ECS Transactions

The Electrochemical Society
Advancing solid state & electrochemical science & technology

Analyzing Three Predictive Algorithms for Diabetes Mellitus Against the Pima Indians Dataset

To cite this article: Aman Darolia and Rajender Singh Chhillar 2022 *ECS Trans.* **107** 2697

View the article online for updates and enhancements.

ECS Transactions, 107 (1) 2697-2704 (2022)
10.1149/10701.2697ecst ©The Electrochemical Society

Analyzing Three Predictive Algorithms for Diabetes Mellitus against the PIMA Indians Dataset

Aman[a], and Rajender Singh Chhillar[b]

[a,b] Department of Computer Science and Applications, Maharshi Dayanand University,
Rohtak, India
[a] sei@live.in, [b] chhillar02@gmail.com

Diabetes Mellitus is a metabolic disease in which the pancreas fails to produce enough insulin required for the processing of blood glucose. Most Medical Institutions analyze Electronic Health Records (EHRs) manually and then predict whether the patient is diabetic or not. The objective of this work is to classify Diabetes and Non-Diabetes patients using Predictive Algorithms/Techniques. These algorithms provide cost, time, and effort-effective solutions for the prognosis and diagnosis of diabetes mellitus. In this work, popular Algorithms like Artificial Neural Network (ANN), Random Forest (RF), and Logistic Regression (LR) have been used against the PIMA Indians Dataset and Analysis has been carried out on open-source software WEKA. In addition, this paper provided state-of-the-art by various researchers related to the said topic. This work concluded that LR outperforms other Algorithms with the accuracy of 77.10% but in the case of Area Under the Curve (0.83), both LR and RF perform equally well.

Introduction

Heath Sector advances at a lightning pace after the adaptation of Information Technologies. The use of Predictive algorithms in Data Mining is not limited solely to informatics (1-4). It has also become a part of the health sector and helps medical specialists to forecast the onsets of several conditions based on certain attributes such as age, blood pressure, level of glucose, insulin, etc.

Diabetes mellitus (DM) is one of the fastest-growing chronic diseases that require an efficient Clinical Predictive System for both diagnosis and prognosis. DM is a disorder in which the body cannot produce adequate insulin to control the blood glucose or the metabolic disease category in which a person has high blood sugar. A variety of serious health problems are more likely to occur in diabetes patients. India has approximately 77 million diabetes patients, making it the world's second-largest hospital, after China. One person in six (17%) with DM worldwide comes from India (5). The international diabetes federation predicts it may rise to 134 million by 2045 (6). There are generally two types of Diabetes. Both types are chronic diseases that impair the body's way of controlling blood sugar or glucose. Glucose is a fuel that feeds the cells of your body, but it requires a key for entering your cells. This is the secret to insulin. Insulin may not exist in people with Type 1 diabetes (T1D). Patients with Type 2 diabetes (T2D) fail to produce the required insulin that leads to a drastic increase in the amount of glucose in their body. Both forms of diabetes can lead to elevated blood sugar levels chronically. That increases the risk of diabetes complications.

In Section "Related Work", discusses the recent and related work by various researchers. Section "Methodology" outlines clear and concise Methodology for prediction of DM discuss in, which includes the description of the PIMA Indians dataset, Techniques/Algorithms, WEKA software, and performance metrics used. In Section "Results and Discussion", Results for the performance analysis shows against the performance metrics.

Related Work

In this section, an overview of the work reported in the field of diabetes prediction by various scholars/researchers are conducted. This section will also discuss the techniques/algorithms, and software used in each work.

Pethunachiyar (7) concluded that SVM algorithm with linear kernel produces a higher accuracy value (100%) than SVM with Radial kernel and Polynomial Kernel in DM prediction. In their model simulation, they collected the PIMA Indians dataset from the UCI repository having 7 feature variables and one target variable. Later Holdout Cross-validation was performed using 70:30 Split. The experiment was conducted by using the R programming language that contains pre-built libraries for SVM with various kernels. Finally, the researchers concluded that SVM with linear kernel fit better than another kernel on PIMA Indians Dataset and thus produces better accuracy.

In Reference (8), they created a new multi-agent system for diabetes diagnosis at an early stage. Three classifiers included in this model as ANN, SVM, and LR. In their proposed MAML (Multi-agent based on Machine Learning); they aggregated the best score of performance metrics by the majority voting approach. Research conducted on an open software WEKA software by using PIMA Indians dataset and provided a better result than a standalone algorithm/technique.

Alshamlan, Taleb, and Al Sahow (9) proposed a predictive model for gene datasets for diabetes. T2D is generally highly linked to genetics. Dataset GSE38642, GSE13760 collected from the Gene Expression Omnibus database and subsequently analyzed on anaconda software. During the processing phase, fisher score and chi2 feature selection were applied to get the most significant features out of the dataset. Later to test the effectiveness of these feature selection techniques, processed data passed to LR and SVM classifiers. Finally, the researchers concluded that the fisher score with LR produces better accuracy (90.23%) than other cases.

In Reference (10), authors conducted a comparative evaluation of various classifiers on the diabetes dataset collected from Kaggle. In this work, LR, KNN (K-nearest neighbor), SVM, Decision Tree (DT), Naïve Bayes (Gaussian, Multinominal, Bernoulli), and AdaBoost classifier used. Finally, research concluded SVM performs better in all aspects.

Geetha Devasena, Kingsy Grace, and Gopu (11), they proposed the PDD (Predictive Diabetes Diagnosis) Model that harnesses the power of K-Mean Clustering and RF. This model initially collected the dataset from the UCI repository and finally concluded that the model produces better accuracy than hierarchical and Bayesian network clustering.

Prabhu and Selvabharathi (12) proposed a DBN (Deep Belief Neural Network) model for the prediction of DM. This model divided into three phases. In the initial phase, PIMA Indians Dataset normalized in the range of 0 and 1 using the min-max normalization function, and then feature extraction done by using PCA (Principal Component Analysis). In the rest phase, DBN is trained and fine-tuned. The research concluded that DBN produces a better trade-off between Recall and Precision than NB, DT, LR, RF, and SVM.

In Reference (13), Researchers suggested using RF for the prediction of DM because it avoids over-fitting on a small dataset. RF harnesses the power of bagging that makes a suitable choice over other algorithms/techniques.

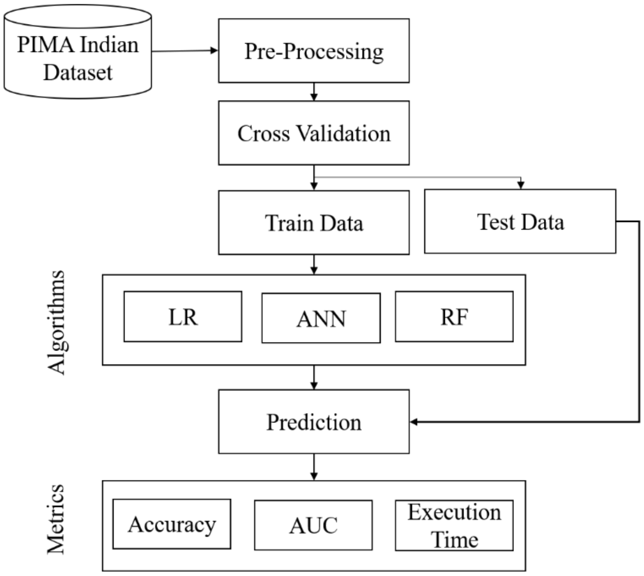


Figure 1. Methodology for Analyzing the Performance of Algorithms.

Methodology

For the Prediction of DM, a systematic and step-by-step process considered for proper analysis/evaluation of algorithms against the PIMA Indians Diabetes dataset (Figure 1). The methodology is very important for systematically understanding and analyzing the algorithms and helps to prevent ambiguity.

Dataset

This research work used the PIMA Indians Diabetes dataset that bundled with WEKA software's repository (14) and also available in the public domain on Kaggle (15). The data package aims to determine whether a patient is suffering from diabetes, based on certain diagnostic measures in the dataset. The collection of such instances from a broader database was subject to many constraints. Each Patient whose data was recorded is of 21 years at least. This dataset contains 9 attributes out of which one is the target variable and the other is feature variables as shown in TABLE I. Attribute "class" is the target variable. If this "YES" that means the patient is tested positive else tested negative for DM. This dataset contains 768 instances.

TABLE I. Structure of PIMA Indians Diabetes Dataset.

Attribute Name	Attribute Type	Data Type	Values
preg	Feature	Numerical	0-77
plas	Feature	Numerical	0-199
pres	Feature	Numerical	0-122
skin	Feature	Numerical	0-99
insu	Feature	Numerical	0-846
mass	Feature	Numerical	0-67
pedi	Feature	Numerical	0-2.45
age	Feature	Numerical	21-81
class	Target	Categorical	YES, NO

Selected Algorithms

The type of dataset, the number of instances, missing values, and the type of target attribute all influence algorithm selection. For this experiment, we have selected the following algorithms after an extensive literature study.

Random Forest. Random forest (i.e., RF) is a classifier and ensemble learning algorithm that uses DTs in a parallel manner and its basic working showed in Figure 2 (16). It founds under Trees > RandomForest in WEKA. Each DT receives input data for training and then the outcome sums up and the majority of the votes is forecast. DT usually suffers from the problem of overfitting; RF helps to avoid this.

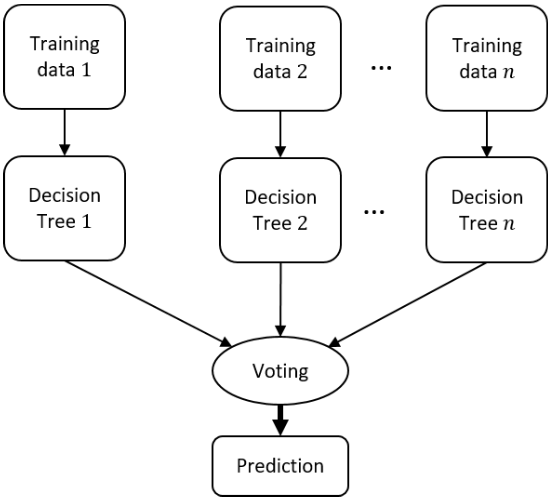


Figure 2. Working of Random Forest using Voting.

Artificial Neural Network. Artificial Neural Network (i.e., ANN) is a data mining algorithm that made up of three layers (Figure 3) (17) and each layer passes its output to the next layer. It is available under Functions > MutilayerPerceptron in WEKA. The nodes in the input layer pass their output to the next layer i.e. Hidden Layer. To Increase the performance of ANN, the number of nodes in the Hidden layer need to be optimized. Later the output of the Hidden layer passes to the output layer.

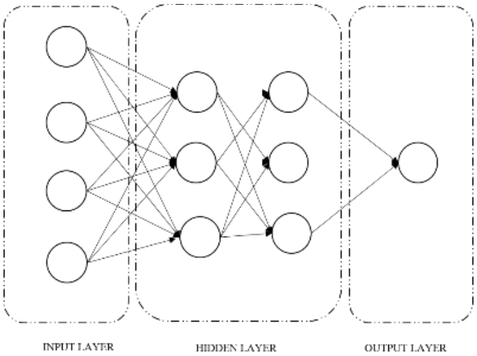


Figure 3. Simple Architectures of Artificial Neural Network.

Logistic Regression. Logistic Regression (i.e., LR) is a transformation of linear regression by the sigmoid function (Figure 4) (18). It is available in WEKA under Features > SimpleLogistics. Figure y represents linear regression and probability p symbolizes LR. The logistic function applies a sigmoid function to restrict the y value from a large scale to within the range (0, 1).

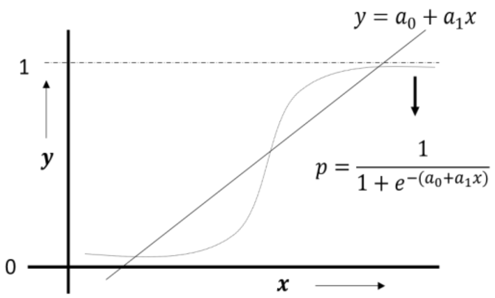
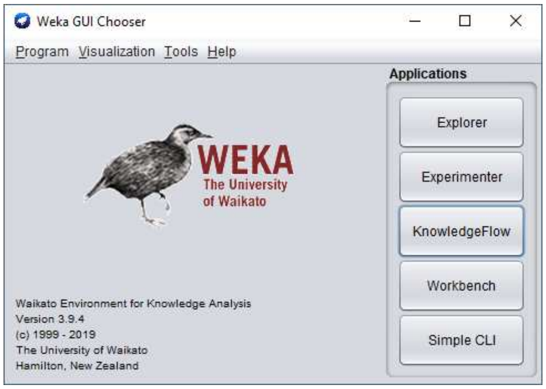


Figure 4. Graphical representation of LR.

Software Used

Waikato Environment for Knowledge Analysis (WEKA) is open-source software that provides the implementation of existing popular algorithms (Figure 5) (19). The framework enables different algorithms to be implemented with data extracts as well as the use of Java programming language to create algorithms from different applications. WEKA provides tools for preliminary data processing, regression, clustering, feature collection, association rules development, and visualization.



Waikato Environment for Knowledge Analysis
Version 3.9.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

Figure 5. The interface of Open-Source Software WEKA.

Performance Metrics Used

Performance of an algorithm measured using metrics. This will decide how better the algorithms in the classification of patients that have DM or not, and how much time it takes to predict the result. For evaluation, we have selected accuracy, execution time, and Area under Curve that described below.

Accuracy. Accuracy is the percentage of the number of correct predictions made by the classifier against the dataset mentioned in Equation 1. Informally, accuracy is a percentage of the results that our model has gotten correct.

$$\text{Accuracy} = \frac{\text{Number of correct Prediction}}{\text{Total Number of Prediction}} \qquad [1]$$

ROC AUC Score. The ROC curve is an assessment metric for the problem of binary classification. The TPR (True positive rate) is traced by the probabilities curve to different thresholds against the FPR (False positive rate) and helps to differentiate the signals from the noises effectively. The Area under Curve Region (AUC) is an indicator of the ability of the classifier to distinguish between different classes. The higher the AUC, the better the model's performance in distinctly positive from negative groups.

Execution Time. It is the time spent by the CPU to train and test the algorithm for a given dataset. Lesser the Execution Time, Lesser the Resources will use by the algorithm.

Results and Discussion

This research is intended to determine whether the patient is diabetic or not. After pre-processing of the data, missing instances replaced with the mean value. The trade-off between precision and recall will be reduced as a result. After handling Missing values, Data records are separated by 10-Fold cross-validation into training sets and test sets. Techniques/algorithms of data mining, namely RF, ANN, and LR, are used. Training data is feed to these classifiers and then Testing data is used to validate the result. Finally, performance of these models analyzed by WEKA software (see TABLE II).

TABLE II. Performance Evaluation of Algorithms.

Algorithms	Accuracy (in %)	Area Under Curve	Execution Time
Random Forest	76.10	0.83	0.24
Artificial Neural Network	74.75	0.80	0.69
Logistic Regression	77.10	0.83	0.08

In Figure 6, Accuracy of algorithms/techniques showed using a bar chart that indicates the LR algorithm outperforms others with an accuracy of 77.10%. In Figure 7, the Radar graph of algorithms concerning Execution time and Area Under Curve showed. This graph showed that Logistic Regression took less time to train against the dataset when compared with RF, and ANN. This also showed that while fitting the data (i.e., AUC) both RF and LR perform equally well with the value of 0.83.

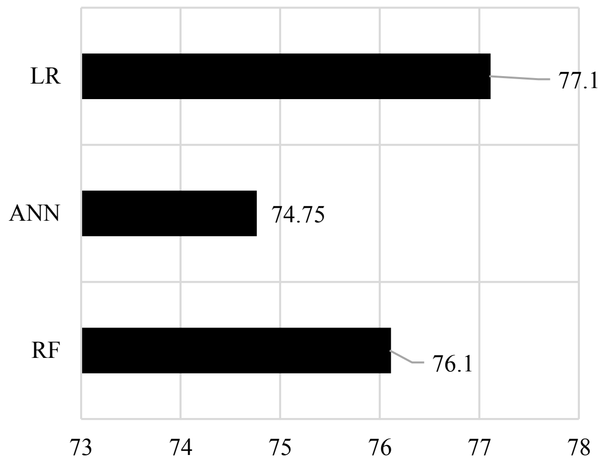


Figure 6. Accuracy (in %) of RF, ANN, and LR.

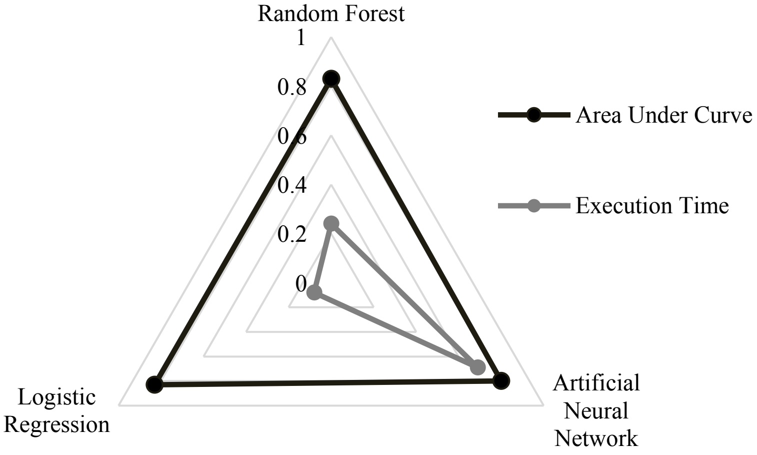


Figure 7. Radar chart showing of AUC, and Execution Time for the algorithms.

Conclusions

The objective of this work is to provide a brief overview and analyze three popular data mining techniques/algorithms namely ANN, RF, and LR for the prediction of DM. The experiment is carried out on WEKA software against the PIMA Indians dataset. In this work, selected algorithms/techniques analyzed and evaluated based on performance metrics like precision, ROC AUC score, and Execution Time. The LR's experimental test showed an accuracy of 77.10% which was higher than the ANN and RF. Limitation for an analysis of algorithm(s) largely depends on the dataset. Since this dataset contains only values related to females whose age less than or equals to 21. This work can improve by considering other datasets having both genders with varying ages. More conventional algorithms and hybrid approaches can consider for better analyzing the performance.

References

1. M. Islam, M. Hasan, X. Wang, H. Germack, and M. Noor-E-Alam, *Healthcare* (2018).
2. Aman and R. S. Chhillar, *SSRG Int. J. Eng. Trends Technol.*, **68**, 52–57 (2020).
3. P. Thareja and R. S. Chhillar, *Int. J. Eng. Trends Technol.*, **68**, 58–62 (2020).
4. Aman and R. S. Chhillar, *Int. J. Adv. Comput. Sci. Appl.*, **12**, 2021 (2021).
5. R. M. Anjana et al., *Lancet Diabetes Endocrinol.* (2017).
6. https://diabetesatlas.org/en/resources/.
7. G. A. Pethunachiyar, *Int. Conf. Comput. Commun. Informatics (ICCCI -2020)*, 22–25 (2020).
8. I. Chakour, Y. El Mourabit, C. Daoui, and M. Baslam, *6th Int. Conf. Optim. Appl. ICOA 2020 - Proc.* (2020).
9. H. Alshamlan, H. Bin Taleb, and A. Al Sahow, *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, 38–41 (2020).
10. R. Pradhan, M. Aggarwal, D. Maheshwari, A. Chaturvedi, and D. K. Sharma, *2020 Int. Conf. Power Electron. IoT Appl. Renew. Energy its Control. PARC 2020*, 133–139 (2020).
11. M. S. Geetha Devasena, R. Kingsy Grace, and G. Gopu, *2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020*, 22–25 (2020).
12. P. Prabhu and S. Selvabharathi, *2019 3rd Int. Conf. Imaging, Signal Process. Commun. ICISPC 2019*, 138–142 (2019).
13. K. Vijiyakumar, B. Lavanya, I. Nirmala, and S. Sofia Caroline, *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, 1–5 (2019).
14. https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff.
15. https://www.kaggle.com/uciml/pima-indians-diabetes-database.
16. L. Breiman, *Mach. Learn.* (2001).
17. O. I. Abiodun et al., *Heliyon* (2018).
18. L. J. Davis and K. P. Offord, in *Emerging Issues and Methods in Personality Assessment*, (2013).
19. https://www.cs.waikato.ac.nz/ml/weka/.