Iraqi Journal of Statistical Sciences

https://stats.uomosul.edu.iq/

A Proposed Method Based on Logistic Regression and Cluster Analysis in Selecting Influential Variables for Kidney Failure Patients

Suhaib Bashar Hameed[1] Mahmood M Taher[2]

[1,2]Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

Article information

Article history:
Received: November 18,2024
Revised:March 19,2025
Accepted : April 25,2025
Available online: December 1,2025

Keywords:
Logistic Regression,
Cluster Analysis,
Kidney Failure

Correspondence:
Mahmood M Taher
Mahmood81_tahr@uomosul.edu.iq
suhaib.23csp146@student.uomosul.edu.iq

Abstract

The research aims to study kidney failure by analyzing the relationship between it and a set of independent variables. To achieve this, a method was proposed that relies on reducing the number of independent variables used in binary logistic regression. The method relies on merging the independent variables with the dependent variable using cluster analysis, to improve the accuracy of the model and obtain the best possible results.The proposed method was applied to a sample of 142 individuals to study the relationship between the response variable (renal and non-renal failure) and independent variables such as gender, age, smoking, urea, creatine, and calcium. The results showed that the proposed method succeeded in reducing the number of independent variables and provided an ideal model that classifies the data with high accuracy. The resulting model focused on the two most influential variables, urea and creatine, and achieved a high classification rate of 94.4%.,The proposed method proved effective in reducing the number of variables and achieving accurate results in classifying data related to kidney failure.

DOI:10.33899/iqjoss.v22i2.54080,

1. Introduction

Logistic regression seeks to identify the most appropriate model that can explain the relationship between the response variable and the set of predictive (explanatory) independent variables by modifying the formula of the regression coefficients to predict the probability of the existence of a logit transformation variable. This model is used in the case of a two-level response variable. (Two classifications). Researchers (Thompson & et al., 2015). studied the prevalent causes of mortality for Albertans with chronic renal disease who passed away between 2002 and 2009. This work was funded by an Alberta Heritage Foundation for Medical Research (AHFMR) Chronic Disease Multidisciplinary Collaboration Team grant, and multinomial logistic regression,(Aqlan, & et al., 2017) A variety of methods, including decision trees and logistic regression, were used to analyze historical patient data related to chronic renal disease in the United States,(Cheng & eta,2020)Using univariate logistic regression to identify relevant variables, a model is created to predict how individuals with diabetic kidney disease will advance to renal

failure, (Bai & et al.,2022) Studying the risk of end-stage kidney disease by applying machine learning using logistic regression, (Khan & et al.,2024) Studying different machine learning models, including logistic regression, random forest, decision tree, nearest neighbor, and support vector machine for predicting chronic kidney disease.

In this research, kidney failure was studied by applying a proposed method based on logistic regression and cluster analysis to reduce the dimensions of the variables

2. Material and Methods

2.2 Logistic Regression

Logistic regression was discussed due to its importance, especially in medical studies, as in many cases the phenomenon to be studied does not represent quantities, but rather represents categories or characteristics. These categories may be multiple (more than two), such as the region variable (north, south, east, west). This variable may be binary (two categories), such as (smoker, non-smoker) or (suffering from kidney failure or not), and many other examples.

2.3 Logistic Regression Model

Logistic regression is one of the popular multivariate statistical analysis models used to establish a multivariate regression relationship between dependent and independent variables (Wubalem & Meten, 2020)(Pradhan & Lee,2010).It can be expressed mathematically (Alshebly & Ahmed,2019)(Ali & Taha,2025)

$$p = {}^{1}/_{1+e^{-z}} \tag{1}$$

Where

P: is the probability that varies from zero to one. Z is the linear combination of the predictors and varies from $-1 < z < 0$ for higher odds to $0 < z < 1$ for odds of higher. Z can be defined as: (Ibrahim, et al.,2020)(Sujatha & Sridhar,2021)(Ebrahimi ,& et al.,2021).

$$Z = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_n x_n \tag{2}$$

where x1, x2, x3 … xn are independent variables, Bo is the intercept of the slope of logistic regression analysis, and $\beta 1$, $\beta 2$, $\beta 3$ … βn are the coefficients of the logistic regression analysis.

2.4 Quality of Conformity Test (Hosmer & Lemeshow)

This test is used to determine whether the model represents the data well or not. The Chi-square test of goodness of fit is used to evaluate the difference between the observed and expected values and to test the following hypotheses. (David, & et al., 2013)

$H0$: The observed cases of kidney failure patients are equal to the predicted cases.

$H1$: The observed cases of kidney failure patients are not equal to the predicted cases.

The test format is as follows

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{p}_k)^2}{n'_k \bar{p}_k - (1 - \bar{p}_k)} \quad ; k = 1,2,3,\ldots g ; \tag{3}$$

whereas

n_k: represents the total number of cases,$O_k = \sum_{i=1}^{n'_k} y_i$ The sum of real values,$\bar{P}_k = \sum_{i=1}^{n'_k} \frac{p_i}{n'_k}$ Average predicted values.

2.5 Cluster Analysis

One of the essential statistical analysis techniques is grouping cases according to specific criteria, and then organizing the cases into groups so that the examples in each group are similar (Essa & Shihab,2023). One of the most popular approaches is this one, which divides the study data into a hierarchy of interconnected clusters rather than multiple clusters in a single step. This explains how the clusters are connected through overlapping series to create a hierarchical shape. The two sections of this analysis are the hierarchical analysis of variables (variables) and the hierarchical analysis of cases (cases). (Härdle& Simar,2015) (Hamad,2023).

2.6 Hierarchical Cluster Analysis

This approach, which is regarded as one of the best ones, shows how to connect the clusters by overlapping the series to create a hierarchical form known as the dendrogram. Rather than breaking the study data up into several clusters in one go, this approach creates a hierarchy of interconnected clusters. This method does not require prior knowledge of the number of clusters; it often works by aggregating tiny clusters into bigger groups (the aggregative method) or partitioning large clusters into smaller groups (the partition method). The following figure shows this process. (Manasa, & et al.,2024)(Sarma & Vardhan.2018) (Adel & Rashed .2021)

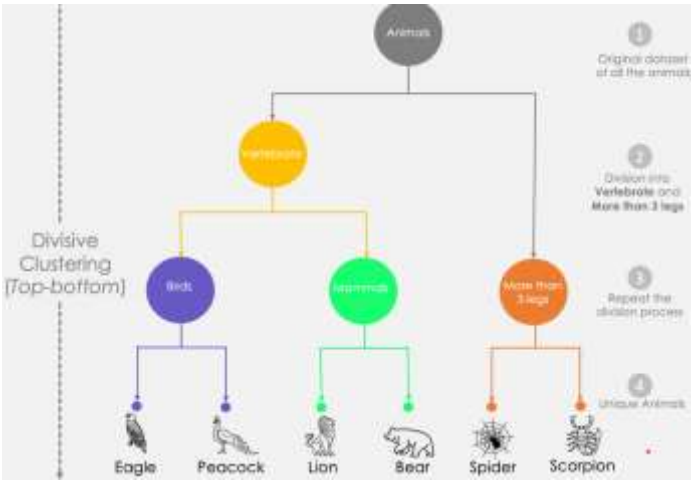


Figure 1. Hierarchical clustering

https://www.datacamp.com/tutorial/introduction-hierarchical-clustering-python

2.6 Typing Techniques for Hierarchical Clustering

Although there are many other kinds of cluster analysis techniques, the following are the most crucial ones:

2.6.1 Single Linkage Clustering

This is one of the simplest and oldest clustering methods. It relies on the distances or similarity coefficients between pairs of elements to determine the links. The process begins by treating each element as a separate cluster, and then the two closest elements (closest in distance in the distance matrix) are combined to form a new cluster. The distance between the new cluster and the rest of the clusters is calculated using the formula (Essa, & et al., 2023):

$d_{ij}^{*} = \min(\, d_{1j} \,, d_{2j})$

where i and j represent the elements in clusters I and, J respectively.

2.6.2 Complete Linkage Clustering

Also known as the Farthest Neighbor method, this method starts by merging the two nearest elements to form a new cluster nucleus. It differs from the Single Linkage method in that it measures the distance between the new cluster and the rest of the clusters using the largest distance between any two points in each of the two clusters. The distance between the two clusters is calculated using the following formula (Essa, & et al., 2023):

$d_{ij}^{*} = \max(\, d_{1j}\, , d_{2j})$

where i and j represent the elements in clusters I and, J respectively.

2.7 Proposed Method

Identify the influential variables and their relationship to the dependent variable based on the concept of cluster analysis through the following steps

1- The independent variables were clustered to determine the number of clusters

2- The independent variables were clustered with the dependent variable

3- Choosing the independent variables that are included in one cluster with the dependent variable

4- Apply logistic regression

5- Determine the significant variables based on the Wald test

6- Selecting significant variables and modeling them using logistic regression

7- Examine the model using the Hosmer and Lemeshow test

8- The model is not significant. Go to step 2

The following diagram shows the proposed method

*The Scheme was prepared by the researcher

Figure (2): Stages of the proposed method

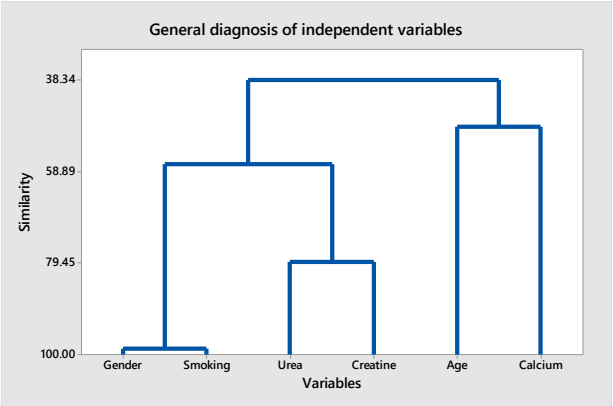
3. Application

The data was obtained from Al-Zahrawi Hospital in the city of Mosul, which represents 6 independent variables. These variables were chosen based on previous studies on kidney failure, and their effect was studied on 142 people.

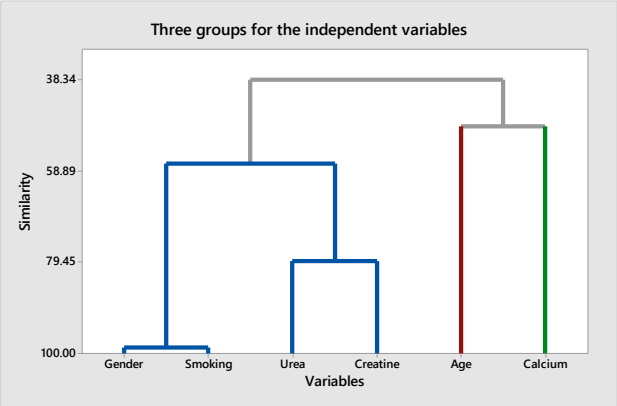
Table (1) Study variables

dependent variables		The description
Y	0	The person does not have kidney failure
	1	The person has kidney failure
Independent variables		**The description**
X1 (Gender)	1	Male
	2	Female
X2 (Age)		The person's age.
X3(Smoking)	0	The person is a non-smoker
	1	The person is a smoker
X4 (Urea)	(10-50)	The normal percentage of urea in the blood
X5(Creatine)	(0.6-1.0)	The normal percentage of Creatinine in the blood
X6 (Calcium)	(8-10)	The normal percentage of Calcium in the blood

Table (1) shows the variables that were recorded for several people, which will be clustered based on Hierarchical cluster analysis. The following figures show the number and stages of cluster formation in addition to the groups of variables that were linked together at each step of the analysis.



Figure(3): clustering variables In one cluster



Figure(4): clustering variables In more than cluster

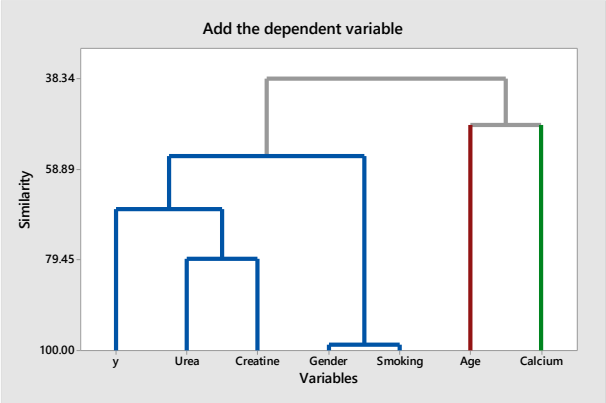
The furthest neighbor approach was applied in the figures (3) and (4). This technique tends to create clusters with comparable diameters and guarantees that all variables in a cluster are within a maximum distance. Figure (3) also shows that the number of clusters for the influential variables is three clusters. This result was adopted and the variables were clustered; figure (4) shows this. The result of the clustering of the variables is shown in the table (2).

Table (2) Cluster aggregates

Cluster	Variables
Cluster 1	Gender Smoking Urea Creatine
Cluster 2	Age
Cluster 3	Calcium

Table (2) represents three clusters, the first cluster includes (Gender Smoking Urea Creatine), the second (Age), and the third cluster (Calcium)

The next step is to enter the dependent variable with the clusters using the furthest neighbor technique, and the figure (5) shows this



Figure(5): Clustering independent variables with the dependent variable

Figure (5) shows the cluster of variables (Gender Smoking Urea Creatine), which represents the first cluster with the dependent variable

Table (3) Cluster totals with the dependent variable

Cluster	Variables
Cluster 1	Y, Gender, Smoking, Urea & Creatine
Cluster 2	Age
Cluster 3	Calcium

Based on the results shown in Table (3), the first cluster was chosen, which represents the variables (y, Gender, Smoking, Urea & Creatine), which is considered the basis for applying the logistic regression model and testing the significance of the variables using the Wald test.

Table (4) Wald test value and parameters values

variables	B	S.E.	Wald	df	Sig.
Urea	.062	.017	13.700	1	.000
Creatine	.601	.250	5.779	1	.016
Gender(1)	-20.023	21516.002	.000	1	.999
Smoking(1)	19.498	21516.002	.000	1	.999

Constant	-9.335	2.512	13.809	1	.000

Table (4) shows the significance of the variables (Urea & Creatine) that will be entered into the binary logistic regression model.

Table (5) Wald test value and parameter values after excluding non-significant variables

	B	S.E.	Wald	df	Sig.	Exp(B)
Urea	.064	.017	14.425	1	.000	1.066
Creatine	.548	.234	5.491	1	.019	1.731
Constant	-9.440	2.437	15.005	1	.000	.000

Table (5) shows the significance of the variables (Urea Creatine), which represent a binary logistic regression model

$$y_i = -9.440 + 0.548x_5 + 0.064x_6 + \varepsilon_i \quad (4)$$

The equation (4) represents the estimated logistic model for studying the disease of kidney failure. To test the significance of the model, the Hosmer and Lemes test was performed, and the test value was (Sig = 0.999), which is a significant value, meaning that the model is significant.

The likelihood ratio for the desired event (O), which is within the period (0,∞), is as follows.

$$O = e^{-9.440+0.548x_5+0.064x_6} \quad (5)$$

The formula for the response probabilities for the logistic regression model within the interval (0,1) is written as follows.

$$p_i = {}^{1}/_{e^{-(-9.440+0.548x_5+0.064x_6)}} \quad (6)$$

Equations (5) and (6) can be interpreted through urea and creatine variables. We find that the urea variable contributes to the effect on the dependent variable (Y) where the effect value is (0.548) and that a change in urea by one unit will lead to an increase in the probability of contracting the disease by (0.548) of the logarithm of the odds ratio. As for the creatine variable, it contributes to the effect on the dependent variable (Y) where the effect value is (0.064) and a change in urea by one unit will lead to an increase in the probability of contracting the disease by (0.064) of the logarithm of the odds ratio.The percentage of description and accuracy are shown in the tables(6)

Table (6) Classification Table

		y		Percentage Correct
		0.00	1.00	
y	.00	12	5	70.6
	1.00	3	122	97.6
	Overall Percentage			94.4

Table (6) shows the classification percentage of the model estimated based on cluster analysis for patients with kidney failure, where the percentage was (94.4)

4.Conclusion

The significance of the urea and creatine variables was revealed through the analysis of the study variables. The effect of their elevation indicates the presence of kidney failure through the efficiency of applying the proposed method, which included

grouping the important independent variables with the dependent variable in one group and excluding the unimportant variables, reducing the independent variables, and obtaining the best model representing the study.

References

1-Thompson, S., James, M., Wiebe, N., Hemmelgarn, B., Manns, B., Klarenbach, S., & Tonelli, M. (2015). Cause of death in patients with reduced kidney function. Journal of the American Society of Nephrology, 26(10), 2504-2511.

DOI: 10.1681/ASN.2014070714

2-Aqlan, F., Markle, R., & Shamsan, A. (2017). Data mining for chronic kidney disease prediction. In IIE Annual Conference. Proceedings (pp. 1789-1794). Institute of Industrial and Systems Engineers (IISE).

https://www.researchgate.net/profile/Abdulrahman-Shamsan/publication/331440652

3-Cheng, Y., Shang, J., Liu, D., Xiao, J., & Zhao, Z. (2020). Development and validation of a predictive model for the progression of diabetic kidney disease to kidney failure. Renal failure, 42(1), 550-559.

https://doi.org/10.1080/0886022X.2020.1772294

4-Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end-stage kidney disease in chronic kidney disease. Scientific reports, 12(1), 8377.

https://doi.org/10.1038/s41598-022-12316-z

5-Khan, N., Raza, M. A., Mirjat, N. H., Balouch, N., Abbas, G., Yousef, A., & Touti, E. (2024). Unveiling the predictive power: a comprehensive study of a machine learning model for anticipating chronic kidney disease. Frontiers in Artificial Intelligence, 6, 1339988.

https://doi.org/10.3389/frai.2023.1339988

6-Wubalem, A., & Meten, M. (2020). Landslide susceptibility mapping using information value and logistic regression models in Goncha Siso Eneses area, northwestern Ethiopia. SN Applied Sciences, 2, 1-19.

https://doi.org/10.1007/s42452-020-2563-0

7-Pradhan, B., & Lee, S. (2010). Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modeling. Environmental Modelling & Software, 25(6), 747-759.

DOI: 10.1016/j.envsoft.2009.10.016

8-Alshebly,O.Q. & Ahmed, R. M,(2019).Prediction and Factors Affecting of Chronic Kidney Disease Diagnosis using Artificial Neural Networks Model and Logistic Regression Model.Journal of Statistical Sciences, 16(1), 140-159.

DOI:10.33899/iqjoss.2019.0164186

9-Ali,Mohammed F.,Taha,Hutheyfa H (2025).Comparison between Logistic Regression and K-Nearest Neighbour Techniques with Application on Thalassemia Patients in Mosul.Iraqi Journal of Statistical Sciences, 22(1), 151-167.

DOI: 10.33899/iqjoss.2025.187789

10-Ibrahim,N.S .Mohammed,N. N. & Mahmood,S. W.(2020).Multicollinearity in Logistic Regression Model -Subject Review- .Journal of Statistical Sciences, 17(1), 46-53.

DOI:10.33899/iqjoss.2020.0165448

11. Sujatha, E. R., & Sridhar, V. (2021). Landslide susceptibility analysis: A logistic regression model case study in Coonoor, India. Hydrology, 8(1), 41.

https://doi.org/10.3390/hydrology8010041

12-Ebrahimi Kalan, M., Jebai, R., Zarafshan, E., & Bursac, Z. (2021). Distinction between two statistical terms: multivariable and multivariate logistic regression. Nicotine and Tobacco Research, 23(8), 1446-1447.

https://d1wqtxts1xzle7.cloudfront.net/92080508/ntaa055-libre.pdf

13-David W. Hosmer, Jr., Stanley Lemeshow and Rodney X. Sturdivant, "Applied Logistic Regression", 3rd Edition. , John Wiley & Sons, (2013). DOI: 10.1080/00401706.1992.10485291

14-Essa, A. K., SH, L. F., & Shihab, D. H. (2023). A comparison between the hierarchical clustering methods for postgraduate students in Iraqi universities for the year 2019-2020 using the cophenetic and delta correlation coefficients. Periodicals of Engineering and Natural Sciences, 11(1), 174-185.

file:///C:/Users/hjzxd/Downloads/174-185_3454-7914-1-RV.pdf

15-Härdle, W. K., & Simar, L. (2015). Applied multivariate statistical analysis. Springer Nature.

http://springer.com/978-3-662-45170-0

16-Hamad, B. A. (2023). Combining Cluster Analysis with Multiple Linear Regression Analysis to Create the Most Accurate Prediction Model for Evaporation in the Kurdistan Region of. Iraqi Journal of Statistical Sciences, 20(2), 188-199.

DOI: 10.33899/iqjoss.2023.181226

17-Manasa, P., Ananth, P., Natarajan, P., Somasundaram, K., Rajkumar, E. R., Ravichandran, K. S., ... & Gandomi, A. H. (2024). An analysis of causative factors for road accidents using partition around medoids and hierarchical clustering techniques. Engineering Reports, e12793. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/eng2.12793

18-Sarma, K. V. S., & Vardhan, R. V. (2018). Multivariate statistics made simple: a practical approach. Chapman and Hall/CRC.https://doi.org/10.1201/9780429465185

19-Adel ,Zainab & Rashed Safwan.(2021).Using the linear and non-linear discriminant function with cluster analysis to study the level of education for the completed stages (governmental – private) In Nineveh Governorate. Iraqi Journal of Statistical Sciences, 18(1), 88-98. DOI: 10.33899/iqjoss.2021.0168377

اسلوب مقترح يعتمد على الانحدار اللوجستي والتحليل العنقودي في اختيار المتغيرات المؤثرة لمرضى الفشل الكلوي

صهيب بشار حميد ، محمود محمد طاهر جادر

قسم الإحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة:يهدف البحث إلى دراسة الفشل الكلوي من خلال تحليل العلاقة بينه وبين مجموعة من المتغيرات المستقلة، ولتحقيق ذلك تم اقتراح طريقة تعتمد على تقليل عدد المتغيرات المستقلة المستخدمة في الانحدار اللوجستي الثنائي، وتعتمد الطريقة على دمج المتغيرات المستقلة مع المتغير التابع باستخدام التحليل العنقودي، لتحسين دقة النموذج والحصول على أفضل النتائج الممكنة ،تم تطبيق الطريقة المقترحة على عينة مكونة من 142 فرد لدراسة العلاقة بين متغير الاستجابة (الفشل الكلوي وغير الكلوي) والمتغيرات المستقلة مثل الجنس والعمر والتدخين واليوريا والكرياتين والكالسيوم، وأظهرت النتائج أن الطريقة المقترحة نجحت في تقليل عدد المتغيرات المستقلة وقدمت نموذج مثالي يصنف البيانات بدقة عالية، وركز النموذج الناتج على المتغيرين الأكثر تأثيرًا وهما اليوريا والكرياتين وحقق معدل تصنيف مرتفع بلغ 94.4%، أثبتت الطريقة المقترحة فاعليتها في تقليل عدد المتغيرات وتحقيق نتائج دقيقة في تصنيف البيانات المتعلقة بالفشل الكلوي.

الكلمات المفتاحية: الانحدار اللوجستي، التحليل العنقودي ، الفشل الكلوي .