EasyChair Preprint

№ 2864

The Prediction of Diabetes in Pima Indian Women Mellitus Based on XGBOOST Ensemble Modeling Using Data Science

Dasari Bhulakshmi and Glory Gandhi

March 5, 2020

The Prediction of Diabetes in Pima Indian women Mellitus Based on XGBOOST Ensemble Modeling using data science

Authors
Dasari Bhulakshmi
Assistant professor
School of computing & information technology
REVA University
Bangalore
Glory Gandhi
Assistant professor
School of Engineering & Technology
Computer Science and Engineering
CMR University
Bangalore

ABSTRACT

Healthcare systems provide personalized services in widespread domains to help patients in fitting themselves into their normal activities of life. This study is focused on the prediction of diabetes in pima Indian women mellitus based on XGBOOST. Types of patients based on their personal and clinical information using a boosting ensemble technique that internally uses random committee classifier. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, Crowd Analytics. These are the most preferred machine learning algorithms today. To evaluate the technique, a real set of data containing 100 records is used. The prediction accuracy obtained is 81.0% based on experiments performed in Weka with 10-fold cross validation.

Keywords - XGBOOST, gradient boosting; medical data mining; machine learning; diabetes mellitus; disease prediction model

I. INTRODUCTION

Diabetes mellitus is a veteran disease that was quoted in Egyptian Palimpsest 3000 years ago. Due to burgeoning nature of big data in health industry, it is mandatory to understand its magnitude into ostensible value with desirable solution. One of the dominant Non-Communicable diseases is Diabetes Mellitus. The subtle nature of diabetes mellitus has long haul complications. Clinical data set engendered from electronic health records plays a sizable role in prediction and diagnosing of Diabetes Mellitus. Support vector machine, Naive bayes and decision tree, Adaboost's algorithm were previously enforced techniques for the prediction and diagnosis of diabetes mellitus. In this paper, we have used gradient boosting algorithm for dexterity and accuracy.

Instead of providing guidance to new sample distribution, the weaker learners are given coaching based on the errors of strong learner. It is done by using gradient boosting escalation process. Based on this inference, it is evident that, gradient boosting is prominent than adaboosting algorithm in this paper, the degree of accuracy for prediction and diagnosing of diabetes mellitus is increased by gradient boosting algorithm. XGBoost is one of the implementation method of gradient boosting models that provides high relation in the performance and speed to the model. It is very feasible with tunable parameters

II. BACKGROUND

Diabetes mellitus can be classified into two types. TYPE 1 and TYPE 2 diabetes. TYPE 1 is the insulin dependent diabetes whereas; TYPE 2 is the noninsulin dependent diabetes. Later is the most prevalent form of diabetes mellitus marked by Hyperglycemia and insulin glitch. Recognition of the cause of disease, precautionary measures should be inaugurated to downturn the dimensions of Diabetes Mellitus. Diabetes mellitus is a group of metabolic disorders where the blood sugar levels are higher than normal for prolonged periods of time [1]. Diabetes is caused either due

to the insufficient production of insulin in the body or due to improper response of the body's cells to Insulin. The former cause of Diabetes is also called Type 1 DM or Insulin-dependent Diabetes mellitus and the latter is known as Type 2 DM or Non-Insulin Dependent DM. Gestational Diabetes is a third type of Diabetes where women not suffering from DM develop high sugar levels during pregnancy. In the United States, 30.3 million Americans were recorded as suffering from Diabetes with 1.5 million being diagnosed with Diabetes every year. Total cost of diagnosed Diabetes in the US in 2017 was $327 billion [2]. Diabetes is especially hard on women as it can affect both the mother and their unborn children during pregnancy. Women with Diabetes have a higher likelihood at having a heart attack, miscarriages or babies born with birth defects [3].

Tests involved in Type 1 diabetes are

1 C peptide test or auto antibodies test

2 A finger stick glucose test

3 Urine tests

Tests involved in type 2 diabetes includes

1 Fasting blood sugar

2 Two-hour post prandial test

3 Random blood sugar

4 Hemoglobin A1C test

5 Oral glucose tolerance tests

Machine learning and data mining methods provides imperative efforts to transmute feasible information into profitable knowledge. According to the statistics,85% of supervised learning and 15% of unsupervised learning approaches were used. Inconsistency can occur due to large volume of data and these discrepancies can be solved by using distinctive classification techniques using gradient boosting. This new statistical machine learning approach would benefit many people to predict diabetes possibilities beforehand thereby which could save them from heart strokes, kidney failure and losing of eyesight etc.

III. RELATED WORK

Prediction and diagnosis of diabetes mellitus has been experimented by different classification and clustering techniques. Few related works are listed below.

Decision support system built using Adaboost's algorithm with decision - stump as base classifier has an accuracy of 80.72%. This system was implemented by WEKA - MATLAB interface tool.

Genetic Algorithm was incorporated in the dataset to analyze and predict the diabetes using MATLAB which produced accuracy rate of 80%.Other Data mining techniques used on different diabetes datasets are Modified J48 Classifier, C4.5,Bayesian Network, Amalgam KNN and ANFIS, Artificial Neural Network and PLS-LDA with various tools such as Tanagara, WEKA,MATLAB, GP Lab tool Box, Clementine etc.

Motivation and Goal of study

Due to increasing incidence rate of diabetes and prediabetes, it is a pressing issue in the health care industry to rightly identify the factors that contribute to the occurrence of Diabetes in people, more so, in Women. From secondary research, factors such as BMI, Blood Pressure, Cholesterol and Glucose levels are important factors that cause Diabetes. In Women, Pregnancy seems to be an additional factor. According to the World Health Organization, people with 2-hour post-load plasma glucose levels at least 200 mg/dl (11.1 mmol/l) at any survey examination were diagnosed with Diabetes [5]. To validate the above hypotheses, identify additional risk factors and build tools that can predict the occurrence of Diabetes, particularly in women, the Pima Indians' Diabetes dataset was chosen.

IV. PROPOSED MODEL

The proposed model uses gradient boosting algorithm to predict diabetes with high accuracy and fast execution time implemented by xgboost. The diagrammatic representation of the proposed model has been shown below. The model is a regularized model and it has been formalized to control over fitting to better performance. It is trained by ensemble method which is composed of multiple trained weak models to make one single model. The Framework used to build this model was win python environment with xgboost package. This model consists of three phases.

The initial phase deals with collection of datasets. second phase deals with splitting of dataset for training and testing the model. The data split is done with the ratio of 8:2 i.e. 80% and 20% for training and testing respectively. Then model is trained using the xgb classifier -gradient boosting trees. After training models are tested by few predictions. Then model is evaluated in terms of performance, execution time, accuracy, error rate etc.

A. COLLECTING OF DATASET

Training and testing phase is done with the data set obtained from UCI repository of machine learning. Population based prevention and preventing diabetes in people at a high risk are the major instances provided by this dataset. The attributes provided by the data set includes.

The following features have been provided to help us predict whether a person is diabetic or not:

Dataset Attributes

The dataset includes data from 768 women with 8 characteristics, in particular:

1. Number of times pregnant
2. Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

Dataset

SNo.	NumTimesPrg	Glucose	BloodPressure(mmHg)	SkinThickness(mm)	Insulin(muU/ml)	BMI	DiabetesPedigreeFunction	Age(yr)	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

First 5 records in the Pima Indians Diabetes Database

B. SPLITTING THE DATASET

The ratio of splitting of the dataset is one important factor that makes up the execution of the model and best usage of the dataset. The purpose of the splitting involves two category.

1) TRAINING - It evolves preparing and training the model with all provided data samples. 80% of the UCI data samples are used for training the model.

2) TESTING - It evolves validating and making new predictions with data samples. 20% of the dataset are used for testing the model.

```
#Split the dataset into train and Test
seed = 7
test size = 0.3
X_trian, X_test, y_train, y_test = train_test_split(X_data, y, test_size=test_size, random_state=seed).
#Train the XGboost Model for Classification
model1 = xgb.XGBClassifier()
model2 = xgb.XGBClassifier(n_estimators=100, max_depth=8, learning_rate=0.1, subsample=0.5)
```

```
train_model1 = model1.fit(X_trian, y_train)
train_model2 = model2.fit(X_trian, y_train)

#prediction and Classification Report
from sklearn.metrics import classification report

pred1 = train_model1.predict(X_test)
pred2 = train_model2.predict(X_test)

print('Model 1 XGboost Report %r' % (classification_report(y_test, pred1)))
print('Model 2 XGboost Report %r' % (classification_report(y_test, pred2)))
```

Model 1 XGBoost Report

	Precision	Recall	F1-score	support
0	0.83	0.82	0.83	147
1	0.69	0.70	0.70	84
avg/total	0.78	0.78	0.78	231

Model 2 XGBoost Report

	Precision	Recall	F1-score	support
0	0.80	0.78	0.79	147
1	0.63	0.65	0.64	84
avg/total	0.74	0.74	0.74	231

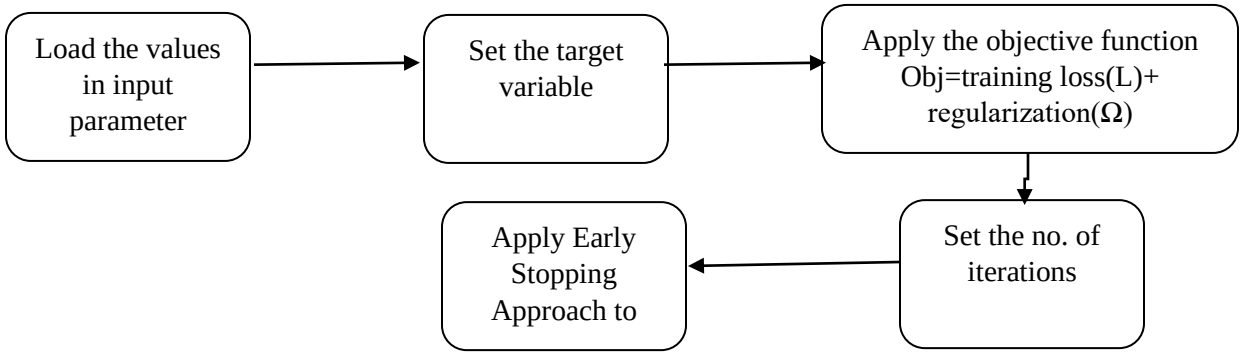
```
#Let's use accuracy score
from sklearn. Metrics import accuracy score

print("Accuracy for model 1: %.2f" % (accuracy score(y_test, pred1) * 100))
print("Accuracy for model 2: %.2f" % (accuracy score(y_test, pred2) * 100))
Accuracy for model 1: 77.92
Accuracy for model 2: 73.59
```

Now that we have transformed the data, we need to split the dataset in two parts: a training dataset and a test dataset. Splitting the dataset is a very important step for supervised machine learning models. Basically we are going to use the first part to train the model (ignoring the column with the pre assigned label), then we use the trained model to make predictions on new data (which is the test dataset, not part of the training set) and compare the predicted value with the pre assigned label.

Training of model is being carried out by gradient boosting machine implemented by xgboost. The decision tree is boosted by means of gradient descent. XGBoost has taken data science competition by storm. XGBoost seems to be a part of an ensemble of classifiers/predictors which are used to win data science competitions.
The training involves all the key features of xgboost to get the best model for predictions of diabetes mellitus. Information are provided to the Parameter of the model to perform xgb classifier

C. TRAINING THE MODEL



Step 1: Input - load the numeric values in the input parameter.
Step 2: Target variable - For classification Problem, the target variable or vector is either 0 or 1.
Step 3: objective - The objective for binary classification is logistic.
objective function = training loss + regularization i.e. Obj = L + Ω
In xgboost, objective function is optimized by the gradient decsenteuation is required
Step 4: Number of iterations - the number of trees to be added to the model
Step 5: Early stopping - To avoid over fitting the dataset into our model during validation this feature will stop the iteration when it does not see improvement in the accuracy and it will also specify the no of iteration to be stopped before over fitting.
Comparing multiple algorithms:
To compare multiple algorithms with the same dataset, there is a very nice utility in sklearn called model_selection. We create a list of algorithms and then we score them using the same comparison method. At the end we pick the one with the best score.

V. RESULT AND ANALYSIS

The performance of the model can be evaluated by means of the xgboost parameters. After the initial iteration, the accuracy of the model was 77%. After many iterations the accuracy keeps increasing. Gradually from 77% to 90%. The xgboost works on generic loss where asadaboost works on exponential loss. The execution time is three times faster than adaboost algorithm.
To be precise on the current trend xgboost is an only implementation to be very fast in execution. Adaboost's tree will punish for its misclassification. Where xgboost will minimize the error of previous tree and regularize them on the next tree which makes the algorithm more effective and controls over.

CONCLUSION

Here the proposed model uses gradient boosting algorithm. This model uses global data set from UCI repository of machine learning. The accuracy of the system can be improved with the implementation of other powerful ensemble methods by using local datasets from various places. This proposed model provides the accuracy and dexterity of 90% for predicting diabetes with less error rates. In Future, Light GBM can be used to improve the scalability to incorporate large data, improve the execution speed and produce 100% accuracy fitting.
The PIMA Indian Women's Database was analyzed and explored in detail. The patterns identified using Data exploration methods were validated using the modeling techniques employed. Classification models such as Logistic Regression, Classification Trees, Random Forest and SVM were built and evaluated to identify best model to predict the occurrence of Diabetes in PIMA Indian women. From the cross-validated performance measure of sensitivity, the Logistic Regression model was concluded as the best performing model.

REFERENCES

1.https://en.wikipedia.org/wiki/Diabetes_mellitus
2.http://www.diabetes.org/diabetes-basics/statistics/
3.http://www.diabetes.org/living-with-diabetes/treatment-and-care/women/
4. Veena vijayan.V, Anjali.C Prediction and Diagnosis of Diabetes Mellitus Machine Learning Approach Published in: Intelligent Advance.
5. Diagnosis of diabetes mellitus using extreme learning machine Published in: Information Technology Systems and Innovation (ICITSI), 2014 International Conference
6. A Novel approach to predict diabetes mellitus using modified Extreme learning machine Published in: Electronics and Communication Systems (ICECS), 2014 International Conference.
7. http://www.saedsayad.com/docs/xgboost.pdf
8. Performance Analysis of Classification Algorithms in Predicting Diabetes, International Journal of Advanced Research in Computer Science Volume 8, No. 3, March –April 2017
9.Tianqi Chen, author of XGBoost, XGBoost: A Scalable Tree Boosting System
10.https://machinelearningmastery.com/gentle-introductionxgboost-applied-machine-learning/
11.
http://xgboost.readthedocs.io/en/latest/model.html#finalwords-on-xgboost
12.https://machinelearningmastery.com/tune-number-sizedecision-trees-xgboost-python