JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A MACHINE LEARNING APPROACH FOR DIABETES PREDICTION

[1]G.Ravindra, [2]P.Mallikarjuna Rao M.E, Ph.D
[1]PG Student, Department of ECE, Andhra University College of Engineering, Visakhapatnam, A.P, India
[2]Professor, Department of ECE, Andhra University College of Engineering, Visakhapatnam, A.P, India

Abstract
Diabetes is a chronic metabolic disorder. Diabetes occurs if the blood glucose levels (BGL) are unduly high in the body. High blood glucose levels gradually lead to cardiovascular diseases, diabetic retinopathy, nephropathy, neuropathy, and foot damage. Early detection is essential to reveal dire effects amongst people, leading to timely medical care and lifestyle changes. Hence, with the boom in machine learning, we can assess whether a person has diabetes or not. In this paper, we use the Pima Indian dataset, which is originally from the "National Institute of Diabetes and Digestive and Kidney Diseases", to build a diabetes prediction model using machine learning. In this work, we will use machine learning classifiers such as Logistic Regression, k-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and Light Gradient Boosting Machine. The algorithm with the most precise result was used for prediction. With an accuracy of 89.86% Gradient Boosting classifier outperforms other classifiers.
Keywords: Diabetes, Logistic Regression, k-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, Light Gradient Boosting Machine

1. Introduction

Diabetes mellitus, commonly known as diabetes, is a group of metabolic disorders characterized by high blood sugar (hyperglycemia) over a long period. Symptoms include frequent urination, increased appetite, and increased thirst. If left untreated, diabetes can cause many health-related problems. Serious long-term complications include stroke, chronic kidney disease, cardiovascular disease, nerve damage, leg ulcers, and visual and cognitive impairment. Diabetes is caused either by the body's cells not responding properly to the insulin produced or by the pancreas not producing enough insulin. Insulin is a hormone that helps glucose from food get into cells to be used for energy.

2. Related Work

P.Sonar and K.JayaMalini developed a model based on categorization methods such as Decision Tree, ANN, Naive Bayes, and SVM algorithms. For Decision Tree, their models give precisions of 85%, for Naive Bayes 77%, and 77.3% for Support Vector Machine.

Malathy.S et al.developed a model that relies on categorization and classification methods like Naive Bayes, Support Vector Machine algorithm and ANN. The main result of this work will be spotting the most effective algorithm which is good at providing good and more accuracy when the classification of a person is

allotted. It's found that the neural network algorithm performs better in comparison to other algorithms for disease prediction.

C.Charitha et al. we used many Machine Learning models such as KNN, Logistic Regression, SVM, Random Forest, LightGBM, and XGBoost for train-test split like 60–40, 70–30, and 80–20 to predict Type-II diabetes mellitus. Among all models, the highest accuracy is obtained as 91.47% from the lightGBM model for the 80–20 train test split.

Dutta et al focused on obtaining an automated tool that will predict the diabetic tendency of a patient. The system proposed by this paper contains two ensemble classifiers- The voting ensemble classifier and the Stacking Ensemble classifier. Both of these methods exhibit better results when compared to other classifiers. The stacking ensemble classifier even performs better than the voting ensemble classifier with an accuracy of 79.87%.

Saxena et al. conducted experiments on the PIMA Indians diabetes dataset using Weka 3.9 and the accuracy achieved for multilayer perceptron is 77.60%, for decision trees is 76.07%, for K-nearest neighbor is 78.58%, and for the random forest is 79.8**%**, which is by far the best accuracy for random forest classifier.

3. Type of Classifiers

In this paper, we proposed a Machine Learning framework for the prediction of diabetes. This section highlights the classification algorithms we have used in our work and further in section 4 reports the workflow of the proposed methodology to predict diabetes.

a. Logistic Regression (LR)

LR is one of the most popular machine learning algorithms that fall under the supervised learning technique. It is used to predict a categorical dependent variable using a given set of independent variables. LR is used to solve classification problems. LR predicts the output of a categorical dependent variable. Therefore, the result must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or false, etc., but instead of an exact value like 0 and 1, it gives probability values that lie between 0 and 1. LR is very similar to linear regression except in how they use.

b. K-Nearest Neighbor (KNN)

K-Nearest Neighbor is one of the simplest machine learning algorithms based on supervised learning techniques. The KNN algorithm assumes similarity between the new case/data and the available cases and assigns the new case to the category most similar to the available categories. The KNN algorithm stores all available data and classifies a new data point based on similarity. This means that when new data appears, it can be easily classified into the appropriate category using the KNN algorithm. The K-NN algorithm can be used for both regression and classification, but it is mostly used for classification problems.

c. Decision Tree (DT)

A decision tree is a non-parametric supervised learning algorithm used for both classification and regression tasks. It has a hierarchical tree structure that consists of a root node, branches, internal nodes, and leaf nodes.

d. Support Vector Machines (SVM)

SVM is a set of supervised learning methods used for classification, regression, and outlier detection. This model is suitable for a small data set that has few outliers. The goal of the SVMs algorithm is to find a hyperplane in the N-dimensional space that clearly classifies the data points. The identified hyperplane separates the two spaces into different domains. Such a domain will consist of similar data types.

e. Random Forest (RF)

RF is a machine learning algorithm and belongs to the supervised learning model. RF, or random decision forests, is an aggregate learning method for classification, regression, and other tasks that works by generating a

large number of decision trees at training time. RF takes the majority vote prediction from all trees and finally predicts the output rather than relying on a single decision tree. Each node of the decision tree queries the data.

f. Extreme Gradient Boosting (XGB)

XGBoost, short for Extreme Gradient Boosting, is a scalable distributed gradient boosted decision tree (GBDT) machine learning library. It provides a parallel tree boosting and is the leading machine learning library for regression, classification, and classification problems.

It is important to understand XGBoost to first understand the machine learning concepts and algorithms XGBoost builds on supervised machine learning, decision trees, ensemble learning, and gradient boosting.

g. LightGBM

LGBM is a gradient boosting framework based on decision trees that increases model efficiency and reduces memory usage. It uses two new techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB), which meet the limitations of the histogram-based algorithm primarily used in all Gradient Boosting Decision Tree (GBDT) frameworks.

4. Proposed Methodology

The workflow in this research can be categorized into six main steps, which are

a. Data Collection

The 'PIMA dataset' originally came from the "National Institute of Diabetes and Digestive and Kidney Diseases". The motive of this data set is to diagnostically predict whether a person has diabetes or not based on certain diagnostic measurements which are included in the data set. Several constraints were placed on the selection of these instances from the larger database. Notably, all patients are women of at least 21 years of age of Pima Indian descent. The dataset consists of several medical predictor variables and a target variable, Outcome. Predictor variables include the following.

Table1. Attributes list

Attribute no.	Attribute
1	Number or times pregnant (NTP)
2	Plasma glucose concentration (PGC)
3	Diastolic blood pressure (mmHg) (DBP)
4	Triceps skin-fold thickness (mm) (TSFT)
5	2-h serum insulin (mu U/mL) (H2SI)
6	Body mass index (kg/m2) (BMI)
7	Diabetes pedigree function (DPF)
8	Age
9	Class 0 and 1 (Diagnosis of type 2 diabetes disease)

b. Data Visualization

Data visualization helps to better recognize the data by placing it in a visual form. In this step, the data distribution of the outcome variable was examined and visualized. In addition, we can observe the data imbalance.

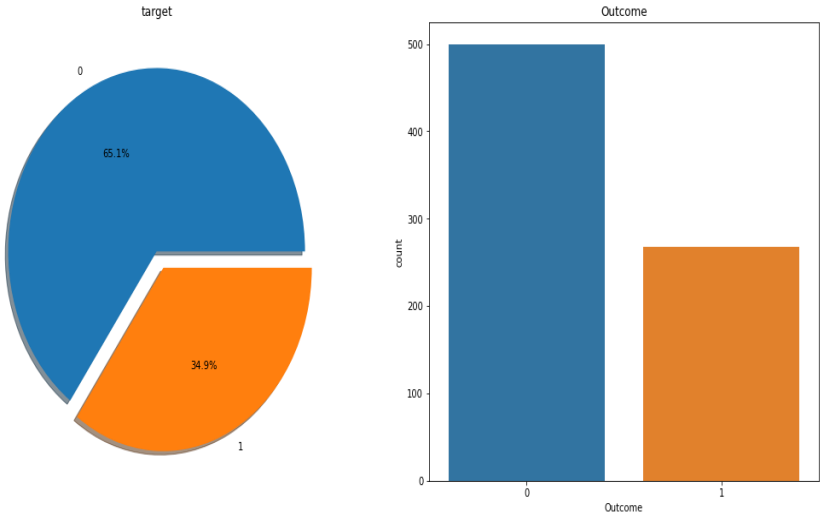

Fig1. The distribution of the output variable in the data was examined and visualize

c. Pre-processing

Prior to implementing classifiers to a data set, the data must be pre-processed and properly organized. The data should be maneuvered with care before processing. In this step, discrepant data is handled, outliers are discarded and data standardization is done to obtain more accurate and precise results. The processed data was used to create the model. This dataset contains missing values. The missing values are filled with the median values of each variable. We then normalized all values by scaling the data set.

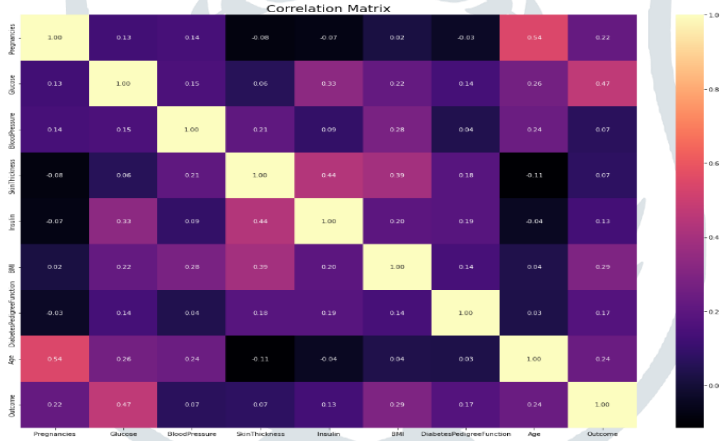

Fig2. The correlation matrix shows the strongest correlations with the outcome trait occurring for the glucose and BMI traits

d. Model Building

Consequently, after pre-processing the data, Machine Learning classifiers are implemented using the sci-kit-learn Python Toolkit. Scikit is a simple set of tools for processing and analyzing data. These toolkits are used in most jobs. Primarily, the dataset is split into training and test datasets using a function such as a test train partitioning for model selection. Due to the limited resource of the dataset, about 80% of the dataset is used for training purposes and the remaining 20% is used for testing by randomly selecting data. Various Machine Learning algorithms like logistic regression, K-Nearest Neighbor, Support Vector Machine, Random Forest, Decision Tree, XGBoost, and LightGBM are then used to predict diabetes. Machine Learning classifiers are acquainted due to their easy usage.

e. Comparison

In this step, we compare all our Machine Learning classification algorithms on the basis of their performance taking accuracy as a measure. After the assessment process, we found XGBoost as one of the better-performing classification algorithms. So, we check whether there are any over-fitting errors. For that, we need to optimize the hyper parameters. Grid search is used for our model tuning.

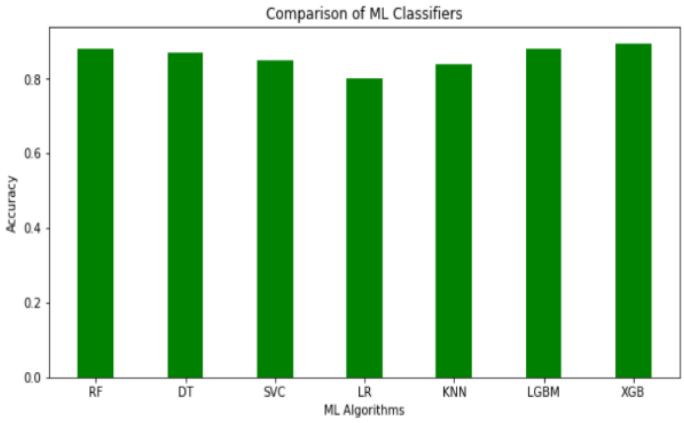

Fig3. Accuracy Comparison of ML algorithms

5. Results and Discussions

Machine learning technique is contemplated worthwhile in disease diagnosis. Early interpretations give patients the advantage of timely medical care. In this paper, we used the 'PIMA dataset' for our prediction model. All machine learning algorithms were forced upon the dataset.

Then it was trained and validated on the test dataset. Extreme Gradient Boosting performed superior to other machine learning algorithms with a cross-validation score of 89.86% in our model implementation. So, it is used for our model building. Further, the performance of our model is evaluated using various performance metrics such as confusion matrix, precision score, and recall score which is used to evaluate the machine learning classifiers' performance. We have achieved precision and recall score of 0.89 on our test dataset. We also observed that Insulin and glucose are the variables which impact for our model's accuracy.

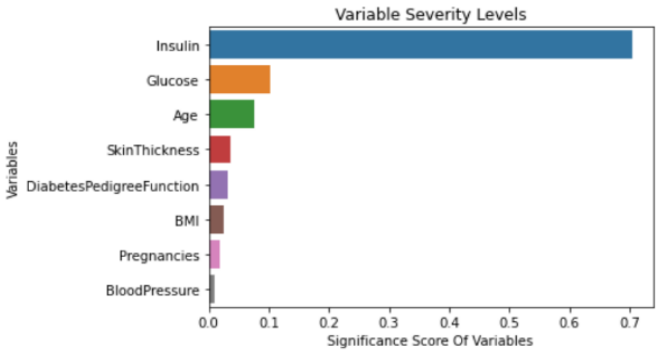

Fig4. Significance Score of Variables for XGBoost Classifier

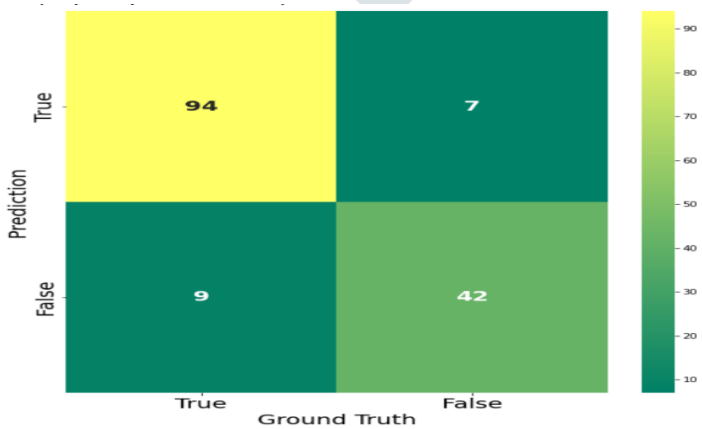

Fig5. Confusion Matrix for XGBoost Classifier

References

[1]. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes
[2]. https://en.wikipedia.org/wiki/Diabetes
[3]. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care, Vol. 31, No. 1, 2008, 55-60, 1935- 5548.
[4]. Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Al Marzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, Basant Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Journal of Healthcare Engineering, vol. 2022, Article ID 1684017, 10 pages, 2022.https://doi.org/10.1155/2022/1684017
[5]. P. Sonar and K. Jaya Malini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 367-371, DOI: 10.1109/ICCMC.2019.8819841.
[6]. V Anuja Kumari and R Chitra. 2013. Classification of diabetes disease using a support vector machine. International Journal of Engineering Research and Applications 3, 2(2013), 1797–1801.
[7]. C. Charitha, A. Devi Chaitrasree, P. C. Varma and C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022, pp. 1-5, DOI: 10.1109/ICCCI54379.2022.9740844.
[8]. Shawni Dutta, Bandyopadhyay Kumar Samir. Diabetes Prediction Using Ensemble Classifier. International Journal of Medical and Health Sciences Int J Med Health Sci. April2020, Vol-9; Issue-226
[9]. M. S, S. M, C. N. Vanitha and K. R. R, "Diabetes Disease Prediction Using Artificial Neural Network with Machine Learning Approaches," 2021 5th International Conference on Electronics, Communication, and Aerospace Technology (ICECA), 2021, pp. 1-5, DOI: 10.1109/ICECA52323.2021.9676094.
[10]. Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta, G. C. Sampada, "A Comprehensive Review of Various