Compression Among Logistic Regression, Classification and Regression Tree and Random Forest in Predicting Kidney Disease

Dr. Elham A. Hussain
Administrative Technical College - Northern Technical University

Abstract

Chronic kidney disease still has a great burden with its negative impact on human life. Increasing incidence has become a global problem, and a major challenge for the health sector in terms of efforts and costs in their quest to save the lives of many patients[1]. A person cannot live without his kidneys for only 18 days or less, after which he needs dialysis. Here lies the importance of biostatistical methods for their ability to predict or classify patients in the early stages of this type of disease to which humans are exposed. These predictions help and enable the medical team to make the best-case management decision. These methods include Logistic Regression, CART (Classification and Regression Tree) and Random Forest (RF). The study was based on a sample of medical data consisting of 153 healthy and healthy patients and (11) independent variables to classify who has kidney disease or is healthy. The data was divided into two parts, 70% for the training and model development stage, and 30% for testing the strength of the models for the three methods. The criteria (Accuracy, sensitivity and specificity and Area under the curve ROC) were used to evaluate the three models in the testing phase. The results were: (86%, 89% and 86%) accuracy (100%, 91%, and 92%) sensitivity, and (67%, 86% and 78%) specificity and (70.6%, 73.4% and 81.1%) AUC of ROC, respectively. Based on the results of the research, the study concluded that the RF method achieved preference among the three methods, and the most important variables that have a clear effect on chronic kidney patients are albumin and urea.

Keywords:
Classification, Logistic Regression, CART, Random Forest, sensitivity, and specificity

الخلاصة

مازال مرض الكلى المزمن له عبئاً كبيراً بتأثيره السلبي على حياة الإنسان. إذا أصبح حدوثه المتنامي معضلة عالمية، ومُشكلاً تحديا كبيراً أمام القطاع الصحي من حيث الجهود والتكاليف في سعيهم بإنقاذ حياة الكثيرين من المرضى. فالإنسان لا يستطيع العيش بدون الكليتين سوى ١٨ عشرة يوما او اقل بعدها يحتاج الى الغسل الكلوي. وهنا تكمن اهمية الطرائق الاحصائية الحيوية بإمكانياتها التنبؤية أو التصنيف المرضي في المراحل المبكرة لهذا النوع من الأمراض التي يتعرض لها الإنسان[1] . وهذه التنبؤات تساعد وتمكن الفريق الطبي باتخاذ القرار الأفضل في إدارة الحالات. ومن هذه الطرائق : طريقة الانحدار اللوجستي Logistic Regression، طريقة Classification and Regression Tree) CART) وطريقة Random Forest (RF). تمت الدراسة بالاعتماد على عينة لبيانات طبية تتكون من ١٥٣ مريض وسليم و (١١) متغير مستقل لتصنيف الشخص اذا كان مصاب

او سليم من مرض الكلى .حيث تم تقسيم البيانات الى قسمين بنسبة ٧٠٪ لمرحلة التدريب وتطوير النماذج و ٣٠٪ لاختبار قوة النماذج للطرق الثلاثة.وقـــــــد تم الاعتماد على معاييـــــــر (Accuracy, sensitivity and specificity and Area under the curve ROC) لتقييم النماذج الثلاثـــــــــة في مرحلة الاختبار. وكانت النتائـــــــــج : صحة النموذج وتوفيقه (٨٦٪، ٨٩٪ و ٨٦%)٩١ والحساسية (١٠٠٪، ٩١٪ و٩٢٪) الدقة (٦٧٪، ٨٦٪ و ٧٨%) وAUC of ROC (٨١,١% ٧٣٤%, ٧٠,٦ %) ،على التوالي. واستناداً الى نتائج البحث توصلت الدراسة ان طريقة الغابة العشوائية RF حققت الافضلية من بين الطرق الثلاثة وإن أهم المتغيرات التي لها تأثير واضح على مرضى الكلى المزمنة هي الألبومين واليوريا.

الكلمات المفتاحية

التصنيف ، الانحدار اللوجستي ، شجرة التصنيف والانحدار،الغابة العشوائية ،الدقة والحساسية.

1.Introduction

Liver disease and patients with heart disease Interpretation of cooling signs and stages of chronic kidney disease (CKD), early identification of disease, and Collaboration between primary care and nephrologists. Because multiple terms have been applied to chronic kidney disease (CKD), for example, chronic renal insufficiency, chronic renal Illness, and chronic renal failure[1-2].

Chronic kidney disease (CKD) is a popular issue that is often unknown until the most finial phase. The happening of CKD is growing due to aging of the population and higher incidence of diseases, e.g., diabetes and blood pressure in the adult population[3]. In the developing countries, the lack of medical equipment and supplies for kidney patients caused the difficulty of detecting and determining the extent of the disease. In the absence of a central medical registry, the only data available is centre based. With awareness growing, more patients are recognizing with CKD; however, the majority requires immediate dialysis and the ethology of CKD in it still a massive speculative. Early diagnosis and suitable management have an essential roles in the prevention of progression of CKD to end-stage renal disease (ESRD)[4]. Classification/ prediction of kidney disease previously could support in correction , that is not always available. To avoid some conditions, management CKD needs to obtain a good understanding of a small number of predictors caused by kidney disease. The main aim of this study is to predict renal disease by analysing data from those patients and implementing three statistical classification methods to predict the disease, then choosing the method with the highest performance rate.

Author uses and compares three classification methods[5]: Logistics regression, classification and regression Tree CART and random forest. R version 3.3.3 Software was used to develop three models (LR, CART and RF). It is conducted by using a sample that includes 153 patients. Then the evaluation performance criteria e.g., Model Accuracy, Model Sensitivity, Model Specificity, and Area under curve (ROC). Finally, the study concluded that the outcomes of Random Forest methods is the best performance.

2.Material and Methods

Logistic regression (or logit regression)

Logistic regression is considered an essential statistical method that is concerned with analysing classified data, especially in the case of the response/ dependent variable, which is related to variables of the nominal or numerical type and consisting of two classifications(binary). Logistic regression works same as linear regression, but with a binomial independent variable (binary: 0,1)[6]

The goal of using logistic regression is to predict the existence of a certain characteristic or phenomenon, depending on the values of a variable or a group of other independent variables that have a relationship with the dependent variable. This type is characterized by the fact that the independent variables can be descriptive or quantitative. The other type is the multiple logistic regression model, which is an elongation of the binary logistic regression model when the dependent variable falls into more than two categories[6-7].

The importance of using logistic analysis has increased day after day because it is concerned with analysing the data with a double response, in which the dependent variable is usually binary, In the case of success, the response variable takes the value (1) and in the case of failure it takes the value (0), The logistic model is used to describe the relationship between the response variable (y) and one independent explanatory variable (x) or several (independent) explanatory variables[5]. The logistic regression model is based on the basic assumption that the dependent variable (y), the response variable that it is interested in studying, is a binary variable that follows the Bernoulli distribution and takes the value (1) with probability (p) and the value (0) with probability (1 - p) i.e., the occurrence of the response and its non-occurrence. The form can be written as follows[5]:

$$logit(p) = B_0 + B_1x_1 + B_2x_2 + \cdots + b_kx_k \qquad (1)$$

where:

p : refers to the probability of an event that could happen.

β_i : are the parameters that related with x_i independent variables (predictors)[8] .

There are many methods for estimating logistic regression parameters, such as least squares method, strong estimation, Bayesian estimation, maximum likelihood estimation, In this research maximum likelihood is used to estimate the parameters of logistic regression model[5] as following :

$$l(\beta_0,\beta) = \sum_{i=1}^{n} y_i logp(x_i) + (1 - y_i)\log 1 - p(x_i)$$

$$= \sum_{i=1}^{n} log1 - p(x_i) + \sum_{i=1}^{n} y_i log\frac{p(x_i)}{1-p(x_i)}$$

$$= \sum_{i=1}^{n} log1 - p(x_i) + \sum_{i=1}^{n} y_i(\beta_0 + x_i\beta)$$

$$= \sum_{i=1}^{n} -log1 - e^{(\beta_0 + x_i\beta)} + \sum_{i=1}^{n} y_i(\beta_0 + x_i\beta) \qquad (2)$$

Now , to get the parameters , we'd differentiate the log likelihood with respect to the parameters β_j , j= 0,1,2, … (equation (2)) as the following :

$$\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{1}{1+e^{(\beta_0 + x_i\beta)}} x_{ij} + \sum_{i=1}^{n} y_i x_{ij}$$

$$= \sum_{i=1}^{n} (y_i - p(x_i; \beta_0, \beta))\, x_{ij} \qquad (3)$$

3. Decision Trees

Decision Tree is a visual method based on explanatory data to determine the course of the decision-making process, by placing all possible possibilities and their outcome. A decision tree consists of many nodes that form the root of the tree, meaning that it is a vector tree with a node called root that has no incoming edges. All other nodes have one incoming edge. Other nodes are called an inner node or test node. All other nodes are called leaves. In a decision tree, each inner node divides into two or more subspaces according to a certain discrete function of the values of the input attributes. In common case, each test takes one attribute which divided according to the value of the attribute. In a numerical attribute, this is referred to the range. There are two types of tree decision. The first one is a classification when the target variable is discrete, and the second type is known as a decision tree regression if a target variable is continuous. The term Classification and Regression Tree (CART) is used to refer to both above procedures. The CART decision tree is a splitting process as binary recursive to processing continuous and nominal attributes as targets and predictors [9,10]. The algorithm of decision tree is called classification and regression tree which is symbolized as Ctree

3.Classification and Regression Tree (CART)

Classification and regression trees are machine learning methods for developing prediction models from data[11]. It is symbolized as CART , it is a nonparametric method that can select from among many variables those and their interactions that are essential in identifying the outcome predictor to be explained[12]. So as to use CART, it is needed to know the number of classes a prior. For building decision trees, CART uses a learning sample, a set of historical data with pre-appointed classes for all observations[11].

CART are prediction models developed by recursively segmenting a data set and fitting a simple model to each partition[13]. The first one is called a root node which divides into two classes, each node can be divided into two terminal nodes and, in turn, each of these terminal nodes can be spilled into additional nodes until it reaches the last end of a branch called a leaf node. Each resulting node is assigned a predicted class and the resulting subgroups should be more homogeneous in terms of the outcome variable. As a result, the segmenting can be represented graphically as a decision tree[13]. The sutable selection is to evaluate the ability of each attribute to create “pure” segmentations, that is, partitions in which each branch is very homogeneous with respect to the class distribution of its examples [6]. The ability of each attribute to create “pure” partitions that must be homogeneous with each other. To measure an impurity of the vector:

$$u = (u_1 , \dots , u_d)$$

We must account how many nodes in each class, into a non-negative scalar. It can be say, one of the most classical impurity measures is the Gini impurity, iGini , which measures the probability of misclassification when an object is determined to class i with probability $\frac{u_i}{\|u\|}$. It is defined as:

$$i_{Gini}(\mathbf{u}) = \sum_{i=1}^{d} \frac{u_i}{\|\mathbf{u}\|_1} \left(1 - \frac{u_i}{\|\mathbf{u}\|_1}\right) . \tag{4}$$

The Gini impurity is used in the CART package in R soft wear for classification and regression decision trees. The Gini impurity is[14] :

$$I_{Gini} = 1 - \sum_{i=1}^{j} p_i^2 \qquad (5)$$

$$I_{Gini} = 1 - (\textit{the probability of target “No”})^2 - (\textit{the probability of target “Y es”})^2$$

Where :
P : the probability value of the target (yes or no) .
The method which is used to get p_i is called (rpart)

4. Random Forest

Random Forests were presented by Leo Breiman in 2001; and Amaratunga .D et al, 2008[15-16]. Breiman defined Random Forest as a set of trees, each tree has a random vector of values which independent to other one and all of trees have the same distribution . CART methodology is used to expand the tree in size without prune. Denote this way by Random Forest RF.As a classification Breiman proposed growing the trees until the terminal nodes were pure or as a regression until there were less than a predetermined number of data points in each terminal node [15]. It can be used if the response variable categorical or continuous, if the response variable categorical ,it is called “classification”, but if the response variable continuous then it is called “regression” .Likewise, the predictor variables can be either categorical or continuous. The algorithm of random forest is called supervised learning. The "forest" it builds is a set of decision trees which are using “bagging” method as a training [17]. Rf structures multiple decision trees and combines the trees together to obtain a more precise and steady prediction. One of the big characteristics of RF is that it can be applied for both classification and regression maters, which form the most of current machine learning systems.

Theoretically, RF adds additional randomness to the model during growing the trees. Instead of looking for the most important predictors while dividing a node, it searches for the best predictors among a random subgroup of predictors. This results in a wide diversity that generally results in a better model[18]. The classifier is a set of algorithms combined to create ensembled algorithms. The idea was that one single algorithm has lower performance, less robust and lower accuracies such as logistic regression and decision tree. Therefore, the ensembled algorithms confer random forest to high performance, robust and high accuracy. This algorithm works by making the prediction of each tree and the outcome is based upon the majority votes which gives a highly accurate prediction [18]. This algorithm is called Boost classification . The formula of RF is[17]:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \qquad (6)$$

Where :

- RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
- normfi sub(ij)= the normalized feature importance for i in tree j
- T = total number of trees

5. Methodology

Regardless of the statistical methodology being used, we need data for each patient. Retrospective data set was used that included 153 patients[19]. The dataset contains data of 153 cases and 11 attributes for diagnosis of chronic kidney disease. The study was based on a sample of medical data consisting of 153 healthy and healthy patients and (11) independent variables to classify who has kidney disease or is healthy. The predictors are: Gender , Age , Smoking and nine predictors have a relationship with measures of blood tests[19]. See Table (1).

Table (1) : The predictors

Predictors	All subjects n = 153 Mean (SD) or n (%)	95% CI	
		Lower	Upper
Gender: **Male** **Female**	83 (54 %) 70 (46 %)	0.460 0.376	0.623 0.539
Age	42.752 (17.391)	39.97	45.53
Smoking: **Male** **Female**	61 (40 %) 92 (60 %)	0.320 0.519	0.480 0.679
Urea	58.653 (53.023)	50.19	67.13
Creatinine	2.011 (2.326)	1.64	2.38
Calcium	8.045 (1.506)	7.80	8.28
Phosphorus	4.404 (1.418)	4.178	4.63
Alka. Phosphates	245.893(125.068)	225.92	265.9
Glucose	122.468 (52.330)	114.11	130.83
Albumin	3.988 (.838)	3.85	4.12
Total Bilirubin	.800 (1.137)	.62	.98

As was previously indicated, the study sample consisted of 153 patients and 11 independent variables. Where it was divided into two separate parts randomly with the same number of predictors variables. The first part of the data was allocated for training and its percentage was 70%, while the second part was devoted to testing the three methods and its percentage was 30%, respectively. That is why the process of building predictive models for the three methods based on training data was called the training phase. While the use of test data is called the test phase, predictive models[20-21].

The test phase aimed to compare the performance of the three methods in classifying those with and without chronic kidney disease. The evaluation process was carried out based on a set of statistical criteria such as evaluation accuracy, specificity, and sensitivity. The classification value appears among patients with symptoms of the disease. The area under the ROC was performed to evaluate each method. The value of ROC shows the power of the model to classify between those patients who have a higher probability of CKD dangers[21 22].

This was undertaken applying the : library(randomForest),library(tidyverse),library(caret),library(dbplyr),library(readxl),library(pROC),library(party)packages in the R.

6. Results

From the table (2), We see that the sensitivity for the CART method (0.91) is less than from the Logit Regression method (1.00), while it is almost very close to FR method (0.92), but the other criteria, we see that the RF and LR methods are more closely related in criteria values than the CART method, however, LR criteria are the best of the three methods.

Table (2): The results of the methods

Methods Criteria	Logistic Regression	CART	RF
Sensitivity	1.00	0.91	0.92
Specificity	0.67	0.86	0.78
Accuracy	0.86	0.89	0.86
P-Value	0.006	0.00002	0.0054
95% CI	(0.664,0.97)	(0.817, 0.927)	(0.64,0.97)

From the table (2), We see that the sensitivity for the CART method (0.91) is less than from the Logit Regression method (1.00), while it is almost very close to FR method (0.92), but the other criteria, we see that the RF and LR methods are more closely related in criteria values than the CART method, however, RF criteria are the best of the three methods.

7. Discussions & Conclusions

In this research three classification methods are used to CKD data set, to develop a predictive model and determine the variables that have an essential effect on explaining kidney disease. Three methods are used when developing a model including stepwise logistic regression, CART and RF methods. The performance of the three methods compares approvingly with the predictive models of kidney disease in the new literature[20]. Based on the literature that reported that the traditional method achieves unwell whether in the general linear model, particularly when[23-26], e.g. Most indicators variables have good illustrative power for their result of interest.

The study's results are approved, the logistic regression method was less accomplished than the classification and regression Tree method and Random Forest method. We also note that the priority in the classification among the methods of the three methods was a Random forest method see table(2) [20]. There are three main results:

RF method has a good performance when determine independent variables of kidney disease patients presenting. This method has arousing achievement in classifying that are clinically relevant to the kidney disease. RF method selected a good variable subsets in prospective modelling of Kidney have been identified as suitable[20]. This method can overcome previous limitations that have to do with traditional method, predicting variables have good explanatory power for the results of interest. Through the conclusions obtained, it can be identified the most important recommendations:

1- We recommend applying the RF machine method that has proven successful and superior to other methods (LR and CART) in the classification and diagnosis Chronic Kidney Disease Patients in future.

2- The use of other statistical models for classification such as the Lasso logistic regression methods, cluster analysis and other statistical methods.

References

1. Control, C. f. D.; Prevention, Chronic kidney disease in the United States, 2021. *US Department of Health and Human Services, Centers for Disease Control and Prevention* **2021**.
2. Lv, J.-C.; Zhang, L.-X., Prevalence and disease burden of chronic kidney disease. *Renal Fibrosis: Mechanisms and Therapies* **2019**, 3-15.
3. Ifraz, G. M.; Rashid, M. H.; Tazin, T.; Bourouis, S.; Khan, M. M., Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods. *Computational and Mathematical Methods in Medicine* **2021,** *2021*.
4. Awad, S. M., Chronic renal failure in Al-Anbar of Iraq. *Saudi Journal of Kidney Diseases and Transplantation* **2011,** *22* (6), 1280.
5. Harrell Jr, F. E., *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer: 2015.
6. David W. Hosmer , J., *Applied Logistic Regression*. third ed.; 2013.
7. Harrell, F. E., Regression modeling strategies, with applications to linear models, survival analysis and logistic regression. *GET ADDRESS: Springer* **2001**.
8. Sperandei, S., Understanding logistic regression analysis. *Biochemia medica* **2014,** *24* (1), 12-18.
9. Xu, M.; Watanachaturaporn, P.; Varshney, P. K.; Arora, M. K., Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment* **2005,** *97* (3), 322-336.
10. Steyerberg, E. W., Clinical prediction models: a practical approach to development, validation, and updating Oxford University Press: 2009.

11.Timofeev, R., Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin* **2004**, 1-40.
12.Hoddinott, J., Operationalizing household food security in development projects: an introduction. *Technical guide* **1999,** *1*, 1-19.
13.Loh, W. Y., Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery* **2011,** *1* (1), 14-23.
14.Laber, E.; Murtinho, L., Minimization of Gini Impurity: NP-completeness and Approximation Algorithm via Connections with the k-means Problem. *Electronic Notes in Theoretical Computer Science* **2019,** *346*, 567-576.
15.Breiman, L., Random forests. *Machine learning* **2001,** *45* (1), 5-32.
16.Amaratunga, D.; Cabrera, J.; Lee, Y.-S., Enriched random forests. *Bioinformatics* **2008,** *24* (18), 2010-2014.
17.Trevor Hastie, R. T., Jerome Friedman, *The Elements of Statistical Learning Data Mining, Inference,and Prediction*. 2008.
18.Kirasich, K.; Smith, T.; Sadler, B., Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review* **2018,** *1* (3), 9.
19. Al- Shebly, Jasim. Omar Qusay (2012)**:"** Using intelligent techniques and logistic regression in classification of chronic kidney disease in Erbil governorate " , unpublished thesis, Statistics and Informatics –University of Mosul
20.Al-shallawi, A. N. S., Applied statistical methods for prediction modelling of upper limb functional recovery after stroke.
21.Steyerberg, E. W.; Vergouwe, Y., Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal* **2014,** *35* (29), 1925-1931.
22.Al-Shallawi, A.; Blana, D.; Pandyan, A. In *Improving variable selection for modelling recovery of upper limb function post-stroke*, International Journal of stroke,sage publications Ltd 1 Olivers yard,55 city road,London EC1 y 1SP,England:2017;pp30-30.
23.Dormann, C. F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J. R. G.; Gruber, B.; Lafourcade, B.; Leitão, P. J., Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013,** *36* (1), 27-46.
24.Guo, C.; Yang, H.; Lv, J., Robust variable selection for generalized linear models with a diverging number of parameters. *Communications in Statistics-Theory and Methods* **2017,** *46* (6), 2967-2981.
25.Kwah, L. K.; Harvey, L. a.; Diong, J.; Herbert, R. D., Models containing age and NIHSS predict recovery of ambulation and upper limb function six months after stroke: An observational study. *Journal of Physiotherapy* **2013,** *59* (3), 189-197.
26.Philp, F.; Al-Shallawi, A.; Kyriacou, T.; Blana, D.; Pandyan, A., Improving predictor selection for injury modelling methods in male footballers. **2019**.