechanics of human beings and predict the complicated problems of living beings. These algorithms have to be implemented for the prediction of complicated diseases using various features and external factors provided from an authentic dataset.

III. REVIEW OF RELATED WORKS

This section will carefully review some of these existing research in line with this project.

Jitranjan Sahool et al. [3] predicting diabetes using Machine Learning Classification Algorithms and this research work shows that, Logistic regression was found outperform all the machine learning algorithm showing the maximum accuracy of 79.17% in comparison to other algorithm.

Nonso et al. [4] presented predicting diabetes novel approach onset: in the supervised learning approach, five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results presented are compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy.

Tejas et al. in a study [2] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression and ANN. Their work project proposes an effective technique for earlier detection of the diabetes disease.

In another study by Deeraj et al. [1] a new approach to diabetes disease prediction proposed using data mining. An Intelligent Diabetes Disease Prediction System is developed that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

Mitushi Soni et al. [5] presented diabetes prediction using different machine learning algorithms(Support vector machine, logistic regression, decision tree, K-Nearest Neighbor and random forest) make comparison between these algorithms.

We used different machine learning algorithms. Which are Logistic Regression, Random forest and Decision tree.

IV. DATASET DESCRIPTION

The data is obtained from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset contains some attributes of 768 patients.

Table 3.1: Dataset Description

S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

V. DATA PREPROCESSING

Data preprocessing is one of the most important process. Mostly, healthcare related data contains missing value and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done.

To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of Iterative Analysis. The set of steps is known as Data Preprocessing. It includes

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Data Preprocessing is necessary because of the presence of unformatted real-world data. Mostly, real-world data is composed of –inaccurate data (missing data) - There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.

The presence of noisy data (erroneous data and outliers) - The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.

Inconsistent data - The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more. For our diabetes dataset we need to perform preprocessing in two steps.

Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through

Volume: 14 Issue: 01 Pages: 5294-5299(2022) ISSN: 0975-0290

eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.

Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.

VI. TRAIN AND TEST DATA CREATION

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

The size of training set will be larger than that of test set. The training set will be trained tested against the test set. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on some diagnostic measurements included in the dataset.

VII. APPLYING MACHINE LEARNING

Machine Learning is a method of statistical learning where each instance in a dataset is described by a set of features or attributes. In contrast, the term "Deep Learning" is a method of statistical learning that extracts features or attributes from raw data.

Deep Learning does this by utilizing neural networks with many hidden layers, big data, and powerful computational resources. The terms seem somewhat interchangeable, however, with Deep Learning method, the algorithm constructs representations of the data automatically. In contrast, data representations are

hard-coded as a set of features in machine learning algorithms, requiring further processes such as feature selection and extraction, (such as PCA).

Both of these terms are in dramatic contrast with another class of classical artificial intelligence algorithms known as Rule-Based Systems where each decision is manually programmed in such a way that it resembles a statistical model.

In Machine Learning and Deep Learning, there are many different models that fall into two different categories, supervised and unsupervised. In unsupervised learning, algorithms such as k-Means, hierarchical clustering, and Gaussian mixture models attempt to learn meaningful structures in the data. Supervised learning involves an output label associated with each instance in the dataset. This output can be discrete/categorical or real-valued. Regression models estimate real-valued outputs, whereas classification models estimate discrete-valued outputs. Simple binary classification models have just two output labels, 1 (positive) and 0 (negative). Some popular supervised learning algorithms that are considered Machine Learning: are linear regression, logistic regression, decision trees, support vector machines, and neural networks, as well as non-parametric models such as k-Nearest Neighbors. In this research we are adopting supervised learning approach.

VIII. LOGISTIC REGRESSION

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. We can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a legit function.

Logistic regression was developed by statistician David Cox in 1958. This binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the Presence of a risk factor increases the probability of a given outcome by a specific percentage.

Sigmoid function P = 1/1+e - (a+bx)

Here P = probability, a and b = parameter of Model.

IX. RANDOM FOREST

It is one of ensemble learning methods and also used for classification and regression tasks.This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

Algorithm

a. The first step is to select the "R" features from the total features "m" where R<<M.

b. Among the "R" features, the node using the best split point.

c. Split the node into sub nodes using the best split.

d. Repeat a to c steps until "l" number of nodes has been reached.

e. Built forest by repeating steps a to d for "a" number of times to create "n" number of trees.

The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place. Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the

Volume: 14 Issue: 01 Pages: 5294-5299(2022) ISSN: 0975-0290

options of Random Forest does correct predictions result for a spread of applications are offered.

X. DECISION TREE

Decision tree is a type of classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc.

Algorithm

a. Construct tree with nodes as input feature.
b. Select feature to predict the output from input feature whose information gain is highest.
c. The highest information gain is calculated for each attribute in each node of tree.
d. Repeat step 2 to form a sub tree using the feature which is not used in above node.

XI. TECHNICAL REQUIREMENTS

1. HARDWARE REQUIREMENTS

- System Processor: Intel or Amd.
- Hard Disk: 500 GB.
- Ram: 4 GB.

2. SOFTWARE REQUIREMENTS

- Operating system: Windows 8 / 10
- Programming Language: Python and Javascript.
- DL Libraries: Jumpy, Pandas, Sci-kit learn, Anaconda, React and Django.

3. LANGUAGE SPECIFICATION

Python programming language is easy to learn, and a powerful programming language. Because of the interactive, interpreted and object-oriented native of python we decided to use python as a programming language of this project.

4. MODEL BUILDING

This is most important phase which includes model building for prediction of diabetes. In this we have implemented machine learning algorithm which is discussed above for diabetes prediction.

5. PROCEDURE

Step1: Import required libraries, Import diabetes dataset.
Step2: Pre-process data to remove missing data.
Step3: Perform percentage split of 70% to divide dataset as Training set and 20% to Test set.
Step4: Choose the machine learning algorithm i.e. Decision Tree, Logistic regression and Random Forest
Step5: Create the model for the mentioned machine learning algorithms based on training set.
Step6: Test the model for the mentioned machine learning algorithms based on test set.
Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
Step8: After analyzing based on various performance measures conclude the best performing algorithm.

6. PERFORMANCE EVALUATION

We will test our model on our prepared dataset and also measure the performance of the algorithms on our dataset. To evaluate the performance of our created classification and make it comparable to current approaches, we use Accuracy to measure the effectiveness of classifiers.

7. PERFORMANCE MEASURES

True Positives (TP) - These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.
True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing. False positives and false negatives, these values occur when your actual class contradicts with the predicted class.
False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.
False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die. Once you understand these four parameters then we can calculate Accuracy, Precision, Recall and F-measure.

RESULTS ANALYSIS AND DISCUSSION

After a great learning experience and going through careful observations of various models, considering much different approach for a particular result, we finally evaluate the major results and explore the right outcomes. The various Machine Learning algorithms taken into account proved to be extremely useful for the collection of right data for cross validation.
For every machine learning model we have created. We will calculate accuracy, precision, recall and f-measure. These are one of the most important performance measures in Machine Learning.

PRECISION

Precision score represents the model's ability to correctly predict the positives out of all the positive prediction it made. Precision score is a useful measure of success of prediction when the classes are very imbalance.

Volume: 14 Issue: 01 Pages: 5294-5299(2022) ISSN: 0975-0290

Algorithms	Accuracy(%)	Precision	Recall	F-measure
Logistic Regression	76%	0.77	0.55	0.64
Random Forest	75%	0.70	0.51	0.59
Decision Trees	73%	0.60	0.58	0.62

Mathematically, it represents the ratio of true positive to the sum of true positive and false.

$$P = \frac{TP}{TP} + FP$$

The same score can also be obtained by using precision_score method from sklearn.metrics.

RECALL

Recall score represents the model's ability to correctly predict the positives out of actual positives. Recall score is useful measure of success of prediction when the classes are very imbalanced. Mathematically, it represents the ratio of true positive to the sum of true positive and false negative.

$$R = \frac{TP}{TP} + FN$$

The same score can also be obtained by using precision_score method from sklearn.metrics.

ACCURACY

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

$$Accuracy\ (Acc) = \frac{TP + TN}{(TP + TN + FP + FN)}$$

The same score can also be obtained by using accurary_score method from sklearn.metrics.

F-MEASURE

F-measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost.

$$F1 = 2 * precision * recall / precision + recall$$

The same score can also be obtained by using f1_score method from sklearn.metrics.

RESULTS

The result of the experiments is presented here. All these algorithms are measured based on accuracy, precision, recall and F-measure. From the results, Logistic Regression outperformed other algorithms in terms of accuracy, precision and F-measure.

Table 4.1: Comparison of Algorithms

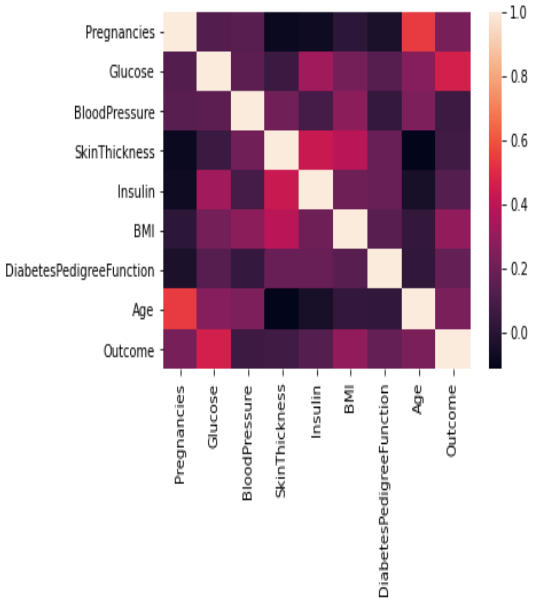

Figure 4 Data correlation

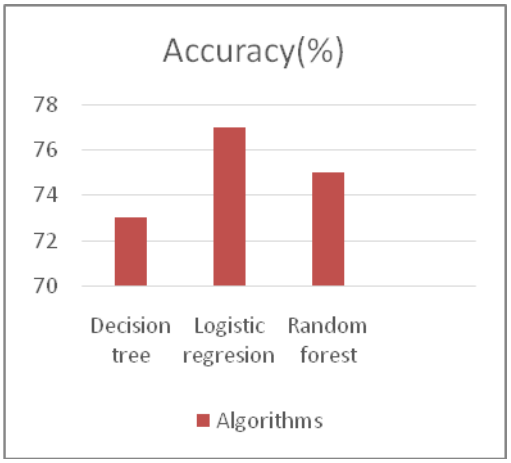

Figure 5 Accuracy result of machine learning algorithms

Volume: 14 Issue: 01 Pages: 5294-5299(2022) ISSN: 0975-0290

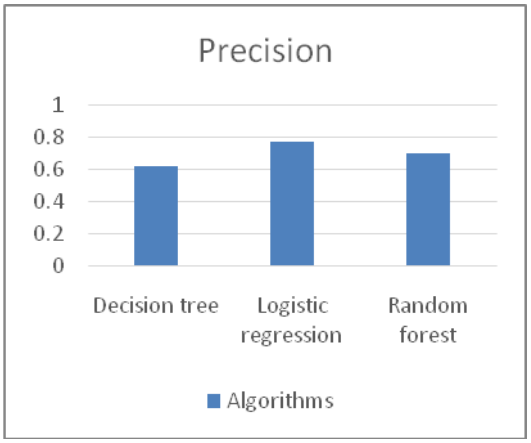

Figure 6 Precision Result Of Machine Learning Algorithms

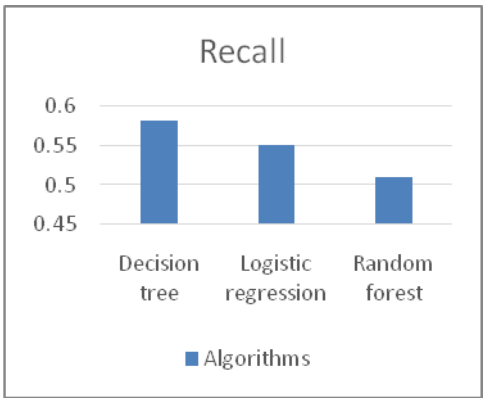

Figure 7 Recall results of machine learning algorithms

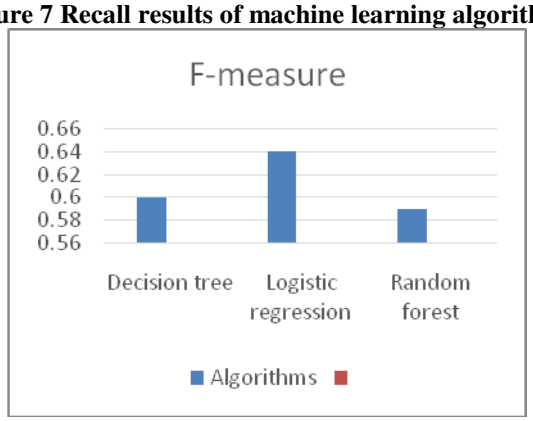

Figure 8 F-measure result of machine learning algorithms

DISCUSSION

A review of the accuracy, F-measure, Recall, and Precision measures in Table 4.1 indicated that Logistic Regression and Random forest were the top performers with many of the algorithms having highly comparable results and Decision tree came least. Given such similar results selection of the choice of algorithm to use is not obvious. As interested parties may have a preference for the choice of algorithm they would like to implement it is important to know if the use of a particular algorithm(s) would result in a statistically lower performance.

CONCLUSION AND FUTRE WORKS

There is a concern among physicians how to detect diabetes at its infancy stage. This research work had made an effort in designing the system in predicting the diabetes using multiple algorithms and comparing their performance. The work implemented three Machine learning algorithms and the evaluation was done on various measures. The experiment was carried on the PIMA Indian Diabetes dataset and the results confirmed that Logistic regression had the best performance which had accuracy score of 76%, precision score of 0.77 and f-measure score of 0.58. This Machine learning algorithm can also be customized in predicting other alternative diseases. The research can be further enhanced in implementing other machine learning algorithm in improving the prediction of diabetes.

REFERNCES

[1] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[2] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13.

[3] Jitranjan Sahoo, Manoranjan Dash & Abhilash Pati, “Diabetes Prediction Using Machine Learning Classification Algorithms”, International Research Journal of Engineering and Technology, Vol. 7, Issue 8, August 2020.

[4]Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[5] Mitushi Soni, ‘Diabetes Prediction using Machine Learning Techniques’, International Journal of Engineering Research & Technology, Vol. 9, Issue 9, September 2020.