Performance Analytics of Classifiers: A Case Study with Diabetic Database

[1]Mohammad Asif Raibag, [2]Rashel Sarkar, [3]Nilakshi Deka, [4]Israfil Hussain

[1]Department of Computer Science and Engineering, Yenepoya Institute of Technology, Moodbidri

[2,3,4] Department of Computer Science and Engineering, The Assam Royal Global University, Assam

Abstract:- This particular work uses artificial intelligence (AI) simulations to evaluate the accuracy of four major classifiers in diabetes prediction. Numerous efforts have been successfully made to use specific approaches to predict the negative effects of diabetes and prevent it before the disease actually manifests itself. Diabetes is extremely difficult to categorise, despite the availability of various categorization algorithms. The primary goal of this study is to compare the effectiveness of the subsequent algorithms: Random Forest (RF), Decision Tree (DT), K Nearest Neighbor (KNN), and Logistic Regression (LR) for diabetes data classification. The Pima diabetic dataset, which makes use of nine features, is made available through the UCI repository that is used in this analysis. All four algorithms' results are assessed using a range of metrics, including recall, precision, accuracy, and F-measure. The acquired results demonstrate that, in comparison to other algorithms, the RF algorithm performs with the highest accuracy of 87.01%. Receiver Operating Characteristic (ROC) curves is used to effectively and meticulously verify these data.

Keywords**:** Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest.

1. Introduction

Diabetes is one among the foremost health tribulations everywhere the planet. A malady ensuing from defects in hormone secretion, hormone action, or both. On the report of World Health Organization (WHO) there are about 92.86 million cases of diabetes in adult population of India. Accompanied with this there's a bigger than before risk of dysfunction and failure of various organs, particularly the eyes, kidneys, nerves, heart and blood vessels. With such entanglements, averting and diagnosing the disease is needed to save human life from early detection. The significant point of this study is to analyze the performance of assorted classification models, which can envision the likelihood of affliction in patients with the best precision. The different AI methodologies help scientists to separate valuable data from colossal clinical datasets which will eventually improve the choice creating method and disease supervision. To attain aforementioned aim various risk factors related to this disease are explored early in prediction. For the purpose of study we have collected diabetic dataset having 9 attributes of 768 sample size. Based on these attributes, we construct expectation models utilizing different AI procedures to examine the presentation of various classifiers in prophesying diabetes.

In order to classify diabetic data for this research four well-known machine learning algorithms: KNN, DT, LR and RF are utilized. The determination of suitable strategies for exhibiting diabetic information dependent on these properties is a serious testing task. As a result, we evaluate the predicted results based on pertinent risk indicators and provide a tactical correlation of different AI techniques here. This is how the remaining portion of the work is structured. The relevant works are reviewed in Section (2). In Section. (3), we give a quick rundown of the algorithms. We provide our experimental results in Section (4). In Section (5), we finally draw this research's conclusion.

2. Background

Many intriguing work has been directed in the zone of diabetic diagnosing by utilizing AI procedures to extricate information from accessible clinical information. Loannis et al. [3] recommended that AI algorithms are essential to prognosticate diverse medical datasets including diabetes dataset. The paper proposed SVM, Logistic Regression and Naive Bayes utilizing ten overlay cross approval to predict diverse diabetes datasets. Bozkurt et al. [1] have utilized distinctive methods such as the Gini calculation from decision trees, Probabilistic neural networks (PNNs), Distributed time delay networks (DTDNs), Artificial Neural Networks (ANN) and compared the performance of these classifications on medical datasets.

For the short term prediction of the disease, Eleni et. al. [2] suggested a system utilizing gaussian process and support vector regression. Methods for grouping and categorization were applied to the long-term prediction of diabetes. They also helped to provide tailored clinical guidance when needed, in addition to ongoing patient monitoring. Reliable patient monitoring and prediction were made possible with the introduction of ML. A fuzzy expert system for diabetes decision assistance was proposed by Lee [4]. They were motivated by the idea that fuzzy ontology could address the shortcomings of conventional ontologies in managing imprecise and vague knowledge for certain real-world applications. They wanted to model knowledge of diabetes and introduce a new network of fuzzy experts to support the application of diabetes decision support. The result showed that the study provided an accuracy of 91.2%. The main goal of the diabetes prognosis method was to forecast whether an applicant would experience difficulties with the disease at a given age. The suggested system's architecture was developed using decision trees and the AI concept. The method's ability to accurately estimate the incidence of diabetes at a specific age is demonstrated by the impressive findings attained here [5]. Genetic programming (GP) was used to train and test a database for diabetes prediction using a diabetic dataset [6]. The findings generated by genetic programming were the most precise in comparison with various other techniques. This method works well for diabetes expectations that call for little to no effort.

Identifying and managing diabetes mellitus three data mining techniques—IB1, NB, and C4.5—were covered using medical datasets [15]. Using the feature selection strategy improved the performance of both NB and IB1. In order to efficiently classify diabetic data the following four supervised learning algorithms were combined: CART, Adaboost, Grading algorithm. Of these, the CART tree technique performed the best, with an accuracy of 78% when compared to the other learning algorithms [9]. PCA and the adaptive Neuro-Fuzzy Inference System (ANFIS) were suggested in [13] as a way to increase the accuracy of diabetic illness diagnosis when combined. Here, the diabetic disease dataset's dimensions are first reduced using PCA, then the adaptive neuro-fuzzy inference system classifier is used to diagnose the condition. The suggested method's attained precision was 89%. A feature selection method in this work that uses supervised model building to pinpoint the critical characteristics influencing diabetes management was proposed [14]. Three complementary classification techniques: Naive Bayes, IB1, and C4.5 were applied to the data once appropriate features were chosen in order to estimate how well the patient's condition was under control. The best prediction accuracy of 95% and sensitivity of 98% were attained in this approach.

A probabilistic neural network structure was employed in a diabetes investigation and a multilayer neural network structure that was trained using the LM method [16] is proposed. As the author suggests, the classification accuracy of MLNN with LM was 79%, which was relatively better than those produced by earlier studies. Here, the author [11] suggested utilizing a fuzzy verdict process to create a fuzzy expert system for diabetes. They conducted studies to offer a straightforward method for diagnosing diabetes. They used a fuzzy verdict process to create a fuzzy expert system framework. The investigation's assessment conclusion indicated that the precision of 85% was higher than that of other comparative technique. They intend in future to undertake the implication and operators for s-norms and t-norms for accuracy improvement of fuzzy verdict mechanism.

3. Proposed Work

The following algorithms are considered for our comparison analysis for prediction of diabetes.

a. Logistic Regression.

b. KNN Classifier.

c. Random Forest.

d. Decision Tree.

a. Logistic Regression.

LR is a statistical model that was developed by Joseph Berkson which in its central structure uses a logistic function to model a binary dependent variable with two possible values labeled as "0" and "1". The reason for popularity of this model is due to the log function that ranges between “0” and “1” hence the estimation of risk will always be between “0” and “1”. LR can be binomial, ordinal or multinomial. Binomial model deals with observing the outcome for a dependent variable that can have only two possible values "0" and "1". Multinomial LR deals with situations where the outcome can have three or more possible outcomes that are not ordered if ordered then it is ordinal LR model. This is how LR is fundamentally set up. A dataset including "N" points is presented to us. An outcome variable that is binary Yi, which is also referred to as a response variable, output variable, dependent variable, or class, is limited to assuming the values "1" (pass) or "0" (fail) for each point "i." A group of "m" input variables comprise each point. The goal of this LR model is to use the dataset to create a prediction model for the outcome variable. Numerous disciplines like healthcare, banking and the social sciences use LR.

b. KNN Classifier

KNN is a directed AI supervised machine learning algorithm which can be utilized for both grouping and regression prescient problems. Nonetheless, it is principally utilized for characterization prescient issues. KNN is characterized by the two significant properties for example it is a languid learning model since it doesn't have a specific preparing stage and uses all the information for preparing while grouping and the other is it's a non-parametric learning model since it doesn't expect anything about the key information.

The following are the important steps in implementing KNN classifier:

i) Selecting K value.

ii) Computing the distance between test data and each row of training data.

iii) Sorting according to distance value in ascending order.

iv) Next to choose from the sorted array, the top K rows.

v) Lastly, assign a class to the test point based on the most common class of those rows.

The benefits of KNN classifier are that the understanding and interpretation of the algorithm is very simple; it is valuable for nonlinear information in light of the fact that there is no supposition about information in this calculation and it has moderately high precision.

The following points go towards the KNN model; it is computationally a bit pricey algorithm due to the fact it stores all the training data, high reminiscence storage required as compared to different supervised learning algorithms, prediction is sluggish in case of huge ‘N’ and the main one is that it is very sensitive to both the data scale and irrelevant features. The functions of KNN are in banking system, to compute the credit ratings, speech recognition, handwriting detection, image/video recognition and so on.

c. Random Forest

RF is a supervised learning algorithm used for both classification and regression but used primarily for classification problems. This algorithm creates decision trees on data samples and then receives the prediction from each of them and eventually selects the excellent solution with the aid of capability of voting. It is an ensemble method which is better than a single decision tree since it decreases the over-fitting by averaging the outcome. It is executed as follows, at first irregular samples from a given dataset is chosen, at that point this algorithm will build a decision tree for each sample and will give expectation result from each decision tree, when forecast is done, casting a ballot will be performed for each anticipated outcome, ultimately the most casted a ballot expectation result will be chosen as the last forecast outcome.

The advantage of using this algorithm is, by combining the results of various decision trees, it eradicates the problem of over fitting, works perfect for large data items and they are robust and have very high precision. The issues with this algorithm are it is excessively unpredictable and tedious to fabricate and colossal computational assets are required for usage. The few areas where this algorithm is implemented are Banking, Medicine, Stock Market and E-commerce.

d. Decision Tree

DT belongs to the class of supervised learning algorithms and can be used for solving both regression and classification problems. It works by creating a training model which is used to predict target variables by learning decision rules inferred from training data. The DT solves the given problem by using tree representation. Each internal node of the tree corresponds to an attribute and each leaf node corresponds to a class label. The procedure is as listed, first we place the best attribute of the dataset at the root of the tree next we split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute this is repeated until leaf nodes in all the branches of the tree are found.

But the primary challenge in the decision tree is to identify which attributes we need to consider as the root node and at each level. This is attributes selection. We use one of the following methods for attribute selection i.e. the Information gain and the Gini index measure. One of the major concern in decision tree approach is it becomes too complex when there are many class labels. Its main use is in decision making applications. The below figure 3.1 illustrates the various components that make up the comparison framework. First, we define the problem and gather pertinent data, in this case the PIMA dataset from the UCI-ML library. After that, we preprocess the data in order to create the prediction model. Then, using the training dataset, we apply the several ML approaches that were previously mentioned. Lastly, the performance of the methods is evaluated to determine which classifier is most effective at predicting the disease.

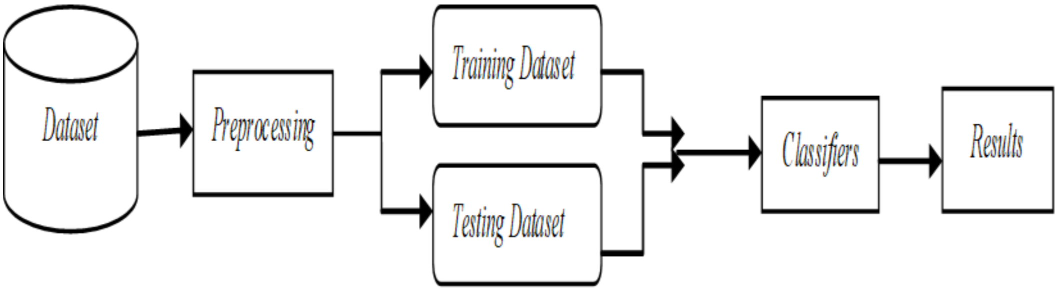


Figure 3.1

4. Experimental Setup and Results

In order to compare and validate the findings, the system is tested on the most commonly used Pima Indian diabetes dataset, which belongs to the National Institute of Diabetes and Digestive and Kidney Diseases. It is part of the UCI machine learning dataset available to researchers. This dataset contains 768 instances and 9 attributes. The input attributes are age, glucose concentration in blood 2 hours after having breakfast (Glucose), serum insulin in blood 2 hours after having breakfast (Insulin), body mass index (BMI), number of pregnancies (NP), triceps skin fold thickness (TSFT), diabetes pedigree function (DPF), and diastolic blood pressure (BP). The output of the system is either 0 or 1. 0 is interpreted as "no diabetes" and 1 is interpreted as "diabetes".

A. Evaluation Metric

In this section several performance metrics like confusion matrix, accuracy, specificity, sensitivity, precision and F-Measure is used to evaluate the performance of the four classifier's. If TP belongs to true positive rate and FP belongs to false positive rate then the first evaluation metric calculated was the accuracy, which is the fraction of true results (both true positives and true negatives) among the total number of cases examined. Following this, specificity, sensitivity, and precision were calculated. Specificity refers to the test's ability to correctly detect patients who do not have diabetes, whereas sensitivity relates to the test's ability to correctly detect patients who do have diabetes. Precision is the proportion of correct positive classifications (TP) from cases that are predicted to be positive. Finally, the F-Measure (also called F-Score) was computed this metric gives the harmonic mean of precision and sensitivity. For calculating these criteria we used the confusion matrix in our calculation process. The general view of confusion matrix is given below.

		Predicted Class	
		YES	*NO*
Actual Class	*YES*	*TP*	*FN*
	NO	*FP*	*TN*

Table4.1 Confusion Matrix

The actual class in a confusion matrix is the one that is provided in the data set, while the predicted class is the one that the classifier predicts. The amount of records that are accurately classified is indicated by True Positive (TP). The number of valid records that are accurately classified is shown by True Negative (TN). False Negative (FN) denotes erroneous classification of the records. False Positive (FP) denotes that the records have been mistakenly assigned a positive classification.

The equations of the performance metrics are given below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \qquad (4.1)$$

$$Specificity = \frac{TN}{FP + TN} * 100 \qquad (4.2)$$

$$Sensitivity = \frac{TP}{TP + FN} * 100 \qquad (4.3)$$

$$Precision = \frac{TP}{TP + FP} * 100 \qquad (4.4)$$

$$F - Measure = \frac{Precision{*}Recall}{Precision + Recall} *100 \qquad (4.5)$$

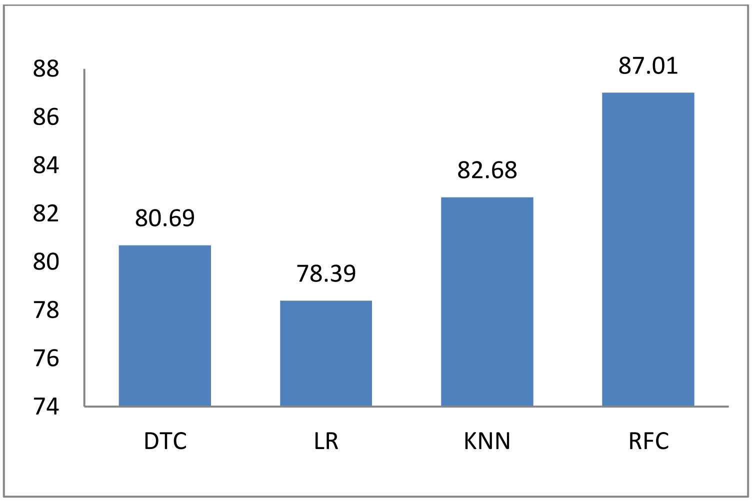
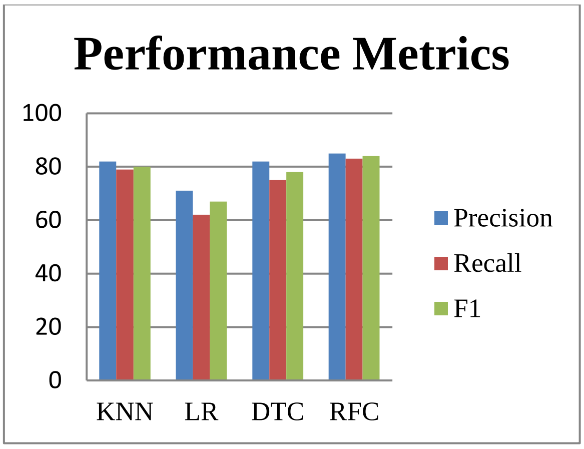


Fig 4.1: Comparison of Accuracy



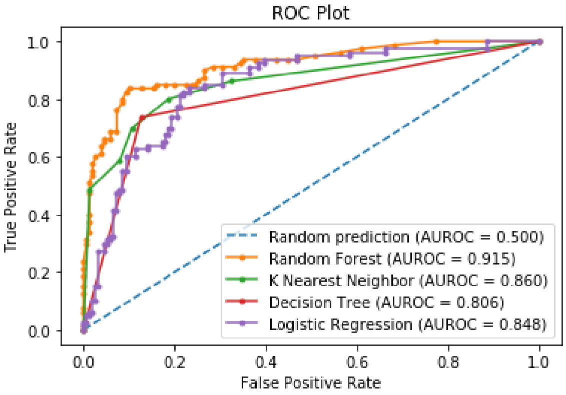


Fig 4.2: Comparison of Precision, Recall and F1

Fig.4.3 ROC curve of Classifiers

Conclusion

Diabetes is the most prevalent illness. Diabetes is a condition that should be detected early, prevented, controlled, and raised awareness of because it can lead to other health issues. In order to predict diabetes in the adult population, a systematic experimental investigation was carried out in this paper employing four well-known ML algorithms: DT (C4.5), KNN, RF and LR. The test findings demonstrate that, when it comes to classifying diabetic data, the C4.5 decision tree performs noticeably better than conventional ML methods. The outcomes of the investigation may help medical professionals avoid diabetes earlier and make better clinical judgements that will ultimately save lives. The proposed model may be used further to develop a diabetic mellitus control strategy.

References

[1] Eleni G. et.al. “Data Mining for Blood Glucose Prediction and Knowledge Discovery in Diabetic Patients: The METABO Diabetes Modeling and Management System”, 31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009.

[2] Pradhan et.al. “A Genetic Programming Approach for Detection of Diabetes”. International Journal of Computational Engineering Research, 91–94, 2012.

[3] Kavakiotis et.al. “Machine Learning and Data Mining methods in Diabetes Research”, Proceedings of Student/Faculty Research Day, CSIS, Pace University, 2005.

[4] Bozkurt, M. R. et.al. “Comparison of Different Methods for Determining Diabetes”. Turkish Journal of Electrical Engineering & Computer Sciences, 22(4), 1044-1055, 2014.

[5] Lee and Wang (2011) – “A fuzzy expert system for diabetes decision support application”. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Volume: 41, Issue: 1, 2011.

[6] Orabi, K. M. et. al. “Early Predictive System for Diabetes Mellitus Disease”, Industrial Conference on Data Mining, Springer. pp. 420–427, 2016.

[7] Yu et. al., “Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes and Pre-Diabetes”. BMC-Medical Informatics and Decision Making, 2010.

[8] Iyer et. al., “Diagnosis of Diabetes Using Classification Mining Techniques”. International Journal of Data Mining & Knowledge Management Process 5, 1–14. 2015.

[9] Yang, H., et. al., “Identification of Secretory Proteins in Mycobacterium Tuberculosis using Pseudo Amino Acid Composition”, Biomed. Res. Int. 2016:5413903, 2016.

[10] Yue, C., et. al., “An intelligent diagnosis to type 2 diabetes based on QPSO algorithm and WLS-SVM,” in Proceedings of the IEEE International Symposium on Intelligent Information Technology, 2008.

[11] Polat, K., and Gunes, S. “An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease”. Digit Signal Process. 17, 702–710, 2007.

[12] Alghamdi, M., et. al., “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project”. PLoS One 12:e0179805, 2017.

[13] M.W. Aslam, A.K. Nandi, “Detection of diabetes using genetic programming”, 18th European Signal Processing Conference, 2010.

[14] T. Daghistani and R. Alshamimar, “Diagnosis of Diabetes by Applying Data Mining Classification Techniques,” IJACSA, vol. 7, no. 7, 2016.

[15] Deepti Sisodiaa et.al., “Prediction of Diabetes using Classification Algorithms”, International Conference on Computational Intelligence and Data Science - ICCIDS 2018 ,Science Direct Procedia Computer Science 132 (2018) 1578–1585.

[16] Han Wu, “Type 2 diabetes mellitus prediction model based on data mining,” in Elsevier journal, 2017, 2352-9148.

[17] Y. Huang, P. et. al., “Feature selection and classification model construction on type 2 diabetic patients data”, Artificial Intelligence in Medicine 41 (3) 251–262, 2015.

[18] Sejdinovic, Dijana, et al. "Classification of Prediabetes and Type 2 Diabetes using Artificial Neural Network." Springer. CMBEBIH 2017.

[19] Nnamoko, Nonso Alex, et al. "Fuzzy Expert System for Type 2 Diabetes Mellitus (T2DM) Management Using Dual Inference Mechanism." AAAI Spring Symposium: Data Driven Wellness. 2013.

[20] Alan Siper et.al., “Machine Learning and Data Mining Methods in Diabetes Research”, Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 6th, 2005.