Pregnant Women Diabetes Prediction using Machine Learning approach

Akash Rakeshkumar Rana
Computer Science and Engineering Department
R.N.G Patel Institute of Technology Bardoli India

Prof. Dhaval J. Rana
Computer Science and Engineering Department
R.N.G Patel Institute of Technology Bardoli India

Abstract*— Now a days the food and diet cycle of any person is so inappropriate. Due to this many long-life disease happened with that person. Diabetes is one of the disease that may end the human life. It is metabolic disease that causes due to high blood sugar. It causes a heart related problem, blood pressure, kidney problem and eye damage. High blood sugar level is the main cause of diabetes. It can also cause various health issues in a person. If it is not treated correctly, it can lead to major issues in a person. I generate a machine learning code for the diabetes prediction specially for pregnant women. This report presents a method for Pregnant Women Diabetes Prediction using a XgBoost and Random Forest Classifier. We Predict the Pregnant Women is Diabatic or not according to some parameters. The accuracy of program is different at two algorithm.

Keywords— Seaborn, Random Forest algorithm, XgBoost algorithm

Introduction.

In the generation of Fast-food and inappropriate routine , a Human life suffer from many long term disease. Those disease are much dangerous that if we do not start proper treatment on time then it take a life of human. Aside of these problem many people are dying in village due to this disease because of lack of knowledge , less medical equipment , less hospital , time consuming reports and many more reason. Those disease are different types of cancer, asthma , Heart disease , diabetes and many more . [1] Diabetes is major disease that it occurs when pancreas does not produce enough of the hormone insulin.

In 2030, the prevalence of diabetes globally could reach 490 billion. In India, it is estimated that there are around 40 million people with diabetes. This figure is higher than the population of Canada and China.

This study aims to predict diabetes by taking into account various attributes related to the condition of diabetes. [2]

Various Machine Learning techniques help to collect knowledge efficiently. These techniques are mainly used for learning various classification and object models.

1. Problem Statement

Diabetes is also major problem for Pregnant women. Gestational diabetes occurs during pregnancy but may resolve after the baby is delivered. Having a family history of diabetes makes it more likely that a woman will develop gestational diabetes, which suggests that genes play a role. But in many cases due to not proper treatment on time both baby and women may die. [3] The main reason is many reports are too much time consuming and due to this we will not start and treatment on time. [4]

2. Proposed Methodology

The main goal of the paper is to find better model for the prediction of Pregnant Women Diabetes with highest accuracy. We use different classification algorithm and ensemble them to get maximum accuracy.

A. *Dataset Discription:*

The Pima Indian Diabetes Dataset(PIDD) has been employed in this study, provided by the UCI Machine Learning Repository. The dataset has been originally collected from the National Institute of diabetes and organic process and urinary organ Diseases. The dataset consists of some medical distinct variables, like gestation record, BMI, internal secretion level, age, aldohexose concentration, heartbeat force per unit area, skeletal muscle skin fold thickness, diabetes pedigree function etc. This Dataset has 768 patient’s data wherever all the patients are feminine and a minimum of twenty-one years previous.The quantity of true cases is 268 (34.90%) and therefore the varieties of false cases are 500(65.10%), severally, within the dataset. [5]

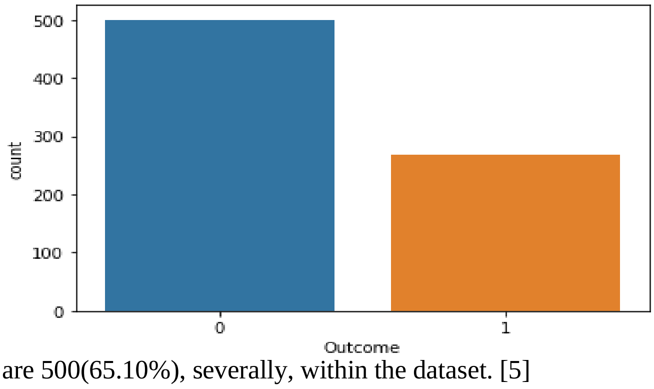


Fig-1(a): Ratio of Diabetic and Non Diabetic Patient

B. *Data Preprocessing:*

Data pre-process is a technique to making a raw data suitable for machine learning model. It is not possible that data is formatted and clean all the time. Sometimes there is null values , inappropriate input , missing values and many more. We need to clean those data for our model. [6]

For PIMA dataset we need to perform Data Preprocessing in two way such as:

I. Remove Null value:

Removing null values from the dataset is one of the important steps in data wrangling. These null values adversely affect the performance and accuracy of any machine learning algorithm. [7]

II. Split data after:

After removing null values the data is now suitable for the model but before use them we need to split it. We split those data for training and testing model. [8]

C. Random Forest:

This method is commonly used for classification and Regression tasks. It provides better accuracy than other models. Random Forest is a type of classifier that combines multiple decision trees on a given dataset. [9]Instead of relying on one tree, it takes the prediction from all the trees and predicts the final output. It able to manage large amount of data.

Algorithm-

- The first step is to select the “R” features from the total features “m” where R<<M.
- Among the “R” features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until ”l” number of nodes has been reached.
- Built forest by repeating steps a to d for “a” number of times to create “n” number of trees.

The random forest finds the best split using the Gin-Index Cost Function which is given by:

$$Gini = \sum_{k=1}^{n} p_k * (1 - p_k)\ Where\ k = Each\ class\ and\ p = proption\ of\ training\ instances$$

D. XgBoost:

Xgboost algorithm is implementation of Gradient boosting. It is dominating algorithm nowadays in machine learning application. It use for getting more speed and performance compare to other algorithm. [10]

Algorithm-

- Consider a sample of target values as P
- Estimate the error in target values.
- Update and adjust the weights to reduce error M.
- P[x] =p[x] +alpha M[x]
- Model Learners are analyzed and calculated by loss function F
- Repeat steps till desired & target result P.

3. BUILD A MODEL:

Step 1: Import suitable libraries and PIMA Dataset.

Step 2: Pre-processing the data and remove null values.

Step 3: Split data such as 70% data to Training set and 30% data to Testing set.

Step 4: Select ML algorithm such as Random Forest and XgBoost.

Step 5: Build a Classifier model based on the Training set.

Step 6: Test the Classifier model for the ML algorithm based on Test data.

Step 7: Perform Comparison evaluation of the experimental performance results obtained from each classifier.

Step 8: After analyze based on various measures conclude the best performing algorithm.

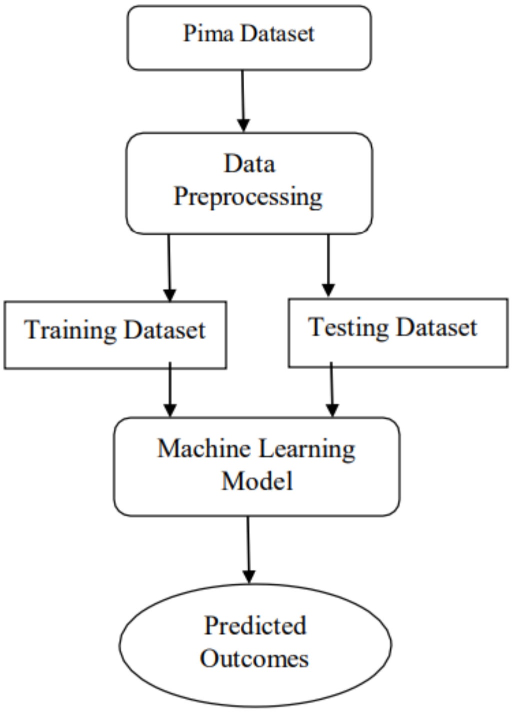
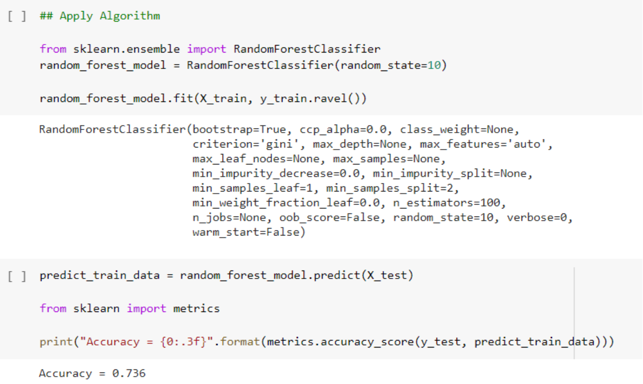


Fig-3(a): Overview of the Process

4. EXPERIMENTAL RESULT:

- Random Forest Algorithm:



```
[ ] ## Apply Algorithm

    from sklearn.ensemble import RandomForestClassifier
    random_forest_model = RandomForestClassifier(random_state=10)

    random_forest_model.fit(X_train, y_train.ravel())

    RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                           criterion='gini', max_depth=None, max_features='auto',
                           max_leaf_nodes=None, max_samples=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=100,
                           n_jobs=None, oob_score=False, random_state=10, verbose=0,
                           warm_start=False)

[ ] predict_train_data = random_forest_model.predict(X_test)

    from sklearn import metrics

    print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, predict_train_data)))

    Accuracy = 0.736
```

Fig-4(a): Random forest algorithm

According to Random Forest algorithm we get 0.736 accuracy. Which means our model is 73.6% accurate.

- XgBoost Algorithm:

According to XgBoost Algorithm we get 0.743 accuracy of model which is higher than Random forest Algorithm. Which means our model is 74.3% accurate.

5. CONCLUSION:

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which Random Forest and Xgboost classifiers are used which are implemented using python. And 74% classification accuracy has been achieved. The Experimental results can be assist health care to take early prediction and make early decision to cure diabetes and save humans life.

6. FUTURE WORK:

In order to diagnose diabetes correctly, the patient must first determine if he or she is in a non-diabetic category or a diabetic one. But, in most cases, this process leads to errors in diagnosis and treatment. In order to severity of such impact there is need to create machine learning model which will provide accurate result and able to make early prediction of diabetes and start a treatment before its late.

7. BIBLIOGRAPHY

[1] "National Institute of diabetes and diagstive and kidney diesease," December 2016. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes.

[2] G. rojlik, "Research Gate," January 2016. [Online]. Available: https://www.researchgate.net/publication/312383296_WHO_Global_report_on_diabetes_A_summary#:~:text=...-,According%20to%20a%20World%20Health%20Organization%20report%2C%20422%20million%20people,2017%20(Roglic%2C%202016)%20..

[3] "Health Line," [Online]. Available: https://www.healthline.com/health/gestational-diabetes.

[4] B. Metzger, "National Institute of Diabetes and Digestive and Kidney Disease," 2017. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/gestational/symptoms-causes.

[5] P. I. D. Database, "kaggle," 2017. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[6] "Javatpoint," 2019. [Online]. Available: https://www.javatpoint.com/data-preprocessing-machine-learning.

[7] N. Kumar, "The Professional Point," [Online]. Available: http://theprofessionalspoint.blogspot.com/2019/03/data-wrangling-removing-null-values.html.

[8] "GeeksforGeeks," 2019. [Online]. Available: https://www.geeksforgeeks.org/splitting-data-for-machine-learning-models/.

[9] "JavaTpoint," 2019. [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm.

[10] "Data Flair," [Online]. Available: https://data-flair.training/blogs/xgboost-algorithm/.