IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

The Random Forest Algorithm and Logistic Regression for Classification with Application

Afiaa Raheem khudhair[1], and Nadia Ali Ayyed[2]

[1]college of Administration and Economics, University of Thi-Qar, Thi-Qar, Iraq
[2]college of Administration and Economics, University of Basra, Basra, Iraq

ARTICLE INFO

Article history:
Received xxxx
Revised xxxx,
Accepted xxxx,
Available online xxxx

Keywords:
Random forest algorithm, Logistic Regression, genetic algorithm, classification, heart disease

ABSTRACT

The increasing use of modern algorithms and their diverse applications in medical and social fields has raised a critical research question if can a genetic algorithm, specifically Random Forest, be effectively applied for classifying the response variable, and how does. added its integration with the traditional Logistic Regression model enhance its performance, in this paper explore the integration of Random Forest with Logistic Regression to classify heart disease data. The results demonstrate that this combined approach significantly improves classification accuracy and reduces error rates compared to the standalone algorithm. We use the real data from heart disease patients, represented by 13 independent variables and a binary response variable (0 indicating no disease, and 1 indicating disease), we provide a comprehensive analysis of the model's performance. The algorithm was implemented and evaluated using the R programming environment, yielding strong results that underscore the power and quality of the combined approach for handling massive data applications.

1. Introduction

Heart disease is a specific abnormal condition that includes heart-related illnesses that impact the bloodstream. It can range from heart rhythm issues to blood vessel disorders. Medical practitioners use a patient's medical history as well as tests like blood pressure, blood sugar, and cholesterol to diagnose heart disease. In addition, sophisticated medical assessments like electrocardiograms (ECGs), exercise stress tests, X-rays, echocardiograms, coronary angiography, radionuclide tests, MRI scans, and CT scans can help determine whether a patient has cardiac disease. Reducing the risk of heart disease and minimizing the rising number of deaths are largely dependent on the application of effective treatment in conjunction with heart disease diagnostics. n the field of early disease diagnosis, machine learning is becoming more and more significant. [1]. A flexible machine learning technique for classification and regression applications is Random Forest. It performs exceptionally well in high-dimensional environments by aggregating the predictions of several randomized decision trees. [2]. Both continuous and categorical data can be handled by the non-parametric approach, which is also resistant to outliers and overfitting. Measures of classification error and variable relevance are provided by Random Forest. It has been effectively used in a number of domains, such as cheminformatics, image categorization, and the prediction of a compound's biological activity. In addition to offering extra features like an integrated performance evaluation and a measure of compound similarity, the algorithm's performance is on par with other cutting-edge techniques. Random Forest is

* Corresponding author.E-**mail** address: afya-rahim@utq.edu.iq
https://doi.org/10.62933/f3g3c233

easily adaptable to different learning tasks and is especially helpful for large-scale challenges. A flexible machine learning approach, Random Forest (RF) can be used for both regression and classification problems. By capturing gene connections and prioritizing significant genes, it has demonstrated promise in pathway-based analysis of microarray data, surpassing single gene-based approaches. (Pang et al., 2006).[3] When it comes to drug sensitivity prediction, RF has outperformed other techniques based on differential gene expression, effectively forecasting drug responses in cancer cell lines (Riddick et al., 2011) [4]. The efficacy of RF extends to the classification of medical data, especially in the diagnosis of diabetes. The algorithm's ensemble approach, which integrates many decision trees by bootstrap aggregating (bagging), is responsible for its effectiveness across these varied applications. This strategy yields lower error rates than other supervised learning techniques. Random Forest is a supervised learning algorithm derived from decision tree analysis, employed bootstrap aggregating (bagging) to enhance accuracy and reduce variance. This method builds multiple decision trees until an optimal 'forest' or model, known as the random forest, is achieved. The growth of trees is conducted using data accurately sampled during the bagging process. [5,6]

Random Forest can provide the highest accuracy among supervised learning algorithms.[3].This research provides evidence that deep learning algorithms can be applied to predict material flow, inventory, and prices more accurately than previous shallow neural network models[7][8]. Numerous machine determine the predicted solution $\hat{y}_i$ based on a threshold:

$$\hat{y}_i = \begin{cases} 0 \; if \; \hat{\pi} < 0.5 \\ 1 \; if \; \hat{\pi} \geq 0.5 \end{cases} \qquad (1)$$

earning approaches mainly consisting of supervised and unsupervised techniques have been investigated in the prediction of heart disease in the literature. As the usage of machine learning in medicine becomes more widespread, it is claimed that certain data sets perform better in the early diagnosis of heart disease [1,9].

The paper content the following **Section 1** describes the Classification, Random Forest and The Logistic Regression Model, Section 2 Methodology Section 3 Results and discussion **Section 4** References.

The aim of the paper:

The objective of this paper aims to be seen as the most appropriate algorithm to give us the ideal classifier by heart attacks are usually caused due to blockages, partially or completely, of the heart's veins or arteries that constrict the flow of blood from or to the heart. We will be comparing the novel Random Forest with Logistic regression the modified and the logistic regression classic to find which of these can give us the best accuracy and confusion matrix and the misclassification error

2. Methodology

2.1. Classification

Classification is the process of splitting a dataset into several parts, mostly according to shared traits. It works with data from two or more classes according to the research variable and is a form of supervised learning. The focus of our research is binary classification. (infected = 1 and uninfected = 0). This means that if the estimated probability ($\hat{\pi}$) of a particular out

come is less than 0.5, the predicted value is 0. If it is greater than or equal to 0.5, the predicted value is 1.[1]

2.2. Random Forest:

The values of a random vector sampled independently and with the same distribution for every tree in the forest determine the values of each tree in a random forest, which is a mixture of tree predictors. As the number of trees in a forest increases, the generalization error for forests converges to a limit. The strength of each tree in the forest and the correlation between them determine the generalization error of a forest of tree classifiers. Each node is split using a random feature selection, producing error rates that are

comparable but more resilient to noise. Internal estimates track correlation, inaccuracy, and strength, and they are used to illustrate how the number of characteristics employed in the. These ideas are also applicable to regression[2]

2.3. The Logistic Regression Model:

Logistic regression is a powerful classification that is relatively simple and robust for differentiating between two classes. Classification analysis involves creating a decision rule based on data (a training dataset) that can automatically categorize new data into one of two or more categories.[9] In a binary case for this model , in which the categorical response has been value as 1/0, least squares regression would produce an estimate for that refers the estimated probability of the outcome value as 1 given X.

it lends to the binary outcome Y = 1 for disease heart and Y = 0 for not disease. [9]

The general formula for logistic regression

$$y_i = \pi(x_i) + \varepsilon_i \ , \ i=1,2,\ldots,n \qquad (2)$$

$$P(X_t; \beta_0, \beta_1 \ldots \beta_j) = p(y_i=1|x_{ij}; \beta_0, \beta_1 \ldots \beta_j) = \pi(x_i) = \frac{exp\,(\beta_0 + x_i^t \beta_j)}{1 + exp\,(\beta_0 + x_i^t \beta_j)} \qquad (3)$$

$$p(y_i=0|x_{ij}; \beta_0, \beta_1 \ldots \beta_j) = 1 - \pi(x_i) = \frac{exp\,(\beta_0 + x_i^t \beta_j)}{1 + exp\,(\beta_0 + x_i^t \beta_j)} \qquad (4)$$

2.4. The algorithm

The following is the random forests algorithm (for both regression and classification):

1. Using the original data, create ntree bootstrap samples.
2. Create an unpruned classification or regression tree for each bootstrap sample, but with the following change: at each node, randomly sample the predictors and select the best split from those variables instead of selecting the best split among all predictors. Assuming mtry = p, the number of predictors, bagging can be considered the special case of random forests. 3. Make predictions for new data by combining the ntree trees' predictions (i.e., average for regression, majority votes for classification). The following can be used to predict the error rate based on the training data:
1. Use the tree created with the bootstrap sample to predict the data that is not in the bootstrap sample (what Breiman refers to as "out-of-bag," or OOB, data) at each bootstrap iteration.
2. Add up the OOB forecasts. (To combine these forecasts, each data point would be out-of-bag around 36% of the time on average.) Determine the error rate and refer to it as the OOB estimate.As long as enough trees have been planted, we have found that the OOB estimate of error rate is rather accurate; otherwise, it may bias upward. [7,10]

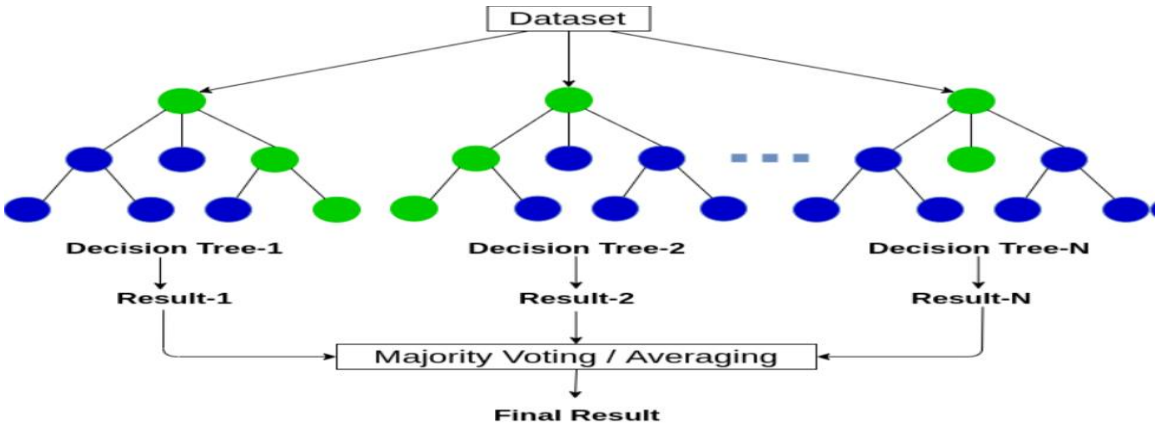

Figure (1) : Structure de l'algorithme random forest [13]

2.5. Data set

The data set used in this research consists of (13 independent variables) that will be explained with (the response variable y) that includes two categories: (0= without heart disease ,1= with heart disease). And n =303 observation.

The following is an explanation of the independent variables from which the model was built:

1. Age: Age of the patient (Integer, years)

2. Sex: Gender (Categorical; 1 = male, 0 = female)

3. CP: Chest pain type (Categorical), (1: typical angina, 2: atypical angina,3: non-anginal pain, 4: asymptomatic)

4.Trestbps: Resting blood pressure (Integer, mm Hg)

5.Chol: Serum cholesterol (Integer, mg/dl)

6.FBS: Fasting blood sugar > 120 mg/dl (Categorical; 1 = true, 0 = false)

7.Restecg: Resting electrocardiographic results (Categorical) ,(0: normal, 1: ST-T wave abnormality 2: probable or definite left ventricular hypertrophy).

8.Thalach: Maximum heart rate achieved (Integer)

9.Exang: Exercise induced angina (Categorical; 1 = yes, 0 = no)

10.Oldpeak: ST depression induced by exercise relative to rest (Integer)

11.Slope: Slope of the peak exercise ST segment (Categorical),1: upsloping,2: flat,3: down sloping

12.Ca: Number of major vessels colored by fluoroscopy (Integer; 0-3)

13.Thal: Thalassemia (Categorical),3: normal,6: fixed defect,7: reversible defect)

3. Results and discussion

We used algorithm random forest for classification heart data Random Forest is considered a genetic algorithm, meaning that it depends on the many decisions resulting from the previous steps, and then the final decision. we used ntree=500,

calculated the Confusion matrix and accuracy and sensitivity and misclassification rate for this model.

When combined with the logistic model to increase the quality and strength of the model, it is as follows: Generating predictions from the Random Forest model:

We use predict (model, data, type = "prob") in Script R to generate probabilistic predictions from the Random Forest model. The second column of results [,2] is used for the probabilities associated with class 1.

Among the standards and formulas that are always compared in topics related to the classification process is the confusion matrix, which includes binary matrix columns according to the dependent variable and y it has (TP (True Positive: Represents cases where the model accurately identifies a positive class or event.)

, TN (True Negative: indicates the number of correctly predicted negative instances) , FN (Fals Negative: It is Representing a failure to detect a positive class.) , FP (Fals Positive : It indicating a false detection of a positive class.)) .[13]

we notice that in table (1), the proposed method or what is called integrated with the logistic model gave very good results compared to the algorithm without the logistic model, as it gave a classification error rate of zero, complete classification accuracy, and sensitivity of one. Appeared Our compute classification accuracy by [14][15] :

1- Classification accuracy (CA) = ((TP + TN) / (TP + FP + FN + TN)) * 100%

2-Misclassification Rate = 1- CA

3-Sensetivity = TP / (TP + FN)

These results indicate the quality of the integration process between the genetic algorithm and the logistic model, a strong and wonderful process, and it led to no ratio error.

Table 1: Comparison criteria between methods

Method	Accuracy	Misclassification Rate	Sensitivity
Random forest algorithm	0.9967	0.003	1
The logistic with RF	1	0	1

Adding predictions as a new feature, and we add the predictions obtained from Random Forest as a new column in the original data. Then we use the logistic model training After training the logistic model using the original variables in addition to the new ones (Random Forest predictions).The predictions extracted from the logistic model are produced And we obtain the probabilistic predictions from the logistic model, then we convert them to binary predictions (0 or 1) based on a threshold of 0.5.Thus, we obtain a process of merging the results of the Random Forest model as an additional feature in the logistic model, which is beneficial and helps in improving the performance of the logistic model and enhances its quality and gives you a more accurate and better view of the relationship between the independent variables and the target variable.

To explain about table 2 it refers to the Logistic Regression model with the initial features + the new Random Forest (RF), can examine its performance based on the confusion matrix, The Logistic Regression model has No False Positives or False Negatives. the model made no incorrect classifications, it classified all instances perfectly. This is a highly favorable outcome, especially in cases where incorrect classification might lead to undesirable consequences. refers to High True Positives (TP) and True Negatives (TN) The model successfully identified 138 and True Positives (TP) and 165 True Negatives (TN). We have proven that (138+165 = 302 observation). These values indicate that the

Logistic Regression model correctly predicted the positive and negative classes at a consistent rate. This reflects its strength in distinguishing between the two classes with high accuracy. we notice incorporation of the new RF feature played a role in improving the performance the Logistic Regression model. The additional information introduced by this feature may have helped the model better capture the patterns in the data, resulting in no errors in classification.

Table 2: confusion matrix for methods

Method	TP	TN	FP	FN
Random forest algorithm	156	137	1	0
The logistic with RF	138	165	0	0

We generally recommend that future researchers modify and change classical models due to the modern scientific trend towards developing artificial intelligence algorithms, modifying them, or adding them to classical models, as well as applying them in various fields, including prediction, classification, clustering, and other fields, as well as changing the type of data to include health, education, environment, industry, and agriculture.

Reference

[1] M. Ozcan and S. Peker, “Healthcare Analytics A classification and regression tree algorithm for heart disease modeling and prediction,” *Healthc. Anal.*, vol. 3, no. December 2022, p. 100130, 2023, doi: 10.1016/j.health.2022.100130.

[2] G. Biau and E. Scornet, “A Random Forest Guided Tour To cite this version : HAL Id : hal-01221748 Introduction,” 2016.

[3] Pang, H., Zhao, H., & Tong, T. (2006). Random forest method for pathway-based analysis of microarray data. BMC Bioinformatics, 7, 49

[4] D. A. Hadi, D. Agustin, and N. Sirodj, "Metode Random Forest untuk Klasifikasi Penyakit Diabetes," pp. 428–435.

[5] Riddick, G., Song, H., Nakai, K., Li, Y., Imoto, S., Shimamura, T., & Tsuda, H. (2011). Predicting in vitro drug sensitivity using random forest. Bioinformatics, 27(17), 2200–2207.

[6] Chen, W., Liu, W., & Zhang, H. (2017). Random forest-based approach identifies differential gene expression in type 2 diabetes. Computational and Structural Biotechnology Journal, 15, 432–439.

[7] A. Ridwan, "Optimizing E-commerce Inventory to prevent Stock Outs using the Random Forest Algorithm Approach," vol. 4, no. April, pp. 107–120, 2024.

[8] Kahya, M. A. (2019). "Classification enhancement of breast cancer histopathological image using penalized logistic regression". Indonesian Journal of Electrical Eng ineering and Computer Science, 13(1), 405-410.)

[9] A. Raheem and S. M. Hussein, "Journal of Economics and Administrative Sciences (JEAS) Performance Classification for Lasso Weights with Penalized Logistic Regression for High-Dimensional Data," vol. 30, no. 139, pp. 149–160, 2024.

[10] Starbuck, C. (2023). Logistic regression. In The fundamentals of people analytics with applications in R (pp. 223–238). Springer. https://doi.org/10.1007/978-3-031-28674-2

[11] Srimaneekarn, N., Hayter, A., Liu, W., & Tantipoj, C. (2022). Binary Response Analysis Using Logistic Regression in Dentistry. International Journal of Dentistry, Volume 2022, Article ID 5358602, 7 pages. https://doi.org/10.1155/2022/5358602

[12] A. Liaw and M. Wiener, "Classification and Regression by randomForest," vol. 2, no. December, pp. 18–22, 2002.

[13] Algamal, Z. Y., & Lee, M. H. (2015). "Applying penalized binary logistic regression with correlation-based elastic net for variables selection". Journal of Modern Applied Statistical Methods, 14(1), 168-179.

[14] Algamal, Z. Y., & Lee, M. H. (2015). "High dimensional logistic regression model using adjusted elastic net penalty". Pakistan Journal of Statistics and Operation Research, 11(4), 667-676.

[15] Kalina, J. (2014). "Classification methods for high-dimensional genetic data". Biocybernetics and Biomedical Engineering, 34(1), 10-18 pp 12-14.