Volume: 7 Issue: IV Month of publication: April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4435

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Stacked Diabetes Prediction using Machine Learning Algorithms

B. Raghavendra Reddy[1], D. Mahendra Reddy[2], A. Mahammad Suhel[3], D. Anusha[4]
[1, 3, 4]Student, Dept of CSE, JNTUACEP, Pulivendula, India, AP
[2]Adhoc Lecturer, Dept of CSE, JNTUACEP, Pulivendula, India

Abstract: Nowadays, diabetes has become a common disease to the mankind from young to the old persons. The growth of the diabetic patients is increasing day-by-day due to various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc. Hence, predicting the diabetes earlier is very essential to save the human life from diabetes. In health care, this analytical process is carried out using machine learning algorithms for analyzing medical data to build the machine learning models to carry out medical diagnosis. Moreover, this is an approach to improve the accuracy in diabetes prediction using medical data with various machine learning algorithms and methods.

In this Project, we proposed various Machine learning approaches such as Logistic regression, Decision Trees, Gradient Boosting and Extra tree Classifier. Here the predictive model is build using the medical data of the people who are lived along the Gila and Salt rivers of Arizona U.S.

Keywords: Machine Learning, Diabetes Patients, Classification algorithms, Decision tree, Gradient Boosting, Stacked Generalization, F1_Score.

I. INTRODUCTION

Machine learning is the science of getting computers to learn without being explicitly programmed. In the past few years, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today such that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI.

A. Machine Learning

The Machine Learning field evolved from the broad field of Artificial Intelligence, which aims to mimic intelligent abilities of humans by machines. In the field of Machine learning one considers the important question of how to make machines able to “learn”. Learning in this context is understood as inductive inference, where one observes examples that represent incomplete information about some “statistical phenomenon”. In unsupervised learning one typically tries to uncover hidden regularities (e.g. clusters) or to detect anomalies in the data (for instance some unusual machine function or a network intrusion). In supervised learning, there is a label associated with each example. It is supposed to be the answer to a question about the example. If the label is discrete, then the task is called classification problem – otherwise, for real valued labels we speak of a regression problem. Based on these examples (including the labels), one is particularly interested to predict the answer for other cases before they are explicitly observed. Hence, learning is not only a question of remembering but also of generalization to unseen cases.

II. MOTIVATION

Diabetes disease is one of the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of Diabetes. Predicting whether an individual having diabetes or not based on the test results of that individual using some classification algorithms with more accuracy and less training time.

Main motivation to take this project is that earlier prediction of Diabetes plays vital role in saving the human life from dreadful diseases like kidney failure, heart diseases etc. Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem. Data classification problem is studied by statisticians and machine learning researchers. Data classification is widely used in variety of engineering and scientific disciplines such as biology, psychology, medicines, marketing, computer vision, and artificial intelligence. The goal of the data classification is to classify objects into number of categories or classes. For a given dataset the task of classification is to assign a class to the data object.

III. RELATED WORK

In recent research works, several data mining models have been developed to predict diabetes disease in the medical field by the physicians such as diagnosis support system , expert system, intelligent diagnosis system, and hybrid intelligent system .The diabetes data warehouse contains the screening the data of diabetes patients. Initially, the data warehouse is pre-processed to make the mining process more efficient. Later on as compared to Machine learning algorithms techniques data mining techniques are lagging behind in performance and accuracy.

IV. PROPOSED SYSTEM

Using data mining techniques predicting patients diabetes disease is a time consuming task which degrades patients survival rate, By Appling different machine learning techniques An early diagnosis of diabetes problems will increase patients' survival rate based on accuracy and F1_score to find the best suitable algorithm for diagnosis of diabetes disease which gives best performance. It added a greater advantage to medical field.

Some of the classification algorithms used are

A. Decision trees Bench Mark
B. Gradient Boosting
C. Stacked Generalization

V. EVALUATION METRICS

The f1_score is one of the evaluation metrics used in the benchmark model as well as the proposed project. It is created by finding the harmonic mean of precision and recall.

A. F1 = 2 x (precision x recall)/(precision + recall)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations

B. Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class – yes

C. Recall = TP/TP+FN

Another metric to be used is the confusion matrix which gives us the estimates of no of true positives, true negatives ,false positives and false negatives. Since our dataset is imbalanced dataset then the F1 score works better when compared to the accuracy. F1 Score 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.

VI. IMPLEMENTATION

A. Data Set & Data Description

The dataset used for this project is the pima-indians-diabetes.csv, which was picked from the website kaggle.com and original owners are National Institute of Diabetes and Digestive and Kidney Diseases . Number of instances-768,Number of attributes-8 plus class(all numeric)

1) feature1 -Number of times pregnant
2) feature2 -Plasma glucose concentration,2 hrs in an oral glucose tolerance test
3) feature3 - diastolic blood pressure (mm Hg)
4) feature4 - Triceps skin fold thickness(mm)
5) feature5 - 2 hr serum insulin (mu U/ml)
6) feature6 -Body mass index (weight in kg/(height in m)^2)
7) feature7 -Diabetes pedigree function
8) feature8 -Age(years)
9) feature9 – Target Class variable(0 or 1)

B. Data Preprocessing

Data Preprocessing The data consisted of zeroes at inappropriate places,it is impossible for some of these features to have a value of zero. While entering the data, the missing values were put as 0 in five of these features, namely, glu_conc, dia_bp,tris_skin,ser_insulin and bmi. The pandas data.describe() function was used to confirm the suspicion and carry out statistical analysis of the features.

Some of these features had their first quartiles and min values as zero. The zeroes in the above mentioned 5 features were replaced by nan, using pandas.replace() function.Individual counting of the no of missing values on each of these features was performed

1) Missing values in glucose Concentration : 5
2) Missing Values in Dia_Blood Pressure : 35
3) Missing values in Triceps skin Thickness : 227
4) Missing values in 2-hour serum Insulin : 374
5) Missing values in body Mass Index: 11 The glucose concentration features missing values were replaced by its mean.

For Diastolic blood pressure, the missing values were replaced by its median.The mean and median were same for this feature.Triceps skin thickness missing values were replaced by its median as the no of missing values is too high, the appropriate metric for this feature was the median, similar replacement was done in the case of serum insulin.

The body mass index was replaced by its mean. After this, the data was separated into two separate variables, X and y, i.e features and labels. The features were scaled together using MInmaxscaler function from the preprocessing module of scikit learn. This data was Stored in X. X and y were later fed into the cross validation module to be split into training and testing states, the split was decided to be 70-30.

VII. CLASSIFICATION TECHNIQUES

A. Decision Tree Classifier

The decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, it's also widely used in machine learning.

1) Advantages

a) Able to handle categorical and numerical data.
b) Doesn't require much data pre-processing, and can handle data which hasn't been normalized, or encoded for Machine Learning Suitability.
c) Simple to understand ,visualize and interpret.

2) Disadvantages

a) Complex Decision Trees do not generalize well to the data and can result in over fitting.
b) Unstable, as small variations in the data can result in a different decision tree.

B. Gradient Boosting

is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

1) Advantages

a) Gradient Boosted Methods generally have 3 parameters to train shrinkage parameter, depth of tree, number of trees. Now each of these parameters should be tuned to get a good fit.
b) It gives better results than Random Forests.

2) Disadvantages

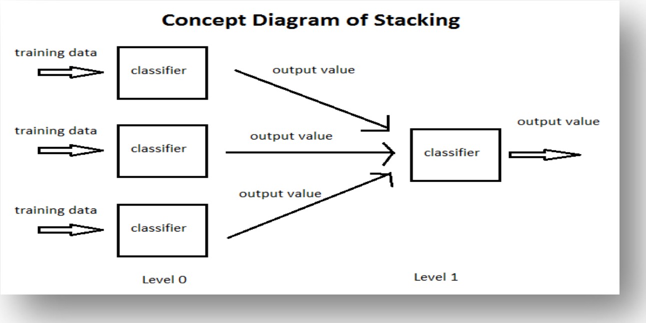
a) Gradient based methods generally give better results but they are harder to fit than Random Forests.

C. Stacked Generalization

Stacking, Blending and and Stacked Generalization are all the same thing with different names. It is a kind of ensemble learning.

In traditional ensemble learning, we have multiple classifiers trying to fit to a training set to approximate the target function. Since each classifier will have its own output, we will need to find a combining mechanism to combine the results. This can be through voting (majority wins), weighted voting (some classifier has more authority than the others), averaging the results, etc. This is the traditional way of ensemble learning.

In stacking, the combining mechanism is that the output of the classifiers (Level 0 classifiers) will be used as training data for another classifier (Level 1 classifier) to approximate the same target function. Basically, you let the Level 1 classifier to figure out the combining mechanism.



The final model is evaluated using F1_score.The input data is split into train and test dataset.
The train dataset is used to used to train the model and then the test dataset is used to see how the model does on previously unseen data.

F1 Score

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

1) *Recall:* Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}}$$

2) *Precision:* To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP).

Precision is given by the relation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}}$$

VIII. CONCLUSION

By using machine learning algorithms we can easily predict the outcomes of diabetes. These models are trained with the dataset and gives the accuracy score. From all the algorithms Stacked Generalization algorithm gives the best accuracy score in the prediction. So it is selected as best algorithm from the others.