RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ
Journal of Advanced Research in Health Sciences /
Sağlık Bilimlerinde İleri Araştırmalar Dergisi 2023
DOI: 10.26650/JARHS2023-1231512

ESTIMATION OF HEALTHY AND LIVER DISEASED INDIVIDUALS BY A LINEAR REGRESSION CLASSIFICATION ALGORITHM

SAĞLIKLI VE KARACİĞER HASTALIĞI OLAN BİREYLERİN DOĞRUSAL REGRESYON SINIFLANDIRMA ALGORİTMASIYLA TAHMİN EDİLMESİ

Handan TANYILDIZI KÖKKÜLÜNK[1]

[1]Altınbas University, Vocational School of Health Sciences, Radiotherapy Program, Istanbul, Turkiye

ORCID ID: H.T.K. 0000-0001-5231-2768

Citation/Atıf: Tanyildizi-Kokkulunk H. Estimation of healthy and liver diseased individuals by a linear regression classification algorithm. Journal of Advanced Research in Health Sciences 2023;6(3):229-233. https://doi.org/10.26650/JARHS2023-1231512

ABSTRACT

Objective: In this study, the aim was to make a categorical estimation of the absent/presence of liver disease by using some blood biochemistry parameters (ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT), gender and the age of healthy individuals, and those with liver disease.
Material and methods: The prediction was obtained with multiple linear regression of machine learning in the R Studio program. Machine learning was improved by selecting parameters that have a high contribution to the prediction by using the Akaike information criterion.
Results: The three strongest parameters with a positive effect on the estimation were AST, BIL, and GGT, respectively; The three strongest parameters with negative effects were CHOL, CHE, and ALB, respectively. The accuracy of the model used was 91%, the precision was 99%, the recall was 0.91, and the F score was 94%. When the correlation relationship graph was examined, it was determined that AST was a strong differential parameter in healthy/liver diseased individuals.
Conclusion: Multiple linear regression is a preferable method for categorical disease classification.
Keywords: Machine learning, liver, classification

ÖZ

Amaç: Bu çalışmada sağlıklı ve karaciğer hastalığı olan bireylere ait bazı kan biyokimya parametreleri (ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT ve PROT), cinsiyet ve yaş bilgileri kullanılarak karaciğer hastalığı yok/var kategorik tahmini yapılması amaçlanmıştır.
Gereç ve Yöntem: R Studio programında makine öğrenmesine ait çoklu doğrusal regresyon ile tahmin elde edilmiştir. Akaike bilgi kriteri kullanılarak tahmin üzerine yüksek katkısı olan parametreler seçilerek makine öğrenmesinde iyileştirilmeye gidilmiştir.
Bulgular: Tahmine pozitif yönlü etkisi olan en güçlü 3 parametre sırasıyla AST, BIL ve GGT; negatif yönlü etkisi olan en güçlü 3 parametre sırasıyla CHOL, CHE ve ALB bulunmuştur. Kullanılan modelin doğruluğu %91, kesinlik %99, geri çağırma 0,91 ve F skoru %94 olarak bulunmuştur. Korelasyon ilişkisi grafiği incelendiğinde AST 'nin sağlıklı/karaciğer hastası bireylerde güçlü bir ayırıcı parametre olduğu tespit edilmiştir.
Sonuç: Çoklu doğrusal regresyonun, kategorik hastalık sınıflandırması için tercih edilebilir bir yöntem olduğu bulunmuştur.
Anahtar Kelimeler: Makine öğrenmesi, karaciğer, sınıflandırma

INTRODUCTION

Machine learning offers strategies, techniques, and resources that can assist in resolving diagnostic and prognostic issues in a range of medical specialties. The significance of clinical indicators and their combinations for prognosis is examined using machine learning. It has developed into a method that is often used to gather medical data for things like planning treatments, outcome studies, and estimating illness progression. Additionally, machine learning is employed for data analysis in the form of smart alerts, continuous data interpretation in intensive care units, and the replication of inaccurate or missing data based on pattern discovery in the available data. It is well recognized that when machine learning techniques are successfully deployed, they aid in the integration of computer-based healthcare systems, present chances to facilitate and enhance the work of medical professionals, and ultimately improve the effectiveness and caliber of medical care (1). Machine learning is being used in the healthcare industry not to replace doctors, but to reduce their workload and give patients feedback more quickly and efficiently.

Machine learning has several important applications in the field of medical diagnosis (2,3). In this approach, doctor-based techniques are used to create hypotheses from patient data. In order to do this, the system is enhanced with symbolic learning

Corresponding Author/Sorumlu Yazar: Handan TANYILDIZI KÖKKÜLÜNKK **E-mail:** handan.kokkulunk@altinbas.edu.tr

Submitted/Başvuru: 09.01.2023 • **Revision Requested/Revizyon Talebi:** 17.01.2023 • **Last Revision Received/Son Revizyon:** 06.07.2023 • **Accepted/Kabul:** 24.07.2023 • **Published Online/Online Yayın:** 10.10.2023

techniques and knowledge management capabilities that are appropriate for the doctor's interpretation of the case. As a result, the forecast is generated using straightforward rules or, most often, a decision tree. The reporting of a medical imaging as a certain radiologist using machine learning is an illustration of this.

Biomedical signal processing is an additional application area (4). With the use of machine learning techniques, it is feasible to model the linear or non-linear relationships that exist between the data and find the fundamental features and information that are concealed in physiological signals or that are likely to be disregarded. Additionally, machine learning is employed in radiography, magnetic resonance imaging, endoscopy, confocal microscopy, computer tomography, and other imaging techniques, particularly for the detection of cancerous regions. In addition to all of these uses, the most typical application of machine learning is to forecast diseases using categorical classification algorithms and patient data (5-7).

The most accessible and practical way for identifying biomarkers that can be used to predict disease is through blood biochemistry characteristics. Some disorders, including diabetes, hypertension, heart disease, hormonal diseases, blood diseases, and liver diseases, can be preliminarily diagnosed using it (8,9). When examining liver illnesses, the blood sample is tested for the presence of biomarkers including ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT (10). ALB acts as a source of amino acids, carries things through the blood, and aids in osmotic pressure maintenance. The liver contains the enzymes ALP, ALT, AST, CHE, and GGT. BIL is a yellow pigment created when red blood cells are broken down. A lipid molecule called CHOL is necessary for a number of physiological activities. A byproduct of muscle metabolism called CREA is eliminated by the kidneys. ALB and other globulins are included in the PROT measurement of the blood's overall protein concentration. Abnormalities in these biomarkers indicate liver disease (11).

In this study, it was aimed to estimate the presence or absence of liver disease by using liver-related parameters, gender and age information from blood biochemistry results of individuals without liver disease (healthy) and diagnosed with liver disease (patient).

MATERIAL and METHODS

A. Dataset and machine learning preparation

The data set was acquired from the open source Kaggle website and included 615 people who were categorized as healthy, hepatitis, fibrosis, and cirrhotic people (12). Age, gender, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT values are all attributes in the data stack that was used. Class values were divided categorically as only healthy (1) and patient (2). By averaging the appropriate column, missing measurements in the dataset were filled in. Detailed information about the attributes in the dataset is given in Table 1.

In the study, the xlsx extension data set was used to code machine learning algorithms through the R Studio program. An amount of 80% of the data set was used to train the algorithm, and 20% of the data set was used in the test set to control the accuracy of the prediction and to evaluate the performance.

B. Forecast ModeL

Machine learning includes a variety of categorization and regression estimation techniques. Multiple linear regression (MLR) was chosen among machine learning regression methods for estimation since the dependent variable in the data set utilized in this study must be estimated as a categorical data type.

B. 1. Multiple Linear Regression (MLR)

One technique for determining the relationship between multiple independent variables (x1–xn) and a dependent variable, y, is known as multiple linear regression (MLR).

Table 1: Some information about the attributes in the dataset

Category		Health (n=540)	Liver Disease (n=75)
		Min-max values	Min-max values
Age	Age	32-77	19-75
Sex	Sex (371 M, 244 F)	-	-
ALB	Albumin Blood Test (g/dL)	14.9-82.2	20-50
ALP	Alkaline phosphatase (U/L)	27-208.2	11.3-416.6
ALT	Alanine Transaminase (U/L)	2.5-325.3	0.9-258
AST	Aspartate Transaminase (U/L)	10.6-188.7	16.7-324
BIL	Bilirubin (mg/dL)	0.8-59.1	5-254
CHE	Acetylcholinesterase (U/g)	3.44-15.43	1.42-16.41
CHOL	Cholesterol (mg/dL)	2.61-9.43	1.43-9.67
CREA	Creatinine (mg/dL)	8-170	45.4-1079,1
GGT	Gamma Glutamyl Transferase (U/L)	4.5-345.6	11.5-650.9
PROT	Protein (g/dL)	44.8-86.5	54.2-90

Multiple linear regression is cited as a typical technique for estimating an unknown variable's value from the known values of two or more other variables (13). The following Eq. 1 for n independent variables, which might be linear or linearized, often expresses this relationship (14).

$$y = b_0 + b_1x_1 + b_2x_2 +\cdots+ b_nx_n \qquad (1)$$

By adjusting the input parameters, 4 alternative fit operations were performed in the study. It is intended to obtain the input parameters with the greatest impact on the estimation using these 4 various procedures. For each fit procedure, the Akaike information criterion (AIC) is determined (15). The AIC is a single numerical value that can be used to identify the best model for a given dataset among the various models. A better forecast is made by the model with the lower AIC value than by the other models.

C. Performance evaluation criteria

Efficiency research of machine learning classification algorithms is measured using metrics such as accuracy, precision, recall, and F-score (16). After making an estimate, the complexity matrix will be determined for this. It displays TP true positive, TN true negative, FP false positive, and FN false negative in order to depict the positive healthy people and the negative persons with liver disease in the complexity matrix and metrics. As a result, Eq. 2-5, as illustrated in the table below, is used to calculate accuracy, precision (P), recall (R), and F-score.

$$Accuracy = \frac{TN+TP}{TP+TN+FP+FN} \qquad (2)$$

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

$$Fscore = \frac{2(R*P)}{R+P} \qquad (5)$$

In order to assess how much the data contribute to the prediction of disease, the correlation matrix, which is the table of correlation coefficients between various factors, will also be visualized using the corrplot tool.

RESULTS and DISCUSSION

ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT and PROT parameters were used in the study, while the search for the presence of liver disease through hemogram biochemistry results was carried out with machine learning. While these parameters are within the typical reference range for healthy people, they do not match the reference range for sick people. In the adult population, the following normal reference ranges are used: 35-52 for ALB, 30-120 for ALP, 0-45 and 0-31 for ALT (male and female), 0-35 and 0-31 for AST (male and female), BIL 3-13, 8-18 for CHE, 0-5.2 for CHOL, 50-110 for CREA, 0-55 and 0-38 (male and female) for GGT, and PROT 60-83 U/L (17). In the data included in the study, the mean age, the ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT and PROT values for healthy individuals were calculated as 42.00, 68.86, 27.61, 27.12, 8.47, 8.38, 5.48, 78.75, 30.62 and 71.87 U/L, respectively. These mean values were calculated as 38.70, 61.35, 34.48, 89.94, 32.41, 6.83, 4.52, 99.53, 103.67 and 73.23 U/L for sick individuals, respectively. The ALB, ALP, ALT, CHOL, CREA, and PROT readings of patients with liver illness were determined to be normal when the hemogram findings were compared with normal reference values. Additionally, it was shown that liver patients had low CHE values and high AST, BIL, and GGT values.

Figure 1 presents a graph illustrating the relationship between machine learning and the features used to predict liver disease. The use of the correlation relationship graph to reduce the amount of features and to exclude those that have little to no impact on the prediction are examples of intermediate operations that may be used to improve machine learning. Since most of the factors that have a strong link with the estimate of the features will be incorporated in the algorithm, increasing the results of metrics like accuracy and precision in machine learning, Figure 1 was also employed as a parameter analysis. It was discovered that the three strongest parameters that have a negative effect on the estimation indicated in red are CHOL, CHE, and ALB, respectively. The three strongest parameters that have a favorable influence on the forecast shown in blue are AST, BIL, and GGT, respectively. It was discovered that other factors contributed less positively or negatively to liver prediction.

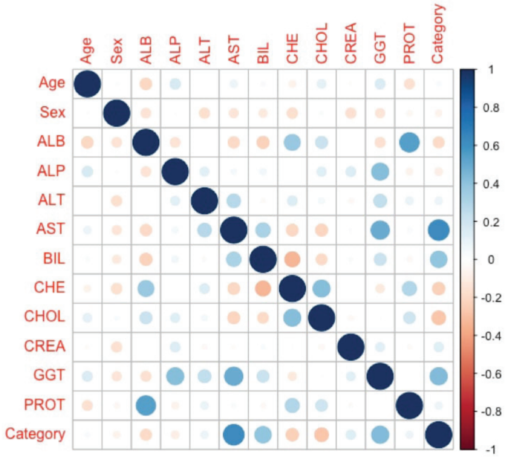

Figure 1. Correlation relationship graph of features with each other. ALB:Albumin, ALP: Alkaline phosphatase, ALT: Alanine Transaminase, AST: Aspartate Transaminase, BIL: Bilirubin, CHE: Acetylcholinesterase, CHOL: Cholesterol, CREA: Creatinine, GGT: Gamma Glutamyl Transferase, PROT: Protein

The glm.fit algorithm was used on 4 different feature collections taking into account the findings of the feature analysis. Table 2 provides details on the glm.fit techniques utilized and the derived AIC results.

Table 2: Machine learning improvement step results

glm.fit number	Attributes	AIC
glm.fit.0	Age + Gender + ALB + ALP + ALT + AST + BIL + CHE + CHOL + CREA + GGT + PROT	151.14
glm.fit.1	Gender + ALB + ALP + ALT + AST + BIL + CHOL + CREA + GGT + PROT	147.82
glm.fit.2	Gender + ALP + ALT + AST + BIL + CHOL + CREA + GGT + PROT	148.41
glm.fit.3	Age + Gender + ALP + ALT + AST + BIL + CHOL + CREA + GGT + PROT	150.33

ALB: Albumin, ALP: Alkaline phosphatase, ALT: Alanine Transaminase, AST: Aspartate Transaminase, BIL: Bilirubin, CHE: Acetylcholinesterase, CHOL: Cholesterol, CREA: Creatinine, GGT: Gamma Glutamyl Transferase, PROT: Protein, glm: General linear model

It is advised to choose the model with the lowest AIC value for predicting the disease. Due to this, the estimation was performed using the features from glm.fit.1, which produced the AIC result with the lowest value (147.82).

Gender, ALB, ALP, ALT, AST, BIL, CHOL, CREA, GGT, and PROT characteristics were used in machine learning utilizing MLR as a healthy/patient estimation method. Table 3 contains the complexity matrix that was produced as a result of the estimation.

The success of the model used to predict healthy individuals with liver illness using the multiple linear regression algorithm of machine learning was found to be 91%. It was determined to be 99% accurate, which indicates how many of the values we anticipated as positive truly are positive. The results showed that the recall and F score were 0.91 and 94%, respectively.

Among the features that excluded AST, Teke discovered that direct bilirubin had the highest correlation with the prediction of liver illness, with a score of 0.87. Additionally, they discovered that the machine learning he created using the logistic regression model had a training accuracy of 87%, test accuracy of 84%, precision of 89%, a F score of 0.78%, and recall of 76% (18). AST was a useful distinguishing characteristic, according to Akter et al., who investigated the liver disease prediction model with machine learning from biochemical test data (19). The accuracy of the random forest and classification-regression trees algorithms was 94% and 95%, respectively (19). Another study used decision trees, logistic regression, random forest, support vector machines, k-near neighbor, and Naive Bayes algorithms to make predictions. The corresponding accuracy percentages for these forecasts were 75%, 74%, 69%, 64%, 62%, and 53% (20). While diagnosing liver disease in machine learning, a comparison of classifications was made with the support vector method or the Naive Bayes-support vector method using several biomarkers (21,22). Similarly, in various studies, methods such as random forest, functional tree, and logistic regression were tried and various liver disease predictions were made using a small number of biomarkers, and the highest success was found to be 82% and the highest accuracy was found to be 87% (18,23).

Table 3: Complexity matrix for MLR estimation of the healthy/patient population

glm.pred_test	Positive	Negative
Positive	110	1
Negative	2	10

glm.pred.test: General linear model predictive test, MLR: Multiple linear regression

The metric having the greatest impact on liver disease was discovered to be AST, and our study was found to be consistent with the literature that uses machine learning to predict liver disease using similar features. Contrary to the literature, the employment of MLR algorithms in machine learning led to the greatest values being attained in all performance evaluation criteria, particularly accuracy. It is believed that getting high accuracy also depends on the quantity of the dataset.

CONCLUSION

The MLR classification algorithm, which is based on machine learning, was used in this study to predict disease based on the categorical classification of healthy/liver disease according to age, gender, and various hemogram biochemical values. The correlation association graph was studied, and it was shown that the AST was a significant differential parameter between healthy and liver-ill people.

The accuracy, precision, recall, and F score of the prediction made by machine learning using the complexity matrix were 91%, 99%, 0.91, and 94%, respectively. Although there are many studies on estimation methods, it has been observed that high success has been achieved by including the features that have a high positive or negative effect on the estimation with the glim.fit function proposed in our study. On this occasion, more efficient results can be obtained even if a single algorithm is used for estimation. If the used model is run on more data, improved accuracy will be possible.

Acknowledgements: I would like to thank Res. Assist. Fatih Okumuş for his valuable suggestions and contributions during the development of this research study.

Ethics Committee Approval: The author declared that this study does not require ethics committee approval.

Peer Review: Externally peer-reviewed.

Conflict of Interest: The author has no conflict of interest to declare.

Financial Disclosure: The author declared that this study has received no financial support.

REFERENCES

1. Magoulas GD, Prentza A. Machine learning in medical applications. In: Paliouras G, Karkaletsis V, Spyropoulos CD, editors. Machine learning and its applications: advanced lectures. Berlin, Heidelberg: Springer; 2001 (cited 2022) p.300–7. (Lecture Notes in Computer Science). https://doi.org/10.1007/3-540-44673-7_19.
2. Stausberg J, Person M. A process model of diagnostic reasoning in medicine. Int J Med Inform 1999;54(1):9-23.
3. B. Zupan, J. Halter, M. Bohanec. Qualitative model approach to computer assisted reasoning in physiology. Computer Science 1998 (cited 2022 September 2) https://www.semanticscholar.org/paper/Qualitative-Model-Approach-to-Computer-Assisted-in-Zupan-Halter/4197bc7fc5af6754e99d39c204eef80a99e324c3
4. Gindi GR, Darken CJ, O'Brien KM, Stetz ML, Deckelbaum LI. Neural network and conventional classifiers for fluorescence-guided laser angioplasty. IEEE Trans Biomed Eng 1991;38(3):246-52.
5. Srinivas S. A machine learning-based approach for predicting patient punctuality in ambulatory care centers. Int J Environ Health Res 2020;17(10):3703.
6. Anusuya V, Gomathi V. An efficient technique for disease prediction by using enhanced machine learning algorithms for categorical medical dataset. I Inf Technol Control 2021;50(1) :102-22.
7. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Inform Decis Mak 2019;19(1):281.
8. Petrie JR, Guzik TJ, Touyz RM. Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. Can J Cardiol 2018;34(5):575-84.
9. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin 2005;55(2):74-108.
10. Jayaswal ANA, Levick C, Selvaraj EA, Dennis A, Booth JC, Collier J, et al. Prognostic value of multiparametric magnetic resonance imaging, transient elastography and blood-based fibrosis markers in patients with chronic liver disease. Liver Int 2020;40(12):3071-82.
11. Abebe M, Melku M, Enawgaw B, Birhan W, Deressa T, Terefe B, et al. Reference intervals of routine clinical chemistry parameters among apparently healthy young adults in Amhara National Regional State, Ethiopia. Plos one 2018;13(8):e0201782.
12. Hepatitis c prediction dataset. 2021 (cited 2022 August 1): 1(1):(1 screen). https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset.
13. Yee MM, Aung EE, Khaing YM. Forecasting stock market using multiple linear regression. IJTSRD 2019;3(5):2174-6.
14. Giacomino A, Abollino O, Malandrino M, Mentasti E. The role of chemometrics in single and sequential extraction assays: a review. Part II. Cluster analysis, multiple linear regression, mixture resolution, experimental design and other techniques. Anal Chim Acta 2011;688(2):122-39.
15. Khalid A, Sarwat AI. Unified univariate-neural network models for lithium-ion battery state-of-charge forecasting using minimized akaike information criterion algorithm. IEEE Access 2021;9:39154-70.
16. Hasan M, Islam MdM, Zarif MII, Hashem MMA. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things 2019;7:100059.
17. Reference ranges for blood tests. In: Wikipedia. 2022 (cited 2022 September 1). https://en.wikipedia.org/w/index.php?title=Reference_ranges_for_blood_tests&oldid=1109845763.
18. Teke M. Prediction of liver diseases with machine learning method. SMUTGD 2022;5(1):115-22.
19. Akter S, Shekhar HU, Akhteruzzaman S. Application of biochemical tests and machine learning techniques to diagnose and evaluate liver disease. Adv Biosci Biotechnol 2021;12(6):154-72.
20. Rahman AKM, Shamrat FM, Tasnim Z, Roy J, Hossain S. A comparative study on liver disease prediction using supervised machine learning algorithms. Int J Sci Technol Res 2019;8(11):419-22.
21. Schiff ER, Maddrey WC, Reddy KR. Schiff's Diseases of the Liver. 12th Edition. USA: Wiley-Blackwell; 2017. pp.135-218.
22. Sorich MJ, Miners JO, McKinnon RA, Winkler DA, Burden FR, Smith PA. Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human udp-glucuronosyltransferase isoforms. J Chem Inf Comput Sci 2003;43(6):2019-24.
23. Saygın E, Baykara M. Karaciğer yetmezliği teşhisinde özellik seçimi kullanarak makine öğrenmesi yöntemlerinin başarılarının ölçülmesi. FÜMBD 2021;33(2):367-77.