Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(2):118-28

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2022-90697

The Classification Capability of Urine Biomarkers in the Diagnosis of Pancreatic Cancer with Logistic Regression Based on Regularized Approaches: A Methodological Research

Pankreas Kanseri Teşhisinde İdrar Biyobelirteçlerinin Düzenlileştirme Yaklaşımlarına Dayalı Lojistik Regresyonla Sınıflandırma Kabiliyeti: Metodolojik Çalışma

Özlem ALPU[a], Güven PEKDEMİR[a]

[a]Department of Statistics, Eskişehir Osmangazi University Faculty of Science and Letters, Eskişehir, Türkiye

ABSTRACT Objective: This study aims to compare the accuracy, reliability, and validity levels of the techniques by using various performance measures applying logistic regression models based on regularization approaches from data mining classification techniques on a dataset. **Material and Methods:** With the development of computerization and technology, machine learning is used in many fields as well as in the field of medicine**.** It has grown in popularity, particularly in cancer diagnosis. A urine biomarkers dataset from the public platform Kaggle database, which is freely available to all researchers, was used to reveal the most appropriate model for diagnosing patients' pancreatic ductal adenocarcinoma (PDAC). Because of the multicollinearity, the following regression models were considered to classify the disease diagnosis: Logistic lasso, logistic ridge, logistic elastic net, logistic adaptive lasso, logistic adaptive elastic net, and logistic adaptive group lasso. The classification success of the methods used was compared using reliability and validity criteria. **Results:** There were three statistically significant variables in all logistic regularization models, according to PDAC diagnostic results. Compared to the estimated model results, the logistic adaptive group lasso regression model appears to perform better in PDAC diagnosis. In addition to the three variables in this model, the variables age and plasma CA19-19 have been identified as important variables in PDAC diagnosis. **Conclusion:** As a result of comparative analyses, the logistic adaptive group lasso regression model outperformed the others in terms of performance measures.

Keywords: Logistic regression; regularization methods; lasso; adaptive group lasso

ÖZET Amaç: Bu çalışma, veri madenciliği sınıflandırma tekniklerinden düzenlileştirme yaklaşımlarına dayalı lojistik regresyon modellerini bir veri kümesi üzerinde uygulayarak tekniklerin doğruluk, güvenirlik ve geçerlilik düzeylerini çeşitli performans ölçüleri aracılığıyla karşılaştırmayı amaçlamaktadır. **Gereç ve Yöntemler:** Makineleşmenin ve teknolojinin gelişmesiyle makine öğrenmesi tıp alanında olduğu gibi birçok alanda kullanılmaktadır. Özellikle kanser teşhisi konusunda artan bir kullanıma sahiptir. Çalışmada pankreatik duktal adenokarsinomunu (PDAC) teşhis etmede en uygun modeli ortaya çıkarmak amacıyla tüm araştırmacıların kullanımına ve erişimine açık olarak sunulan Kaggle veri tabanından bir idrar biyobelirteçleri veri kümesi kullanıldı. Veri kümesindeki değişkenler arasında çoklu doğrusal bağıntı problemi olması nedeniyle, hastalık teşhisini sınıflandırmada lojistik lasso, lojistik ridge, lojistik elastik ağ, lojistik uyarlamalı lasso, lojistik uyarlamalı elastik ağ ve lojistik uyarlamalı grup lasso regresyon modelleri ele alınmıştır. Kullanılan modeller sınıflandırma başarısı, güvenirlik ve geçerlik kriterleri kullanılarak karşılaştırılmıştır. **Bulgular:** Tüm düzenlileştirme tahmin modellerindeki PDAC teşhisi sonuçlarına göre istatistiksel olarak anlamlı bulunan üç değişken belirlenmiştir. Tahmin edilen model sonuçları karşılaştırıldığında, PDAC teşhisinde lojistik uyarlamalı grup lasso regresyon modelinin daha iyi sonuç verdiği görülmektedir. Bu modelde tüm modellerde anlamlı bulunan değişkenlere ilave olarak yaş ve plazma CA19-19 değişkenlerinin de önemli değişkenler olarak belirlendiği görülmektedir. **Sonuç:** Karşılaştırmalı analizler sonucunda performans ölçülerine göre lojistik uyarlamalı grup lasso regresyon modelinin en iyi performansı gösterdiği gözlenmiştir.

Anahtar Kelimeler: Lojistik regresyon; düzenlileştirme yöntemleri; lasso; uyarlamalı grup lasso

Correspondence: Özlem ALPU
Department of Statistics, Eskişehir Osmangazi University Faculty of Science and Letters, Eskişehir, Türkiye
E-mail: oalpu@ogu.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 19 Apr 2022 ***Received in revised form:*** 23 May 2022 ***Accepted:*** 24 May 2022 ***Available online:*** 13 Jun 2022

Cancer is still a disease that causes a lot of deaths in the 21st century. Although there are very promising clinical trials, the survival rate in some types of cancer, such as pancreatic cancer, is unfortunately not very high. One of the most important steps in the treatment of cancer is early diagnosis. Unfortunately, when clinical findings are seen, cancer has already spread throughout the body. Pancreatic cancer, one of the cancer types, ranks fourteenth among the most common cancers in the world and seventh in cancer-related deaths according to 2020 GLOBOCAN data.[1] It is forecasted that pancreatic cancer will be the third leading cause of cancer due to slow progress in delayed diagnosis and treatment options until 2030.[2]

More than 90% of pancreatic cancers constitute ductal adenocarcinomas. Pancreatic ductal adenocarcinoma cancer (PDAC) is known as the deadliest and most aggressive of pancreatic cancer types. It is said that it is not possible to detect it in its early stages, since there is no marker as the exact cause of the disease. In this context, intensive scientific studies are carried out to detect pancreatic ductal adenocarcinoma cancer at an early stage for the quality of life of people. It is said that the number of people who have contracted the disease and lived for more than 5 years is less than 10%, and that poor results depend on the late diagnosis, but if the diagnosis of the disease is detected in its early stages and the tumors are still small, the 5-year survival rate can be increased by up to 32%.[3-5] For the diagnosis of pancreatic cancer, the carbohydrate antigen 19-9 (CA19-9) is currently utilized as the only blood-based biomarker. Researchers currently aim to diagnose pancreatic cancer before clinical findings with a urine test. Mayerle et al. state that pancreatic cancer is hard to distinguish from chronic pancreatitis, with current diagnosis accuracy ranging from 50% to 60%.

The diagnosis of diseases on time is critical for the selection and implementation of treatment options quickly. This importance has been inevitable to take advantage of the information sector in the field of medicine. By using information technologies and health information systems, it is possible to shorten the time allocated to disease processes, provide quality service and achieve positive results.

Even in the early stages of pancreatic cancer, because it is a compelling disease for physicians, it is of great importance to realize the diagnosis of pancreatic cancer in accurate and high performance. With the development of technology, machine learning has been used in many areas as in medicine field. It is widely used for cancer diagnosis.

Methods used in the diagnosis of cancer include clustering methods, classification methods, regression analysis, artificial neural networks, decision trees, fuzzy logic, Naive Bayes, Support Vector Machines (SVM)[7,12], logistic regression and some hybrid methods in the literature.[2,3,6-18] Some produce better results than others, but there has been still no comprehensive work to compare the possible feature selection methods and classifiers.

In many fields of study, such as medicine and epidemiology, it is very important to estimate the probability of an event occurring, which is a level of the categorical variable, with the explanatory variables associated with that variable. Most of the methods used in the diagnosis of any disease are based on classification. Researchers are working to improve the accuracy and precision of these classification methods during the diagnostic process.

In this study, it was aimed to select the most appropriate classification method in terms of validity and reliability criteria in the diagnosis of PDAC on the Urine Biomarkers dataset for Pancreatic Cancer. When there is a multicollinearity problem between the explanatory variables, overfitting and/or statistical errors occur in the models. To cope with the overfitting and multicollinearity, regularized models were included in the analysis. The study focused on the use of five regularization methods with the logistic model (logistic ridge, logistic lasso, logistic elastic net, logistic adaptive lasso, logistic adaptive elastic net, and logistic adaptive group lasso). The results of the analysis were compared with various classification performance measures.

MATERIAL AND METHODS

CLASSIFICATION

If the response variable is categorical, logistic regression models are often preferred by researchers as a classification method when statistically modeling the relationship between the response and explanatory variables. Logistic regression aims to predict the probability of the response variable's categories. The model prediction between response and explanatory variables is being realized through a link function. The lack of multicollinearity among explanatory variables and working with a sufficient sample size will ensure that parameter estimations are more efficient in logistic regression. Various methods have been developed to cope with the multicollinearity problem caused by explanatory variables with a high correlation between them due to the nature of the variables studied. Some of these methods will be included in the following parts of the study.

LOGISTIC REGRESSION BASED ON REGULARIZATION TECHNIQUES

Ridge regression, proposed to dealing with multicollinearity, is built on the basis of minimizing the sum of squares of residuals to achieve β coefficients under the L2 norm-restricted.[19] The aim is to eliminate the collinearity among explanatory variables by taking the regression model coefficients closer to zero. The problem of multicollinearity is also common in logistic regression. Several estimators have been developed to cope with this difficulty in logistic regression. The ridge estimator was first proposed by Schaefer et al.[20]

For the use of a ridge estimator in logistic regression models, the model parameters are estimated by adding the regularization term to the log-likelihood function. The optimization function used to get the coefficients of logistic ridge regression is given below:

$$\hat{\beta}_{Logistic-Ridge} = \underset{\beta}{argmin} \sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Where $\lambda \geq 0$, is the regularization constant that controls how much shrinkage occurs. Parameter $l_2 = \sum_{j=1}^{p} \beta_j^2$ is the ridge regularization function.[21]

Lasso is one of the regularization approaches used in the multicollinearity problem, which performs variable selection and parameter estimation together.[22] The objective function of the method is based on minimizing the sum of the squares of residuals under the L1 norm. In this method, the lasso regularization term is added to the negative log-likelihood function of logistic regression and the model coefficients are estimated.

The optimization function used to get the coefficients of logistic lasso regression is given below:

$$\hat{\beta}_{Logistic-Lasso} = \underset{\beta}{argmin} \left[-\sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

Where $l_1 = \sum_{j=1}^{p} |\beta_j|$ is the lasso regularization function. The regularization procedure brings the coefficients of $\hat{\beta}_{Lasso}$ with multicollinearity problems closer to zero, ensuring that statistically meaningful variables remain in the model.

Zou and Hastie suggested Elastic Net (EN) to address some of the drawbacks of lasso (such as randomly selecting one of the highly correlated variables and ignoring others or allowing as many variables as the maximum size of the sample in datasets where the number of variables is greater than the sample

size).[23] It is an extension of the Lasso method, which is robust to the high correlation between the explanatory variables.

Elastic Net regression is a hybrid of Lasso and ridge regression. Therefore, the model will attempt to bring the variable coefficients closer to zero while also indirectly selecting variables by equalizing some variable coefficients to zero. The elastic net method uses a mixture of ridge and lasso regularization parameters. The optimization function of logistic elastic net regression coefficients is given below.[24]

$$\hat{\beta}_{Logistic-Elastic} = \underset{\beta}{argmin}\left[-\sum_{i=1}^{n}\{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda\left[\tfrac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right]\right]$$

There are two parameters (λ ve α) that must be determined in the regularization function used in logistic-elastic net regression. The parameter α can take a value between zero and one ($0 \leq \alpha \leq 1$). When the regularization parameter of the regularized regression model is set too high, the model becomes quite simple and underfitting occurs. If this parameter is set to low, the problem of overfitting occurs by neglecting regularization. Therefore, the selection of these parameters is of great importance.

LOGISTIC REGRESSION BASED ON ADAPTIVE REGULARIZATION TECHNIQUES

By applying identical regularization to all regression coefficients, lasso leads estimates to be biased. To solve this problem, Zou suggested an adaptive lasso approach that assigns large weight to small regression coefficients and low weight to large ones.[25] By regularizing large coefficients, adaptive lasso prevents overfitting as a regularization method. Furthermore, it has the same advantage as lasso: it can perform the variable selection with regularization because it reduces some coefficients to exactly zero.

The optimization function of the adaptive logistic lasso regression model is given below.[26]

$$\hat{\beta}_{Logistic-ALasso} = \underset{\beta}{argmin}\left[-\sum_{i=1}^{n}\{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\} + \lambda\sum_{j=1}^{p} w_j|\beta_j|\right]$$

Adaptive elastic net was introduced by Zou, Zhang, and Ghosh.[27,28] Elastic net regression and adaptive lasso regression have been combined to form adaptive elastic net. In other words, it is a mixture of adaptive l_1 norm and l_2 norm regularization functions. It is noted that this method performs well when partial correlations between variables are too high. Ridge regression can be used to calculate adaptive weights.[24]

The coefficients of the logistic adaptive elastic net regression model are estimated using the following optimization problem:[26]

$$\hat{\beta}_{Logistic-AElastic} = \underset{\beta}{argmin}\left[-\sum_{i=1}^{n}\{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\} + \lambda_1\sum_{j=1}^{p} w_j|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right].$$

Wang and Leng proposed the adaptive group lasso method because of the inefficient estimates and variable selection inconsistency of the group lasso method developed by Yuan and Lin.[29,30] The main difference is that the adaptive group lasso method applies different parameter settings for different variables and produces a different amount of shrinkage for different factors.[29] Logistic adaptive group lasso model coefficients were estimated in the study by adding the regularization function recommended by Wang and Leng to the negative log-likelihood function of logistic regression.

The optimization function for the model is given below.[31]

$$\hat{\beta}_{Logistic-AGroupLasso} = \underset{\beta}{argmin}\left[-\sum_{i=1}^{n}\{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + n\sum_{j=1}^{p}\lambda_j\|\beta_j\|\right].$$

The adaptive group lasso estimator is obtained by minimizing this objective function. Different selection criteria, such as cross-validation, generalized cross-validation, AIC, or BIC, can be used to select regularization parameters.[30]

DATASET

The urine biomarkers for pancreatic cancer dataset, published publicly by Debernardi et al. on the Kaggle public data platform, was used in the study.[3] It is a dataset that can be used to reveal some demographic characteristics of the patient and the effects of biomarkers on disease diagnosis. The dataset consists of a total of 13 variables and 590 observations. The variables in the dataset can be listed as a patient group, the name of the institution from which the data is taken, age, gender, disease diagnosis, disease stage, benign hepatobiliary diseases, blood plasma levels of CA19-9 monoclonal antibody, creatinine value for urine biomarkers of kidney function, urine levels of lymphatic vascular endothelial hyaluronan receptor 1 (LVYE1) that is a protein that can play a role in tumor metastasis, urine levels of the protein associated with pancreatic regeneration (REG1B), levels of urine Trefoil factor 1 (TFF1), which may be associated with the repair of the urinary tract, and urine levels of a protein (REG1A) that may be associated with pancreatic regeneration.

The study aims to determine the probability that the individual will especially be PDAC with some explanatory variables in the diagnosis of pancreatic cancer. Therefore, the study focuses specifically on the diagnosis of a level (PDAC) of the categorical response variable.

PREPROCESSING OF DATASET

In the preprocessing of dataset, some variables (i.e., patient group, institution from which the data was taken, disease stage, and benign sample diagnosis) were removed from the data because they contained a high percentage of missing values (about 65%) or were not for the purpose of the research. The CART approach in the *mice* package in the R programming language was used to solve the missing value problem. The outliers were identified using multivariate robust mahalanobis distance (observations 92, 142, 182, 294 and 362) and retrieved from the data set. Logarithmic transformation has been applied to Plasma CA19-9, Creatinine, LVYE1, REG1B, TFF1 and REG1A continuous variables.

Table 1 shows the variables used in the study.

TABLE 1: Patients characteristics.

Variables	Levels of the variables
1. Diagnosis (Response variable)	1=Control, 2=Benign, 3=PDAC
2. Sex	Female, Male
3. Age	26, ... ,89
4. Stage	0, IA, IB, II, IIA, IIB, III, IV
5. Benign sample diagnosis	Abdominal, Pancreatitis, …, Simple benign liver cyst
6. Plasma CA19-9 U/ml	117, … ,1488
7. Creatinine mg/ml	18322, ... ,150423
8. LVYE1ng/ml	8932192, … ,8200958
9. REG1B ng/ml	5294884, … ,411938275
10. TFF1 ng/ml	654282174, … ,2021321078
11. REG1A ng/ml	1262, … ,13200

LVYE: Lymphatic vascular endothelial hyaluronan receptor 1 (LVYE1), REG18: Protein associated with pancreatic regeneration, TFF1: Urine Trefoil factor 1 (TFF1), PDAC: Pancreatic ductal adenocarcinoma cancer.

The problem of missing values is solved in RStudio ver.2021.09.2 using the CART technique in the *mice* package.[32,33] Logarithmic transformation was implemented on variables such as plasma CA19-9, creatinine, LVYE1, REG1B, TFF1, and REG1A. The data was then standardized, and outliers were removed from the set. Thus, the considerations on the dataset were continued with 585 observations and 11 variables. Frequencies of the categories of diagnosis, which is the categorical response variable in the study, have been reorganized 180, 206, 199, respectively. The generalized Shapiro-Wilk test for multivariate normality by Villasenor-Alva and Gonzalez-Estrada was applied in the *mvShapiroTest* package to learn more about the data and identify the methods to be used, it was found that the data were not distributed normally (Generalized Shapiro-Wilk test statistic value =0.9312, p-value < 0.01).[34]

The Spearman correlation was then computed among all variables, and the correlation matrix is shown in Table 2.

TABLE 2: Spearman correlation matrix of the variables.

	Age	Plasma CA19-9 U/ml	Creatinine mg/ml	LVYE1ng/ml	REG1B ng/ml	TFF1 ng/ml	REG1A ng/ml
Age	-	0.0981	-0.0403	0.1190***	0.1782***	0.1496***	0.2385***
Plasma CA19-9		-	-0.0138	0.0051	0.2168***	0.2631***	0.0295
Creatinine mg/ml			-	0.2703***	0.2254***	0.2611***	0.1190***
LVYE1ng/ml				-	0.3474***	0.3937***	0.1444***
					(2.2e-16)		
REG1B ng/ml					-	0.6610***	0.0530
TFF1 ng/ml						-	-0.0108
REG1A ng/ml							-

*** p-value<0.01
LVYE: Lymphatic vascular endothelial hyaluronan receptor 1 (LVYE1), REG18: Protein associated with pancreatic regeneration, TFF1: Urine Trefoil factor 1 (TFF1).

According to Table 2, the correlation coefficient between REG1B and TFF1 variables appears to have a moderate positive relationship of 0.6610 and is statistically significant at a level of 0.01. To ensure the degree of the relationship between variables, the variance inflation factor (VIF) values are also given in Table 3.

TABLE 3: VIF values of the variables.

Variables	VIF values
Age	20.53†
Plasma CA19-9 U/ml	3.33
Creatinine mg/ml	60.64†
LVYE1ng/ml	72.05†
REG1B ng/ml	33.95†
TFF1 ng/ml	34.09†
REG1A ng/ml	8.80

†VIF value>10
VIF: Inflation factor, LVYE: Lymphatic vascular endothelial hyaluronan receptor 1 (LVYE1), REG18: Protein associated with pancreatic regeneration, TFF1: Urine Trefoil factor 1 (TFF1).

Table 3 shows that variables with a VIF value greater than 10 have a multicollinearity problem. To cope with this problem, lasso, ridge, elastic net, adaptive lasso, adaptive elastic net, adaptive lasso, adaptive elastic net, adaptive lasso, adaptive elastic net, and adaptive group lasso regularization methods were used in the study. All analyses in the study were also carried out in RStudio ver.2021.09.2.

RESULTS

Except for the dependent variable, categorical variables were removed from the dataset for the regularization approaches to be used with logistic regression, and 8 variables and 585 observations were preprocessed. The dataset was then randomly divided into two parts: 65% training and 35% test.

To compare various machine learning methods, the accuracy, recall, precision, area under the ROC curve, Kappa, and F1 Score values were determined using the 10-fold cross-validation method.

Because the categories of the dependent variable were the control group, benign hepatobiliary disease, and PDAC, multinomial logistic regularization regression models were estimated in the study. The regularization parameter is chosen for all models by adding one standard error to the parameter value corresponding to the smallest 10-fold cross validated error.

TABLE 4: Logistic regularization regression coefficients for PDAC.

Variable	Logistic Lasso Coef.	Logistic Ridge Coef.	Logistic Elastic Net Coef.	Logistic Adaptive Lasso Coef.	Logistic Adaptive Elastic Net Coef.	Logistic Adaptive Group Lasso Coef.
Age	0.0562	0.0149	0.0254	.	0.0372	0.0011
Plasma CA19 19	.	-0.0364	.	.	.	-0.0397
Creatinine	-0.2195	0.0346	0.0529	-0.1740	-0.0025	.
LVYE1	0.2965	0.1369	0.1934	0.1829	0.1703	0.0578
REG1B	0.6879	0.1845	0.3036	0.5700	0.4046	0.3142
TFF1	.	0.1034	0.0841	.	.	.
REG1A	0.7781	0.1541	0.2639	0.6837	0.3318	0.1594

LVYE: Lymphatic vascular endothelial hyaluronan receptor 1 (LVYE1), REG18: Protein associated with pancreatic regeneration, TFF1: Urine Trefoil factor 1 (TFF1), PDAC: pancreatic ductal adenocarcinoma cancer.

Logistic lasso regression results

The minimum value of the lambda parameter is 0.0088 and the value after adding the one SE is 0.0427 in the logistic lasso regression model. These values are calculated at –4.7330 and -3.1536, respectively, using logarithmic transformations. Figure 1a has a "Log (λ) versus Multinomial Deviance" graphic showing the cross-validation error. The minimum lambda value and the value after adding the one SE are represented by vertical lines on the x-axis of the graph. As the penalty is applied, the number of variables remaining in the model is presented at the top of the graph. According to the value of log (λ) = –4.7330 selected in the model, it is seen that the coefficient of the two variables is zero and five variables remain in the model.

Table 4 shows the coefficients of the variables in the model based on the minimum log (λ) value chosen.

The coefficients of plasma CA19-19 and TFF1 are equal to zero, according to the logistic lasso regression model results in Table 4. In other words, they were excluded from the model by selecting variables. This implies that Plasma CA19-19 and TFF1 are not significant variables in the model.

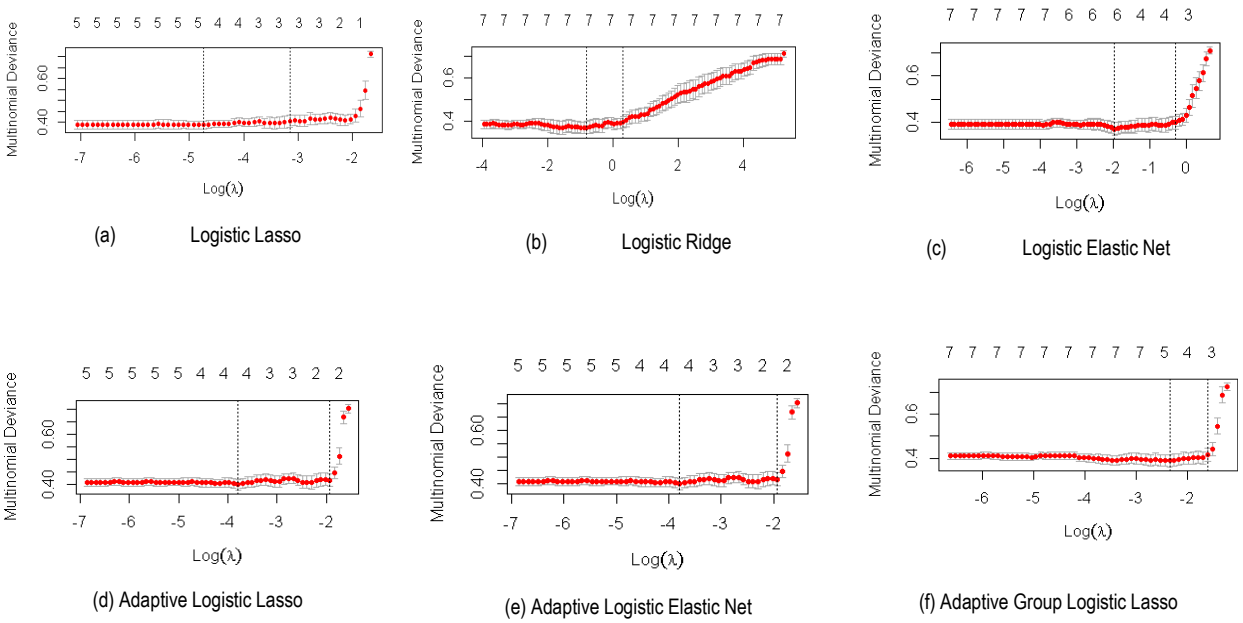

FIGURE 1: Log (λ) vs Multinomial Deviance Graphs.

Logistic ridge regression results

The minimum value of the parameter λ, which corresponds to the smallest value of the 10-folded CV error, was 0.4472 and this value was 1.3658 after applying one SE rule. These values were obtained using logarithmic transformations as –0.8047 and 0.3117, respectively. Figure 1b shows the "log (λ) versus Multinomial Deviance" graph showing the cross-validation error. The number of variables in the model has not decreased with the penalty applied, as can be seen from the top of the graph, and remains at 7, the number of variables for this dataset. Table 4 shows the coefficients of variables in the model based on the minimum log (λ). The coefficients of all variables differ from zero, as expected, according to the logistic ridge regression model results.

Logistic elastic net regression results

Logistic elastic net regression parameters, α and λ, are 0.38 and 0.1398, respectively, with one standard error value of 0.7460. These values were determined as –1.9675 and –0.2930, respectively, after logarithmic transformation. Figure 1c shows the " log (λ) vs multinomial deviance" graph with the cross-validation error. The number of variables in the model reduces with regularization, and the number of variables for this dataset is 6, as seen at the top of the graph. Table 4 shows the coefficients of the model's variables based on the minimum log (λ) value.

According to the results of the logistic elastic net regression model, the coefficient of the plasma CA19-19 variable is equal to zero. In other words, the less significant variable in the model implies that plasma CA19-19 was excluded from the model.

Logistic adaptive lasso regression results

For the multinomial logistic adaptive lasso regression model, the minimum value of the parameter λ is 0.0226 and this value is 0.1455 after applying one SE rule. After logarithmic transformation, these values were computed at -3,7882 and -1.9276, respectively. The log (λ) versus Multinomial Deviance graph with the CV error is given in Figure 1d.

Four variables remain in the model, according to log (λ) = –3.7882. Table 4 shows the coefficients of the variables in the logistic adaptive lasso model. According to the results of the logistic adaptive lasso regression model in Table 4, the coefficients of age, plasma CA19-19, and TFF1 are equal to zero. This means that the variables age, plasma CA19-19, and TFF1 are less significant in the model, but the variables REG1a and REG1b are important indicators of PDAC detection.

Logistic adaptive elastic net regression results

The study uses 10-fold CV to identify the optimum values (α and λ) of the multinomial logistic adaptive elastic net regression parameters, whereas logistic ridge regression is being used to compute the weights. The minimum value of the parameter λ is 0,0675 and this value is 0,5739 after applying one SE rule. These values were calculated utilizing a logarithmic transformation as –2.6956 and –0.5553, respectively. Figure 1e shows the log (λ) versus Multinomial Deviance graph with the CV error.

Six variables remain in the model, according to log (λ)= –2.6956. Table 4 shows the coefficients of the variables in the model.

According to the results of the adaptive logistic elastic net regression model in Table 4, plasma CA19-19 and TFF1 are equal to zero, indicating that the variables of less importance in the model are them. The results also demonstrate that the variables REG1A and REG1B are important in detecting PDAC.

Logistic adaptive group lasso regression results

The minimum value of the parameter λ is 0.0949 and its value is 0.1997 after applying one SE rule. When logarithmic transformations were performed, these values were calculated as –2.3549 and -1.6109, respectively. Figure 1f displays the cross-validation error as a graph of log (λ) versus multinomial deviance.

Based on the log (λ)=–2.3549, it is seen that five variables remain in the model. Table 4 shows the coefficients of variables in the model in terms of the minimal log (λ) value.

According to the results of the multinomial logistic adaptive group lasso regression model in Table 4, the coefficients of the variables creatinine and TFF1 are equal to zero. This implies that creatinine and TFF1 are less important variables in the model, and that the REG1A and REG1B variables are significant indicators in PDAC detection.

Validity and reliability results of PDAC

Some performance measures were computed to determine the validity and reliability of the models created using observations that were specified as a test set and covered 35% of the dataset.

The complexity matrix was used to determine accuracy, recall, precision, sensitivity, Kappa statistics, AUC, and F1-score values.

Table 5 shows the validity and reliability performance measures for all models that were employed in the analysis.

TABLE 5: Validity and reliability results for PDAC.

Measures	Logistic Lasso Regression	Logistic Ridge Regression	Logistic Elastic net Regression	Logistic Adaptive Lasso Regression	Logistic Adaptive Elastic Net Regression	Logistic Adaptive Group Lasso Regression
Accuracy	0.7244	0.7372	0.7488	0.7555	0.7405	**0.7621**
Recall	0.6154	0.6220	0.6310	**0.6538**	0.6329	0.6364
Precision	0.8333	0.8525	0.8667	0.8571	0.8480	**0.8879**
F1-Score	0.7080	0.7192	0.7303	**0.7418**	0.7248	**0.7414**
Kappa	0.4160	0.4740	0.4815	0.4748	0.4526	**0.5029**
AUC	0.6895	0.7240	0.7439	0.7360	0.7204	**0.7668**

PDAC: Pancreatic ductal adenocarcinoma cancer, AUC: Area under the curve.

The logistic adaptive group lasso regression has a better performance in predicting the likelihood of becoming PDAC, as shown in Table 5. Friedman test statistic, which was calculated to test for differences between results, was found H=25.905, p-value = 9.311e-05. At the 1% significance level, at least one of the model results is statistically different from the others. According to the Nemenyi test, there is a statistically significant difference at 1% level between logistic EN and logistic lasso, logistic adaptive group lasso and logistic lasso, logistic adaptive group lasso and logistic lasso, logistic adaptive group lasso and logistic ridge.

DISCUSSION

In this study, the regularized regression methods based on the logistic model were compared on a dataset in terms of various validity and reliability criteria.

Because the multicollinearity problem was identified among the variables in the dataset as a consequence of the study's preprocessing, regularized regression methods were chosen to address the problem as well as categorize disease diagnosis and perform variable selection. For this purpose, lasso, ridge, elastic net, adaptive lasso, adaptive elastic net, and adaptive group lasso regularization methods based on the logistic model were used in the study.

In addition, as the categories of disease diagnosis identified as dependent variables in the dataset had three categories with the control group, benign hepatobiliary disease, and PDAC, regularization methods based on the multinomial logistic model were used.

The dataset was randomly assigned 65% to the training set and the remaining 35% to the test set. With observations of 411 in the training set, classification models were trained as logistic lasso regression, logistic ridge regression, logistic elastic net regression adaptation, adaptive logistic elastic net regression, and adaptive group logistic lasso regression.

The performance of the models was assessed in terms of several criteria of validity and reliability on the test set containing 204 observation values after the regularization parameters were determined using cross-validation. Similar results in terms of AUC values from the performance measures are seen in the results of the study of Hsieh et al, and Lee et al.[16,17] Moreover, the literature on the use of logistic regression backs up the findings of the study.[14-17] The difference of this study from previous studies is that build models to predict PDAC diagnosis with regularization techniques that take into account the presence of multicollinearity in logistic regression.

CONCLUSION

REG1B, REG1A, and LVYE1 were shown to be statistically significant factors based on the results of PDAC diagnosis in all regularization models in the classification of PDAC detection of the three-level dependent variable. The logistic adaptive group lasso regression model performs better in PDAC diagnosis when compared to the estimated model results for the urine biomarkers dataset. Age and plasma CA19 19 are identified as key variables in PDAC diagnosis in this model, in addition to REG1A, REG1B, and LVYE1. Additionally, except for ‘Recall’, the logistic adaptive group lasso regression model outperformed all other performance measures.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Özlem Alpu, Güven Pekdemir;* ***Design:*** *Özlem Alpu, Güven Pekdemir;* ***Control/Supervision:*** *Özlem Alpu;* ***Data Collection and/or Processing:*** *Güven Pekdemir;* ***Analysis and/or Interpretation:*** *Özlem Alpu, Güven Pekdemir;* ***Literature Review:*** *Güven Pekdemir, Özlem Alpu;* ***Writing the Article:*** *Özlem Alpu, Güven Pekdemir;* ***Critical Review:*** *Özlem Alpu;* ***References and Fundings:*** *Özlem Alpu, Güven Pekdemir;* ***Materials:*** *Güven Pekdemir.*

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-49. [Crossref] [PubMed]
2. Mayerle J, Kalthoff H, Reszka R, Kamlage B, Peter E, Schniewind B, et al. Metabolic biomarker signature to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis. Gut. 2018;67(1):128-37. Erratum in: Gut. 2018;67(5):994. [Crossref] [PubMed] [PMC]
3. Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD, Plješa-Ercegovac M, et al. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: a case-control study. PLoS Med. 2020;17(12):e1003489. [Crossref] [PubMed] [PMC]
4. Shimizu Y, Yasui K, Matsueda K, Yanagisawa A, Yamao K. Small carcinoma of the pancreas is curable: new computed tomography finding, pathological study and postoperative results from a single institute. J Gastroenterol Hepatol. 2005;20(10):1591-4. [Crossref] [PubMed]
5. Sarantis P, Koustas E, Papadimitropoulou A, Papavassiliou AG, Karamouzis MV. Pancreatic ductal adenocarcinoma: treatment hurdles, tumor microenvironment and immunotherapy. World J Gastrointest Oncol. 2020;12(2):173-81. [Crossref] [PubMed] [PMC]
6. Sanoob M, Madhu A, Ajesh K, Varghese SM. Artificial neural network for diagnosis of pancreatic cancer. Int J Cybernet Inform. 2016;5:41-9. [Crossref]
7. Blyuss O, Zaikin A, Cherepanova V, Munblit D, Kiseleva EM, Prytomanova OM, et al. Development of PancRISK, a urine biomarker-based risk score for stratified screening of pancreatic cancer patients. Br J Cancer. 2020;122(5):692-6. [Crossref] [PubMed] [PMC]
8. Ge G, Wong GW. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. BMC Bioinformatics. 2008;9:275. [Crossref] [PubMed] [PMC]
9. Marengo E, Robotti E, Cecconi D, Scarpa A, Righetti PG. Application of fuzzy logic principles to the classification of 2D-PAGE maps belonging to human pancreatic cancers treated with Trichostatin-A. 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542). 2004;1:359-64. [Link]
10. Torshabi AE, Riboldi M, Pella A, Negarestani A, Rahnema M, Baroni G. A clinical application of fuzzy logic. In: Dadios EP, ed. Emerging Technologies and Applications. 1st ed. Rijeka, Croatia: InTech; 2012. p.3-18. ISBN:978-953-51-0337-0 [Crossref] [PubMed]
11. Jiang W, Shen Y, Ding Y, Ye C, Zheng Y, Zhao P, et al. A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system. Int J Cancer. 2018;142(2):357-68. Erratum in: Int J Cancer. 2018;143(1):E2. [Crossref] [PubMed]
12. Sadewo W, Rustam Z, Hamidah H, Chusmarsyah AR. Pancreatic cancer early detection using twin support vector machine based on kernel. Symmetry. 2020;12(4):667. [Crossref]
13. Chang C-L, Hsu M-Y. The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer, Expert Systems with Applications. 2009;36(7):10663-72. [Crossref]
14. Baecker A, Kim S, Risch HA, Nuckols TK, Wu BU, Hendifar AE, et al. Do changes in health reveal the possibility of undiagnosed pancreatic cancer? Development of a risk-prediction model based on healthcare claims data. PLoS One. 2019;14(6):e0218580. [Crossref] [PubMed] [PMC]
15. Appelbaum L, Cambronero JP, Stevens JP, Horng S, Pollick K, Silva G, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: an observational study. Eur J Cancer. 2021;143:19-30. [Crossref] [PubMed]
16. Hsieh MH, Sun LM, Lin CL, Hsieh MJ, Hsu CY, Kao CH. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Cancer Manag Res. 2018;10:6317-24. [Crossref] [PubMed] [PMC]
17. Lee HA, Chen KW, Hsu CY. Prediction model for pancreatic cancer-a population-based study from NHIRD. Cancers (Basel). 2022;14(4):882. [Crossref] [PubMed] [PMC]
18. Cho S-B, Won H-H. Machine learning in DNA microarray analysis for cancer classification. Proceedings of the First Asia-Pacific Bioinformatics Conference. 2003;34:189-98.
19. Hoerl A, Kennard R. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55-67. [Crossref]
20. Schaefer RL, Roi LD, Wolfe RA. A ridge logistic estimator. Communications in Statistics: Theory and Methods. 1984;3:99-113. [Crossref]
21. Çiftsüren MN, Akkol S. Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, lasso and elastic net. Archives Animal Breeding. 2018;61(3):279-84. [Crossref]
22. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996;58(1):267-88. [Crossref]
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67:301-20. [Crossref]
24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1-22. [Crossref] [PubMed] [PMC]
25. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006;101(476):1418-29. [Crossref]
26. Algamal ZY, Lee MH. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Comput Biol Med. 2015;67:136-45. [Crossref] [PubMed]
27. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics. 2009;37(4):1733-51. [Crossref] [PubMed] [PMC]
28. Ghosh S. On the grouped selection and model complexity of the adaptive elastic net. Stat Comput. 2011;21:451-62. [Crossref]
29. Wang H, Leng C. A note on adaptive group lasso. Computational Statistics and Data Analysis. 2008;52(12):5277-86. [Crossref]
30. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006;68(1):49-67. [Crossref]
31. Yang X, Tong Y, Meng X, Zhao S, Xu Z, Li Y, et al. Adaptive logistic group Lasso method for predicting the no-reflow among the multiple types of high-dimensional variables with missing data. 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). 2016;1085-9. [Crossref]
32. RStudio [Internet]. Copyright © 2022 RStudio, PBC [Accessing Date:19.04.2022]. Available from: [Link]
33. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1-67. [Crossref]
34. Gonzalez Estrada E, Villasenor-Alva JA. mvShapiroTest: Generalized Shapiro Wilk test for multivariate normality. R package version 1.0. 2013.