International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 5, ISSN No. 2455-2143, Pages 151-154
Published Online September 2021 in IJEAST (http://www.ijeast.com)

DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS

Rishab Bothra
Department of Information Technology
Motilal Nehru National Institute of Technology Allahabad, Prayagraj

Abstract— **Diabetes mellitus is a disease in which blood sugars level is abnormally high due to inability of the body to produce or respond normally to insulin. It is among the critical disease and lots of people are suffering from this disease. Due to age, lack of exercise, hereditary diabetes, bad diet, high blood pressure etc. can cause this disease. Healthcare Industries have large volume of databases so by Big Data Analytics we can extract meaningful insights such as hidden patterns, unknown correlations to discover knowledge from the data and predict the outcome accordingly. In this paper we have proposed a diabetes prediction model using Machine Learning algorithm for better classification prediction. We have tried different Machine Learning algorithms to find which gives the better accuracy of classification.**

Keywords— **Diabetes mellitus, Healthcare, Machine Learning, Data Analytics**

I. INTRODUCTION

Over the last decade, technology has been evolved at an incredible rate and it has been interwoven into nearly every aspect of our life including at health sectors. Diabetes mellitus is the most common endocrine disease which is characterized by metabolic abnormalities and by long term complications in the eyes, kidneys, nerves and blood vessels. People having diabetes have high risk of other diseases like heart disease, kidney disease, eye problem, nerve damage, etc. also people with uncontrolled diabetes may develop poor circulation which makes blood to move more slowly so it makes difficult for the body to deliver nutrients to wounds which delays the injury to heal.

There are two major types of Diabetes **Type I diabetes** and **Type II diabetes**. [1] **Type I diabetes** is usually diagnosed in children and young adults and was previously known as juvenile diabetes. It is also known as Insulin-Dependent Diabetes (IDDM). Inability of human’s body to generate sufficient insulin is the reason behind this type of DM and hence patient requires insulin to inject in a body. [2] **Type II diabetes** also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM) is the most common form of DM characterized by hyperglycemia, insulin resistance, and relative insulin deficiency. It is seen when body cells are not able to use insulin properly.

Predictive Analysis incorporates a variety of machine learning algorithms, data mining techniques and statistical analysis uses the current and past data to find knowledge and predict future events.[3] Data Mining includes iterative series like Data Cleaning (removal of noise and inconsistent data), Data Integration (combining data from different sources), Data Selection (selection of relevant data for analysis), Data Transformation, Pattern Evaluation and Knowledge presentation (use of knowledge presentation technique for presenting the minded knowledge to users). Machine Learning can be explained as automating and improving computers learning process based on their experiences. ML is one of the parts of AI. In ML we build a model based on sample data, known as ‘training data’, in order to make predictions or decisions without being explicitly programmed to do so and evaluate the algorithm using metrics.

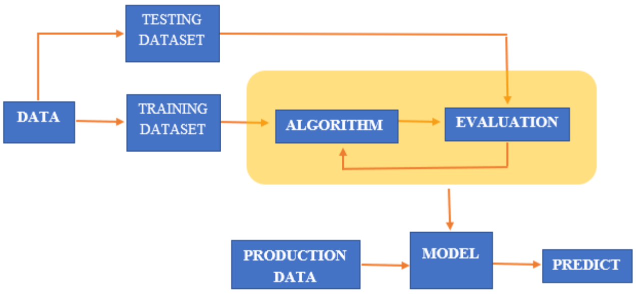


Fig1: Overview of the Workflow of ML

Various prediction models have been developed and implemented by various researchers using variants of machine learning algorithms and deep learning (ANN). Some researchers like-

Muhammad Waqar Aslam (2010) [4] carried out the prediction model in two stages. In first stage, genetic programming to generate an individual from training data that converts the possible characteristics to a single feature such that it has distinct values for healthy and infected (Diabetes) data. Then in second stage test data is used for testing of that individual features.

K. Rajesh and V. Sangeetha (2012) [5] used classification technique. They used decision tree algorithm to find hidden patterns from the dataset for classifying efficiently.

Aiswarya Iyer (2015) [6] used Naive Bayes and Decision Trees Algorithms in the model and comparison was made for performance and effectiveness of both algorithms which was shown as a result.

Rashid (2016). [7] created a prediction model with two combinations of modules to predict diabetes as diabetic or non-diabetic. The first module is ANN (Artificial Neural Network), and the second module is Decision Tree. And it turned out to be FBS (Fasting Blood Sugar) is the critical factor to predict signs of diabetes.

Humar Kahramanli and Novruz Allahverdi (2008) [8] used Artificial neural network (ANN) with fuzzy logic instead of Boolean logic to predict diabetes. Fuzzy logic uses the degree of truth approach rather than usual (True or False) Boolean logic.

B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) [9] proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm, in order to build classifiers decision tree algorithm was used.

Nongyao Nai-arun and Punnee Sittidech, [10] In this paper they explained the role of Adaboost and Bagging Ensemble techniques using decision tree as the basis for classifying the patients as diabetic or non-diabetic. Their experiment results prove that Adaboost machine learning ensemble technique performs well in terms of prediction comparatively bagging as well as a decision tree

II. PROPOSED ALGORITHM

Prediction using Random Forest –

In a single Decision Tree we get low bias and high variance output so to convert high variance into low variance we combine multiple Decision tree (i.e., Random Forest). It works on four steps: -

a) Random samples selected from a given dataset.
b) Construct Decision tree for each sample and get the predicted result for each.
c) Perform a vote for each predicted result
d) Select the prediction result with the greatest number of votes as the Final Prediction.

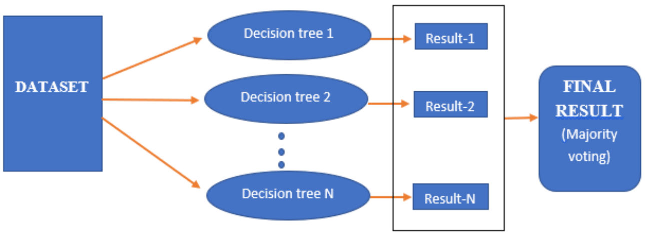


Fig2: Workflow of Random Forest

XGBoost Feature –

It is a decision tree based ensemble technique which uses gradient boosting framework. The two major reason for using xgboost are Execution speed and Model performance.

The three major methods of the gradient boosting that are supported:

a) Gradient Boosting method or algorithm, which is also known as gradient boosting machine together with learning rate.
b) Stochastic Gradient Boosting using sub sampling at column, row, and the column per split level.
c) Regularized Gradient Boosting, which is using both the L1 and L2 regularization.

Support Vector Machine –

It is used to solve both classification and regression types of problem. In SVM there are parallel lines to the hyperplane that separates the classification feature and it passes to one of the nearest point of samples. The samples closest to the margin that were selected to determine the hyper plane is known as support vectors. We can draw many hyperplanes but we need the plane which has maximum margin distance (for linearly separatable points).

III. PROPOSED METHOD

a) Dataset collection – It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries 1405 rows i.e., total number of data and 10 columns i.e., total number of features. Features include Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age & Id.

b) Data Pre-processing – This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id is inconsistent so we dropped the feature. This dataset doesn't contain missing values. So, we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then data was scaled using StandardScaler. Since there were a smaller number of features and important for prediction so no feature selection was done.

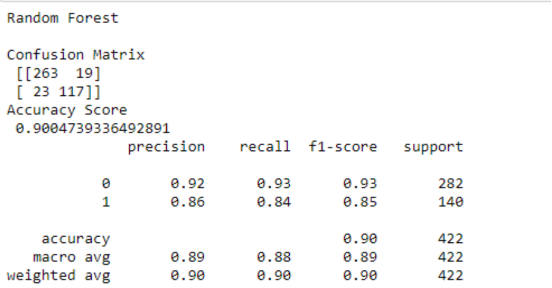
c) Model Building – In this step we tried various algorithms like Logistic Regression, Random Forest, SVM (Support vector machine), xgboost & KNearestNeighbour but according to the problem statement we analyzed the confusion matrix and concluded the false negative (FN) value should kept minimum i.e., if person is actually infected but predicted False then it might be dangerous also other factors like Accuracy & Precision was considered.

Algorithm: - Diabetes Prediction using Random Forest

- Generate train set and test set randomly.
- random_forest_model = RandomForestClassifier(random_state=4219) random_forest_model.fit(X_train, y_train) y_pred1=random_forest_model.predict(X_test)
- **For evaluation** print('Confusion Matrix\n',confusion_matrix(y_test,y_pred1)) print('Accuracy Score\n',accuracy_score(y_test,y_pred1)) print(classification_report(y_test,y_pred1))

IV. Experiment and Result

a) Random Forest



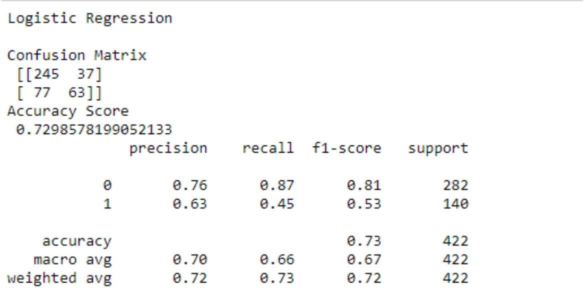
```
Random Forest

Confusion Matrix
 [[263  19]
 [ 23 117]]
Accuracy Score
 0.9004739336492891
              precision    recall  f1-score   support

           0       0.92      0.93      0.93       282
           1       0.86      0.84      0.85       140

    accuracy                           0.90       422
   macro avg       0.89      0.88      0.89       422
weighted avg       0.90      0.90      0.90       422
```

b) Logistic Regression



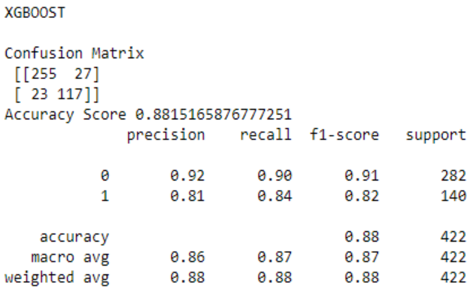
```
Logistic Regression

Confusion Matrix
 [[245  37]
 [ 77  63]]
Accuracy Score
 0.7298578199052133
              precision    recall  f1-score   support

           0       0.76      0.87      0.81       282
           1       0.63      0.45      0.53       140

    accuracy                           0.73       422
   macro avg       0.70      0.66      0.67       422
weighted avg       0.72      0.73      0.72       422
```

C) xgboost



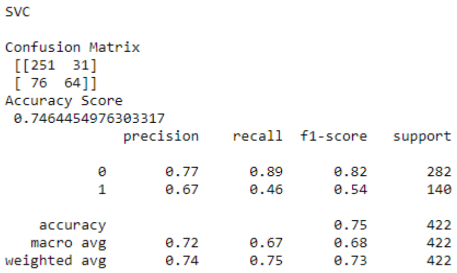
```
XGBOOST

Confusion Matrix
 [[255  27]
 [ 23 117]]
Accuracy Score 0.8815165876777251
              precision    recall  f1-score   support

           0       0.92      0.90      0.91       282
           1       0.81      0.84      0.82       140

    accuracy                           0.88       422
   macro avg       0.86      0.87      0.87       422
weighted avg       0.88      0.88      0.88       422
```

d) Support Vector Machine



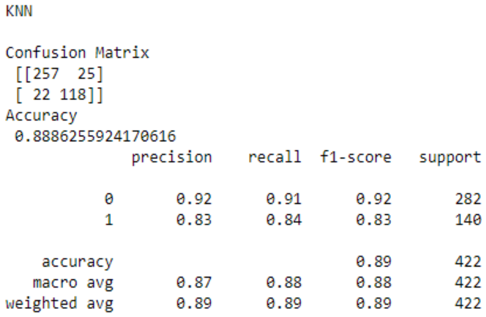
```
SVC

Confusion Matrix
 [[251  31]
 [ 76  64]]
Accuracy Score
 0.7464454976303317
              precision    recall  f1-score   support

           0       0.77      0.89      0.82       282
           1       0.67      0.46      0.54       140

    accuracy                           0.75       422
   macro avg       0.72      0.67      0.68       422
weighted avg       0.74      0.75      0.73       422
```

e) KNearestNeighbour



```
KNN

Confusion Matrix
 [[257  25]
 [ 22 118]]
Accuracy
 0.8886255924170616
              precision    recall  f1-score   support

           0       0.92      0.91      0.92       282
           1       0.83      0.84      0.83       140

    accuracy                           0.89       422
   macro avg       0.87      0.88      0.88       422
weighted avg       0.89      0.89      0.89       422
```

Accuracy Table

ALGORITHM	ACCURACY
• Random Forest	90 %
• Logistic Regression	73 %
• XGBoost	88 %
• SVM	74 %
• KNN	89 %

V. CONCLUSION

In this study, various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms of which Random Forest gives highest accuracy of 90%. We have seen comparison of machine learning algorithm accuracies and also by comparing confusion Matrix in order to keep False Negative value as less as possible. Further we can extend the research by finding whether the non-diabetic person is likely to have diabetes in next few years or not.

VI. REFERENCE

[1] Mark A Atkinson, PhD, Prof, George S Eisenbarth, MD, Prof, and Aaron W Michels, MD ‘*Type 1 diabetes’.*

[2] Abdulfatai B.Olokoba,[1,*] Olusegun A. Obateru,[2] and Lateefat B. Olokoba[3] *‘Type 2 Diabetes Mellitus: A Review of Current Trends’.*

[3] Rabina , Er. Anshu Chopra ‘*DIABETES PREDICTION BY SUPERVISED AND UNSUPERVISED LEARNING WITH FEATURE SELECTION’.*

[4] Santi Wulan Purnami, Jasni Mohamad Zain and Abdullah Embong, “*Data mining techniques for medical diagnosis using a new smooth SVM*”, Communications in Computer and Information Science, Vol 88, Part 1, pp.15-27,2010.

[5] K. Rajesh and V. Sangeetha, “*Application of Data Mining Methods and Techniques for Diabetes Diagnosis*”, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[6] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, *“Diagnosis of Diabetes Using Classification Mining Techniques”,* International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[7] Abdullah, R.M., Tarik A., Rashid, S.M.A., Abstract, 2016. *“An Intelligent Approach for Diabetes Classification, Prediction and Description”.* Advances in Intelligent Systems and Computing.

[8] Humar Kahramanli and Novruz Allahverdi “*Design of a Hybrid System for the Diabetes and Heart Disease*”, Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[9] B.M. Patil, R.C. Joshi and Durga Toshniwal*, “Association Rule for Classification of Type-2 Diabetic Patients”*, ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[10] Nai-Arun, N., Sittidech, P., 2014. *“Ensemble Learning Model for Diabetes Classification”.* Advanced Materials Research 931 - 932,