Original Article

IJCIH
Converging Healthcare & Technology

INTERNATIONAL JOURNAL OF CONVERGENCE IN HEALTHCARE

Published by
IJCIH & Pratyaksh Medicare LLP www.ijcih.com

Performance Evaluation of Various Classifiers for Diabetes Detection: A Comparative Approach

Bhavya Sharma

Student, Department of Computer Science & Engineering, Inderprasha Engineering College, APJ Abdul Kalam Technical University, Uttar Pradesh, India

Abstract

Various Supervised learning algorithms or techniques viz. Random Forest, Naïve Bayes Classifier, Logistic Regression (LR), K-Nearest Neighbour (KNN) algorithm, Support Vector Machines (SVM), etc are used for the purpose of data classification.. But the question is which of the classification technique accurately identifies this sensitive disorders like Diabetes. The accuracy, specificity and sensitivity, are some of the important performance evaluation parameters, which are required to be analysed for every machine learning algorithm. In the performed work, the various classification techniques viz. (NB) Naïve Bayes Classifier, (LR) Logistic Regression, (KNN) K-Nearest Neighbour algorithm, (SVM) Support Vector Machines, and (RF) Random Forest are compared on the basis of the accuracy, sensitivity and specificity as the performance evaluation parameters. The classifiers were exposed to the Pima Indian dataset for classification of diabetes, and their respective performance metrics Accuracy, Sensitivity, and Specificity were compared. It is found that on account of accuracy sensitivity and specificity the Random Forest performed the best on the Pima Indian Dataset for the Diabetes detection.

Keywords: *Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbor (KNN) algorithm, Support Vector Machines (SVM), Random Forest(RF), Classifier, PIMA Indian dataset.*

Introduction

The performed work relates to the comparison of the performance of various classifiers when subjected to the well-known dataset (Pima Indian) from UCI repository [1], it carries eight features viz. Pregnancy, Plasma Glucose Concentration, Diastolic Blood Pressure, Triceps Skin fold thickness, 2-hour serum insulin, Body Mass Index, Diabetes pedigree function, and Age. The dataset is exposed to various classifiers and the respective confusion matrix is determined. Thereafter, the comparison is performed on the grounds of various performance evaluation parameters viz. Accuracy, Sensitivity and Specificity But, prime focus is on the Accuracy of the classifier, as it is used to Classify/identify that the patient as diabetic or not. Alternatively, Accuracy is defined as the rate at which the investigated cases are correctly classified by a classifier. Based on the parameters derived from the confusion matrix of the respective classifiers the performance evaluation parameters for each classifier are evaluated, the respective formulation of the performance evaluation parameters under consideration are as follows:

Corresponding Author:
Bhavya Sharma
Student, Department of Computer Science & Engineering, Inderprasha Engineering College, APJ Abdul Kalam Technical University, Uttar Pradesh, India
e-mail: bhavyasharmacse7@gmail.com

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{TN + FP}$$

Where,

TP (True Positives): positive cases classified correctly

TN (True Negatives) negative cases classified correctly.

FP (False positives):negative cases classified wrongly.

FN (False negatives):positive cases classified wrongly.

In this paper the accuracy of the classifier is considered as the prime factor of consideration because Accuracy relates to proximity of measurement to be correct, whereas the Sensitivity relates to the quantification of positives being correctly identified e.g. correct identification of ill-people, and Specificity relates to the quantification of negatives being correctly identified e.g. correct identification of healthy people. None the less the Specificity and Sensitivity are also important performance measures for binary classification, especially while studying medical datasets[2]. Thus the performed work focused on the explicit coverage of these performance evaluation parameters for the evaluation of the classifiers viz. (NB) Naïve Bayes Classifier, (LR) Logistic Regression, (KNN) K-Nearest Neighbor algorithm, (SVM) Support Vector Machines, and (RF) Random Forest

Methodology

The performed work, firstly addressed the problem of handling the missing data values and data outliers[3], by using winsorization (by 2%) technique for handling outliers and class-wise mean for data imputation. After performing data pre-processing, the K-fold cross validation technique (with K=10) was used for validating the models. Where, the initial dataset was partitioned into 10 Partitions. Among the 10, single partition is used for testing purpose and the remaining 9 partitions are used for the purpose of training. This process is repeatedly performed 10 times, and various parameters like accuracy, sensitivity, specificity etc. are recorded in the confusion matrix of respective classifiers. Then average Accuracy, Sensitivity, and Specificity were determined as the final parameter for each classifier. Subsequently the comparison is performed on the grounds of various performance evaluation parameters i.e. Accuracy, Sensitivity and Specificity.

Results and Discussion

Table-1 consolidates the resulting parameters for the classifiers in terms of Accuracy, Sensitivity and Specificity after handling outliers & using Class wise mean as data imputation technique, the analysis of the data reveals that Random Forest performed the best classifier among the rest of the classifiers, for the classification of Diabetes patients, using PIMA Indian dataset.

Table-1: Resulting parameters for the classifiers in terms of Accuracy, Sensitivity and Specificity after handling outliers & using Class wise mean as data imputation technique

Classifier	Accuracy(%)	Sensitivity(%)	Specificity(%)
LR	85.28	78.83	89.23
KNN	84.89	78.22	89.17
NB	84.64	78.51	86.25
SVM	84.88	78.42	88.66
RF	88.01	79.69	92.2

However the results tabulated in Table-2 above are further verified by analyzing the polynomial trend line equations and the R^2 values for the curves of the performance evaluation parameters i.e. Accuracy, Sensitivity and Specificity.

Table-2: Polynomial Trend Line Equation and R^2 Values for the Curves Performance Evaluation Parameters of Various Classifiers

Performance Evaluation Parameter	Polynomial Trend Line Equation	R^2 Value
Accuracy (%)	$y = 0.537x^2 – 2.682x + 87.67$	$R^2 = 0.895$
Sensitivity (%)	$y = 0.241x^2 – 1.256x + 79.84$	$R^2 = 0.886$
Specificity (%)	$y = 0.895x^2 – 4.827x + 93.73$	$R^2 = 0.789$

The analysis of the Polynomial Trend Line Equation for the respective parameters explains that the observed curve is parabolic pointing upwards and the curve for each classifier is skewed towards Right i.e. towards the Random Forest Classifier, which is found true when we cross referred it with the data given in Table-2. Further, the R^2 value determined for the respective classifiers elaborates that 89.5% of data is in synchronization with the results of Accuracy of the classifiers, 88.6% of data is in synchronization with the results of Sensitivity of the classifiers and 78.9% of data is in synchronization with the results of Specificity of the classifiers. The observed results are verified with the Trend line equations and the Statistical Parameter R^2.

Conclusion

Based on the analysis of the data tabulated in Table-1 and Table-2, and the graphical representation of the curves of the performance evaluation parameters of the various classifiers, it is found that performance of the Random Forest Classifier supersedes the performance of the rest of the classifiers when it is applied over the PIMA Indian dataset for the diabetes patients.

Ethical Clearance: Taken

Source of Funding: Self

Conflict of Interest: Nil

References

1. UCI repository, Pima Indian Dataset.
2. Souad Larabi-Marie-Sainte, Linah Aburahmah, Rana Almohaini and Tanzila Saba, ”Current Techniques for Diabetes Prediction: Review and Case Study”, Appl. Sci. 2019, 9, 4604; doi:10.3390/app9214604 www.mdpi.com/journal/applsci, October 2019
3. Sofia Goel and Sudhansh Sharma,” Advanced Data Imputation Techniques for Predicting Type 2 Diabetes using Machine Learning, “International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, vol.9,issue 2, December 2019.