Comparing the Performance of Machine Learning Classifiers for Diabetes Detection and Feature Selection

ISBN: 978-1-943295-20-3

Bibaswan Basu
M. P. Sebastian
Indian Institute of Management
(bibaswanb14phd@iimk.ac.in)
(sebasmp@iimk.ac.in)

Diabetes is a global epidemic. Though many studies have differentiated diabetes and non-diabetic individuals, none of them is successful in identifying all the important predictors of diabetes. This research is an attempt to identify better the predictors of diabetes. The UCI ML repository's "PIMA Indian" dataset is used for our research. Both baseline and advanced classifiers were used in the experimentation. Using Artificial Neural Network (ANN), various baseline classifiers, such as, decision tree (DT), logistic regression (LR), naive Bayes (NB), k-nearest neighbour (KNN), support vector machine (SVM), and various ensemble classifiers, such as, random forest (RF), adaboost, and gradient boosting (GB) the authors have classified the diabetic and non-diabetic instances and hence, discovered the top five predictors of diabetes which include glucose levels in blood and diabetes pedigree.

Keywords: AdaBoost, Artificial Neural Network, Bagging, Boosting, Classification, Decision Tree, Diabetes Detection, Ensemble Classifiers, Important Feature Identification & Validation, Gradient Boosting, K-Nearest Neighbour, Logistic Regression, Machine Learning, Naive Bayes, Random Forest, Support Vector Machine.

1. Introduction

According to a WHO survey related to pregnant women (Rajput et al., 2013; Seshiah et al., 2008), 2-17.8% develop "gestational diabetes". Diabetes Mellitus is one of the most pressing issues in the scientific and medical domain due to the significant societal impact of the condition, which necessarily generates massive volumes of data. Certainly, when it comes to Diabetes Mellitus ML, and other data-driven Artificial Intelligence (AI) techniques have significant breakthroughs (Kumari et al., 2021).

Diabetes Mellitus is a complicated condition related to metabolism, characterized by chronic "hyperglycemia" caused by deficiencies in the production of insulin hormone (Holt & Hanley, 2021). Diabetes has been categorized into two types: type 1, i.e., insulin-dependent, and type 2, i.e., non-insulin-dependent. The dataset used in this research is made up of samples collected from a "Pima Indian" population (Blake, 1998). These are distinct type-2, having both, diabetic and non-diabetic, kinds of instances. According to some experts, the "Pima Indians of Arizona", USA, have the world's highest documented diabetes occurrence (Knowler et al., 1979). Furthermore, the prevalent Type-2 diabetes in their community is thought to start slowly and gradually. As a result, according to Holt and Hanley, the "standard diagnosis" approach, which is largely based on the "plasma glucose test", may be delayed by up to 10 years (Holt & Hanley, 2021).

In this paper, the authors applied 5 baseline classifiers: DT, KNN, LR, NB, and SVM; 1 advanced ML classifier: ANN; some ensemble classifiers: AdaBoost, Gradient boosting, and Random Forest. The performances of these classifiers in terms of Accuracy and ROC score have been compared. In terms of accuracy, decision tree, adaboost with decision tree, and random forest; ROC score, ANN is the best. The authors also tried to find the most contributing features with multiple classifiers, so the report could be validated. All three classifiers, DT, LR and NB report that Glucose level in blood is one of the top features. Also, BMI level, Age could be potential features. Finally, this paper is concluded with discussions and future research agenda.

2. Related Literature

In recent years, a remarkable quantity of research in the field of diabetes patient detection utilizing ML and Data Mining (DM) techniques. The ensemble approach was applied by the researchers, which combines numerous single models to get superior prediction results. (Vijayan & Ravikumar, 2014) used different DM techniques for Diabetes Mellitus. And in 2017, they discussed the significance of AdaBoost and ML-based Bagging approaches utilizing J48 as the foundation for the prediction of diabetes. It accurately distinguishes diabetic and non-diabetic people based on various diabetes-related risk factors. It was discovered that the AdaBoost outperforms Bagging and the J48 method (Fatima & Pasha, 2017). (Smith et al., 1988) created a Neural Network based ADAP technique for building an associative model using randomly picked data for training, with a 76% accuracy. (Quinlan, 2014) employed a C4.5 model, which showed very good performance with an accuracy of 71.1%. Diabetes (type 2) was diagnosed using NB, J48, radial basis function (RBF), and ANN. NB had an accuracy of 76.95% and surpassed J48 and RBF, which had accuracies of 76.52% and 74.34%, respectively (Nai-Arun & Moungmai, 2015). (Nai-Arun &

Moungmai, 2015) created a model that predicts the risk of diabetes. It compared the 4 classifiers, ANN, LR, NB, and DT. Bagging and Boosting both have been applied to improve performance. Random Forest was the winner. (Şahan et al., 2005) obtained an accuracy of 75.87% using a 10-fold CV procedure using a weighted artificial immune system (Soltani & Jafarian, 2016). (Anand & Shakti, 2015) used the CART model with 75% accuracy (Anand & Shakti, 2015; Şahan et al., 2005). (Rani & Jyothi, 2016) introduced various ensemble techniques for classification that employ the various baseline as well as advanced classifiers along with CV-based hyperparameter tuning and reported a 77% accuracy. (Li, 2014) suggested a weight-based search. (Bashir et al., 2014) developed an ensembled model of CART, ID3, and C4.5 with a 76.5% accuracy.

3. Methodology

This study has 2 major parts: A comparison of the performance of various classifiers and important feature detection and validation. The framework is shown in Figure 1.

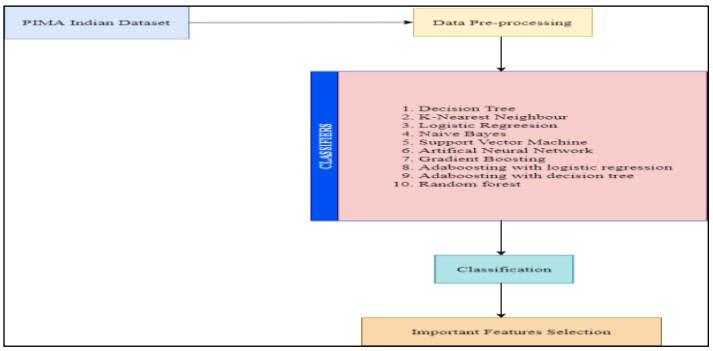

Figure 1 *Framework*

3.1. Description of Dataset

The chosen dataset consists of 8 feature columns and 1 label (Output) column with binary specification indicating whether the person is diabetic ("1") or non-diabetic ("0"). It comprises 768 instances out of which 500 are non-diabetic and the rest 268 are diabetic. The dataset contains eight feature columns, including no. of times of pregnancies, the concentration of glucose, plasma, blood pressure (mm Hg), fold thickness of triceps skin (mm), the quantity of insulin (μ U/ml), BMI (wt. in kg/(ht. in m2)), Diabetes Pedigree function, age of the patient, and the output column (0 or 1). The pair plots and correlations of the variables are shown in Figure 2 and Figure 3 respectively.

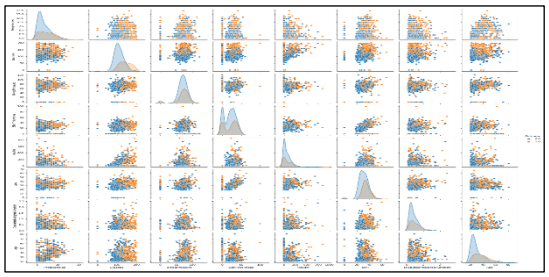

Figure 2 *Pair-Plots between all Variables*

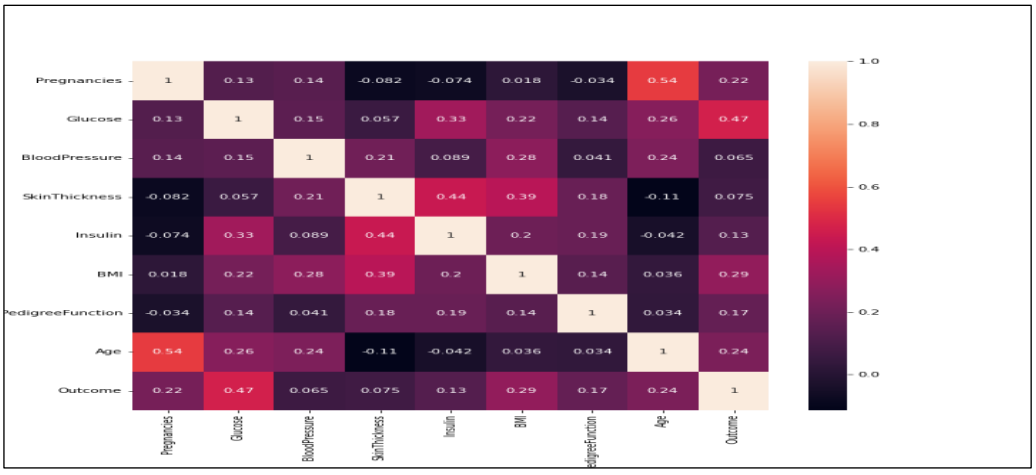

Figure 3 *Heatmap of Correlations between all Variables*

3.2. Pre-processing

3.2.1. Resampling

It is a very crucial step for AI-based data-driven research (Estabrooks et al., 2004). It is found that this dataset is imbalanced (Figure. 4). The authors up-sampled the diabetic instances to 500. So, the total number of instances becomes 1000.

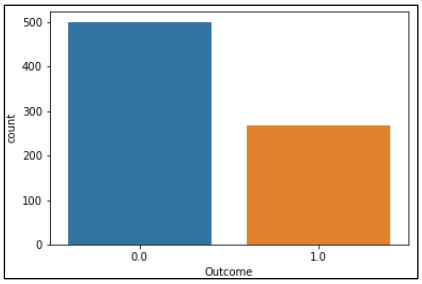

Figure. 4 *Count-Plot of Types of Instances*

3.2.2. Standardization

It is rescaled to a typical normal distribution with a zero mean and unit variance. Standardization reduces data distribution skewness. The standardization(R) is as follows

$$R(x) = \frac{x-\bar{x}}{\sigma} \quad (1)$$

Where symbols carry usual meanings. Many ML models, such as tree-based models, may not benefit from feature standardization (Hasan et al., 2020).

3.2.3. Train-Test Split

It is an important stage in ML. The recommended splits are (80%, 20%), (67%, 33%), (70% ,30%) etc. By executing a random split on the data, one can verify that the training and testing sets are representative samples of the whole dataset. For this work, the authors split the dataset into 80:20 ratio.

3.3. Classifiers Applied

The authors applied the following models

3.3.1. Artificial Neural Network

ANN models have the distinct benefit of being "universal approximators", capable of approximating a huge class of functions with high accuracy over a broad range of parameters. (Khashei et al., 2009) describes the mathematical equation

$$y_t = w_0 + \sum_{j=1}^{Q} w_j g\left(w_{0,j} + \sum_{i=1}^{P} w_{i,j} \cdot y_{t-i}\right) + e_t \quad (2)$$

where $w_{i,j}$(i=0, 1, 2, . . . , P; j=1, 2, . . . , Q) and w_j(j=0, 1, 2,. . . , Q) are model parameters often called connection weights; P is the number of input nodes, and Q is the number of hidden nodes and g is the activation function.

3.3.2. Decision Tree

Many prominent classifiers build class models using decision trees, which are then fed into other classifiers. To increase accuracy and minimize overfitting, these classifiers first build a DT and then prune the subtrees from it (Rastogi & Shim, 2000).

3.3.3. Naïve bayes

To classify data using probabilistic methods, a basic Bayes classifier applies Bayes' theorem to each row. Here, C represents the instance's class, and X the attribute values. Let c represent a class label and x represent an observed attribute value. According to the independence assumption, qualities X_1..., X_n are conditionally independent. The independence theory states such. The mathematical equation is as follows

$$\theta_{i,j} = P\left(X = x_i \middle| C = c_j\right) \quad (3)$$

For each class in the Bayes' theorem, the probability of each class may be calculated given a test case

$$\text{x:} p(C = c|X = x) = \frac{p(C=c)\cdot p(X = x|C = c)}{p(X=x)} \quad (4)$$

then guessing which class will occur. Because the event is a collection of attribute-value assignments, and the conditional independence assumption holds, the following equation holds

$$p(X = x|C = c) = \prod_i p(X_i = x_i|C = c) \quad (5)$$

For training and test data, this is a simple computation (Soria et al., 2011).

3.3.4. Support vector machine

With SVMs, the greatest generalization ability is achieved by maximizing the margin, and reducing training error, while avoiding overfitting (Boser et al., 1992; Cristianini & Shawe-Taylor, 2000; Vapnik, 1999).

3.3.5. Logistic regression

In statistical modelling, logistic regression (LR) links a category's probability to a set of explanatory factors (Liao et al., 2005).

$$z = a_0 + \sum_{i=1}^{n} a_i x_i \quad (6)$$

$$P(z) = \frac{e^z}{1+e^z} \quad (7)$$

Where x_i s are explanatory variables.

3.3.6. K-nearest neighbour

Since its inception, it has been possibly the most well-known supervised learning method in ML. When this learns, it simply keeps the full training set and gives a class to each query based on the majority label of its "k-nearest neighbours" in the training set. When k = 1, it becomes only the "Nearest Neighbour". KNN provides numerous key advantages, including simplicity, efficacy, intuitiveness, and competitive classification performance across a wide range of applications (Gou et al., 2012).

3.3.7. AdaBoost

AdaBoost prioritizes the wrongly classified instances by assigning more weights, hoping that these instances will be learnt by the classifier at the next stage. It starts with equal weights to all instances. The final prediction is calculated based on the weighted sum of the classifiers (Schapire, 1999)

$$F(X_i) = Sign(\sum_{k=1}^{K} \alpha_k f_k(X_i)) \quad (8)$$

Where, K is the total number of classifiers, $f_k(x_i)$ is the prediction of the weak classifier k for input feature vector X_i, and α_k is the weight assigned to the k^{th} classifier

$$\alpha_k = \frac{1}{2} \ln \frac{1-\varepsilon_k}{\varepsilon_k} \quad (9)$$

Where, ε_k is error rate of the k^{th} classifier's error rate.

3.3.8. Gradient boosting

Unlike AdaBoost, gradient boosting classifier learns from the residuals calculated in the previous stage, instead of learning from the misclassified instances. The equation of the prediction as follows (Zhao et al., 2016):

$$F(X_i) = (\sum_{k=1}^{K} \alpha_k F_k(X_i)) \quad (10)$$

Where, α_k is the learning rate or accuracy of the k^{th} classifier. F_k is the prediction at the k^{th} stage.

3.3.9. Random forest

It is one of the bagging techniques, which consists of ensemble of decision trees formed from bootstrap samples, sampling without replacement, and randomly selected subset of features without replacement. The decision trees generally expand without pruning (Speiser et al., 2019).

3.4. Performance Comparison

3.4.1. Baseline Classifiers

The performances of all the baseline classifiers in the resampled dataset are presented in Table I.

Table I *Performance Metrics of all Baseline Classifiers*

Model	Precision	Recall	F1-score	Accuracy	AUC-ROC score
ANN	74%	74%	73%	73%	**85%**
DT	**80%**	**80%**	**80%**	**80%**	80%
NB	**80%**	79%	79%	79%	79%

SVM	76%	76%	76%	77%	76%
LR	74%	74%	74%	74%	74%
KNN	70%	70%	70%	70%	70%

3.4.2. Ensemble classifiers

To decide the optimal hyperparameters, we applied the 5-fold cross-validation technique with the help of Grid Search CV package of Python. We applied logistic regression and decision tree as the two weak classifiers for AdaBoost classifiers. For logistic regression as a weak classifier, the optimal criteria come out to be maximum iteration as 250, penalty l2, learning rate 1, and number of estimators 75. For decision tree, the splitting criterion "entropy", maximum depth 8, maximum features the square root of the total no of features, learning rate 1, and number of estimators 75. For gradient boosting, the residual function friedman_mse, loss function deviance, learning rate 1, maximum depth 16, maximum features logarithmic value of the number of features, and number of estimators 75. The comparison is shown in the Table II.

Table II *Performance Metrics of all Ensemble Classifiers*

Model	Precision	Recall	F1-score	Accuracy	AUC-ROC score
AdaBoost with logistic regression	**81%**	79%	79%	79%	**79%**
AdaBoost with decision tree	80%	**80%**	**80%**	**80%**	75%
Gradient Boosting	76%	77%	74%	77%	66%
Random forest	80%	**80%**	**80%**	**80%**	75%

3.5. Identification of Important Features

The authors identified the top 5 important features and the absolute value of their weights using DT (Figure. 5), NB (Figure 6), and LR (Figure 7).

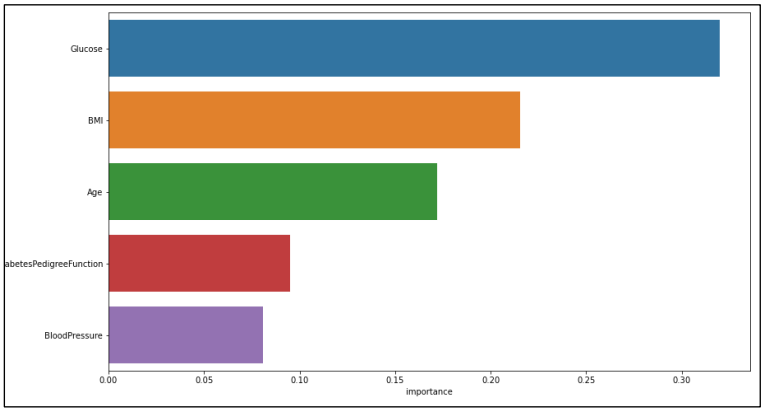

Figure 5 *Important Features Identified by DT*

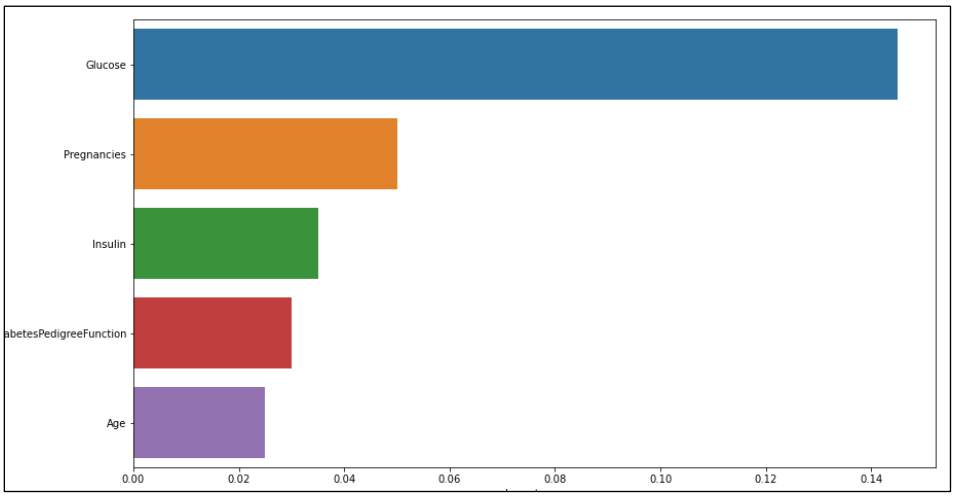

Figure 6 *Important Features Identified by NB*

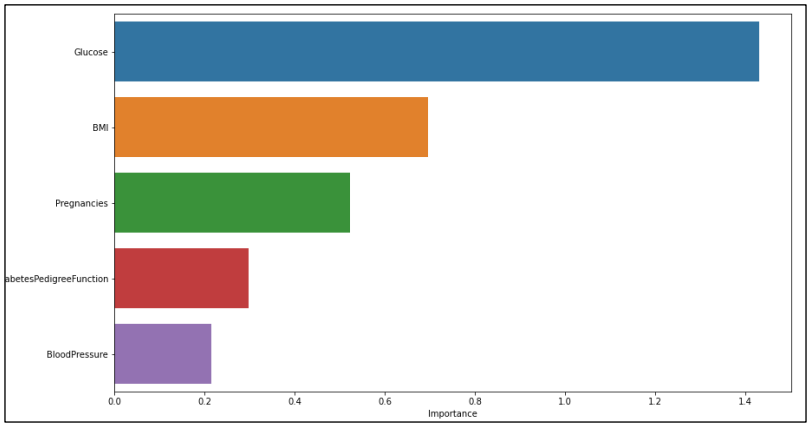

Figure 7 *Important Features Identified by LR*

4. Discussion and Conclusion

Among all the classifiers, decision tree, adaboost with decision tree, and random forest reported in highest accuracy score (80%) and ANN reported the highest ROC score (85%). KNN reported in lowest Accuracy score (70%) and gradient boosting reported lowest ROC score (66%).

Also, it is not sufficient to detect diabetes but to identify the main parameters which are the main reasons causes of diabetes. Having some checks and balances on these parameters can help to treat diabetes and also indicate which type of people could be prone to diabetes. Therefore, the authors identified the top 5 features using 3 classifiers, DT, NB, and LR. Each of the 3 classifiers identified glucose levels in blood and pedigree are the main features of diabetes. BMI levels have been identified by DT and LR. Age has been identified by DT and NB. Blood pressure has been identified by DT and LR. No. of times of pregnancies has been identified by NB and LR. It could be observed that the top features identified by these classifiers have a lot in common. These features could be used by the healthcare providers for early detection and treatment, thereby improving the longevity of the affected people.

5. References

1. Anand, A., & Shakti, D. (2015). Prediction of diabetes based on personal lifestyle indicators. *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 673–676.
2. Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014). An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. *2014 12th International Conference on Frontiers of Information Technology*, 226–231.
3. Blake, C. (1998). UCI repository of machine learning databases. *Http://Www. Ics. Uci. Edu/~ Mlearn/MLRepository. Html*.
4. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152.
5. Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
6. Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18–36.
7. Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1.
8. Gou, J., Du, L., Zhang, Y., & Xiong, T. (2012). A new distance-weighted k-nearest neighbor classifier. *J. Inf. Comput. Sci*, *9*(6), 1429–1436.
9. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516–76531.
10. Holt, R. I. G., & Hanley, N. A. (2021). *Essential endocrinology and diabetes*. John Wiley & Sons.
11. Khashei, M., Bijari, M., & Ardali, G. A. R. (2009). Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs). *Neurocomputing*, *72*(4–6), 956–967.
12. Knowler, W. C., Bennett, P. H., Bottazzo, G. F., & Doniach, D. (1979). Islet cell antibodies and diabetes mellitus in Pima Indians. *Diabetologia*, *17*(3), 161–164.
13. Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40–46.
14. Li, L. (2014). Diagnosis of diabetes using a weight-adjusted voting approach. *2014 IEEE International Conference on Bioinformatics and Bioengineering*, 320–324.
15. Liao, X., Xue, Y., & Carin, L. (2005). Logistic regression with an auxiliary data source. *Proceedings of the 22nd International Conference on Machine Learning*, 505–512.

16. Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, *69*, 132–142.
17. Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
18. Rajput, R., Yadav, Y., Nanda, S., & Rajput, M. (2013). Prevalence of gestational diabetes mellitus & associated risk factors at a tertiary care hospital in Haryana. *The Indian Journal of Medical Research*, *137*(4), 728.
19. Rani, A. S., & Jyothi, S. (2016). Performance analysis of classification algorithms under different datasets. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1584–1589.
20. Rastogi, R., & Shim, K. (2000). PUBLIC: A decision tree classifier that integrates building and pruning. *Data Mining and Knowledge Discovery*, *4*(4), 315–344.
21. Şahan, S., Polat, K., Kodaz, H., & Güneş, S. (2005). The medical applications of attribute weighted artificial immune system (AWAIS): diagnosis of heart and diabetes diseases. *International Conference on Artificial Immune Systems*, 456–468.
22. Schapire, R. E. (1999). A brief introduction to boosting. *Ijcai*, *99*, 1401–1406.
23. Seshiah, V., Balaji, V., Balaji, M. S., Paneerselvam, A., Arthi, T., Thamizharasi, M., & Datta, M. (2008). Prevalence of gestational diabetes mellitus in South India (Tamil Nadu): a community-based study. *JAPI*, *56*, 329–333.
24. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261.
25. Soltani, Z., & Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *International Journal of Advanced Computer Science and Applications*, *7*(6).
26. Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M., & Ellis, I. O. (2011). A 'non-parametric'version of the naive Bayes classifier. *Knowledge-Based Systems*, *24*(6), 775–784.
27. Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101.
28. Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.
29. Vijayan, V., & Ravikumar, A. (2014). Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. *International Journal of Computer Applications*, *95*(17).
30. Zhao, Y., Gong, L., Zhou, B., Huang, Y., & Liu, C. (2016). Detecting tomatoes in greenhouse scenes by combining AdaBoost classifier and colour analysis. *Biosystems Engineering*, *148*, 127–137.