Penalized Logistic Regression Model to Predict a Results of RT-PCR by Using Blood Laboratory Test

A Dwinata\textsuperscript{1,2}, K A Notodiputro\textsuperscript{2} and B Sartono\textsuperscript{2}*,

\textsuperscript{1}Mathematics Education Study Program, Raja Ali Haji Maritime University, Indonesia
\textsuperscript{2}Department of Statistics, IPB University, Indonesia

*Corresponding author: bagusco@apps.ipb.ac.id

Abstract. Statistical modelling to determine the effect of several predictor variables on the binary response variable is known as multiple logistic regression model. The addition of a penalty function to the model is done to improve prediction accuracy. Penalized logistic regression shrinks the regression coefficient to zero. This penalized logistic regression model will be used to predict a result of RT-PCR by using the features of blood laboratory tests. This research uses LASSO and elastic net penalties function. This study aims to determine the prediction performance of the RT-PCR test using logistic regression with LASSO and elastic net penalties. The data from the RT-PCR test were used as the binary response variable. Patient age quantile and 27 features of laboratory blood test were used as predictor variables. The results of this research showed that prediction performance of a RT-PCR test using LASSO logistic regression was better than elastic net logistic regression. The LASSO logistic regression model had a good performance for predicting the RT-PCR test with 88% accuracy and 93% AUC. Based on the result of LASSO logistic regression model, the features of laboratory blood tests that affect a RT-PCR test were leukocytes, basophils, RDW and C-reactive protein.

1. Introduction
Statistical modelling to determine the effect of several predictor variables on the binary response variable is known as multiple logistic regression model. Multiple logistic regression coefficients are estimated using the maximum likelihood method [1]. The more predictor variables used in modelling will increase the chances of multicollinearity between the predictor variables. Multicollinearity tends to increase the variance of regression coefficients [1]. The addition of predictor variables that are not related to the response variable in modelling will cause overfitting problem [2]. So, prediction error will increase if the predictor variables in modelling are not related to the response variables. The addition of a penalty function to a model is done to improve a prediction accuracy [2]. Adding a penalty to the logistic regression model is done by adding the penalty function to the log likelihood function. Penalized logistic regression shrinks the regression coefficient to zero. There are three types of penalties, namely ridge regression, Least Absolute Shrinkage and Selection Operators (LASSO) and Elastic Net [3].

Ridge regression use $L_2$ penalty that is $\lambda \sum_{j=1}^{p} \beta_j^2$, LASSO use $L_1$ penalty that is $\lambda \sum_{j=1}^{p} |\beta_j|$ and elastic net combines $L_1$ and $L_2$ penalties that is $\lambda \sum_{j=1}^{p} (1 - \alpha) \beta_j^2 + \alpha |\beta_j|$. In general, ridge regression involves all predictor variables in the model, but the regression coefficient value is smaller than a model without penalty [2]. LASSO reduces overfitting and also selects predictor variables [4]. Elastic
net uses the ridge regression penalty to handle the problem of high correlation and the LASSO penalty to do variables selection [3]. The shrinkage parameter (\(\lambda\)) directly regulates the bias and variance and regulates the shrinkage strength of the regression coefficient. The value of \(\lambda\) can be calculated using cross validation [1]. The parameter \(\alpha\) on the elastic net adjusts for the penalty, when \(\alpha = 0\) a ridge regression penalty is obtained and when \(\alpha\) approaches 1, a LASSO penalty is obtained. Elastic net penalty which is a linear combination of the ridge regression penalty and LASSO is used as a penalty function in the R programming system with the glmnet package [5]. The solution of the elastic net log likelihood function is obtained using descent coordinates [3]. The addition of LASSO and elastic net penalties can shrink the regression coefficients and select variables in modeling. The correlation between the response variable and certain predictor variables can also depend on the values of the other predictor variables. When important covariates are associated with correlated variables, the simulation results show that the predicted performance of LASSO and elastic net increases [6]. The logistic regression model with LASSO and elastic net penalty have been performed on gene expression data of patients that suffer colorectal cancer and hepatocellular carcinoma [7]. The results of this study have a high accuracy.

In this study, penalized logistic regression models will be used to predict a result of RT-PCR by using the features of blood laboratory tests. The RT-PCR stands for Reverse Transcription Polymerase Chain Reaction. The RT-PCR is the most reliable test tool for coronavirus disease (COVID-19). The COVID-19 is spreading rapidly from China to countries in the world. The transmission of COVID-19 has exponential growth. The true level of spread of COVID-19 is not yet known due to various limitations in conducting the COVID-19 test. The RT-PCR test takes a long time approximately three until four days to find out the results. While, the status of a patient infected with COVID-19 or not is quickly required for prevention of transmission and clinical treatment measures. Apart from the RT-PCR test, a rapid diagnostic test can provide results in a short time. However, rapid diagnostic tests are very susceptible to accuracy problems so that the tests performed are not efficient. Research to overcome the limitations of the COVID-19 test kit is still ongoing. Models using routine laboratory test results offer opportunities for early and rapid identification of patients infected with COVID-19 [8]. The blood laboratory test results can be known within two to three hours. The blood laboratory test results from a patient include many features. This research interested in selecting the features of blood test laboratory results using logistic regression with LASSO and elastic net penalties. So, the logistic regression model with LASSO and elastic net penalties can be used to obtain a simpler model. This research aims at the model performance of the RT-PCR test result predictions by using the blood laboratory test.

2. Data and Research Method

2.1. Data

The data in this study were secondary data published by Einstein Data4u on March 28, 2020 which can be accessed via https://www.kaggle.com/einsteindata4u/covid19. This data was the RT-PCR test results and laboratory test results from the patient at the Israelita Albert Einstein Hospital, in São Paulo, Brazil. Many variables in the data contain missing values. Before this data was used, the data cleaning process was carried out. Firstly, we deleted data on patients who did not perform laboratory tests. The data cleaning process was continued by deleting variables containing more than 80% missing values. At the final stage of the data cleaning process, variables with zero diversity were not used. After performing the data cleaning process, we obtained the data on 75 patients who performed RT-PCR tests and laboratory tests. This study used 28 predictor variables in modelling, patient age quantile and 27 variables of blood laboratory test results. The features of blood laboratory test results were the haematocrit, haemoglobin, platelets, mean platelet volume, red blood cells, lymphocytes, mean corpuscular haemoglobin concentration (MCHC), leukocytes, basophils, mean corpuscular haemoglobin (MCH), eosinophils, mean corpuscular volume. (MCV), monocytes, red blood cell
distribution width (RDW), neutrophils, urea, C-reactive protein, creatinine, potassium, sodium, pCO₂ (venous blood gas analysis), Hb saturation (venous blood gas analysis), base excess (venous blood gas analysis), pO₂ (venous blood gas analysis), total CO₂ (venous blood gas analysis), pH (venous blood gas analysis), HCO₃ (venous blood gas analysis).

2.2. Research Method

The results of the RT-PCR test were currently very reliable for examining a person who was infected by COVID-19. The data from the RT-PCR test were used as the binary response variable. A positive RT-PCR test result was categorized as one and a negative RT-PCR test result was categorized as zero. The predictor variables were the patient age quantile and 27 features of blood laboratory test results from patients who performed a RT-PCR test. This study used a logistic regression model with LASSO and Elastic Net penalties to select the variables that affect the RT-PCR test results. Penalized logistic regression model with LASSO and elastic net penalties used the R programming system with the glmnet package. LASSO logistic regression modelling with α = 1 was carried out in the following steps:

a. the data was divided based on the leave one out cross validation procedure, 80% as modeling data and 20% as validation data
b. we determined optimal λ (shrinkage parameter) with 10-fold cross validation
c. we estimated the coefficient β using modelling data
d. we performed analysis on the predictor variables of modelling results

Furthermore, we conducted elastic net logistic regression model with 0 < α < 1, which was carried out in the following steps:

a. the data was divided based on the leave one out cross validation procedure, 80% as modeling data and 20% as validation data
b. we determined the optimal α and λ with 10-fold cross validation
c. we estimated the coefficient β using modelling data
d. we performed analysis on the predictor variables of modelling results

This study also investigated the selection of the best model to predict the RT-PCR test results. Selection of the best model was based on accuracy and AUC performances. The steps for selecting the best model were as follows:

a. we performed logistic regression modelling using LASSO and Elastic Net penalties with 30 repetitions on modelling data. A sample taken as modelling data and validation data were arranged as one observation. So, there were 30 pairs of modelling and validation data
b. we measured model performance with accuracy and AUC values using validation data for 30 repetitions
c. we performed the mean difference test for accuracy and AUC values with paired sample t-test
d. we determined the best model to predict the results of a RT-PCR test

3. Results and Discussion

3.1. Results

3.1.1. LASSO Logistic Regression Model

The data of the patient was randomly partitioned. They were 61 patient data (80%) as modelling data and 14 patient data (20%) as validation data. LASSO logistic regression required λ (shrinkage parameter) to shrink the regression coefficient. The optimum λ was obtained by 10-fold cross-validation.
Figure 1. Cross validation plot for optimizing the LASSO shrinkage parameter

Figure 1 shows the binomial deviance value for each log $\lambda$. The vertical line showed the optimum log $\lambda$. The first vertical line represented the value of $\lambda_{\text{min}}$ which was 0.05 and the second vertical line represented the value of $\lambda_{\text{1se}}$ which was 0.13. The optimum $\lambda$ selected in this study when the binomial deviance was minimum. So, $\lambda_{\text{min}}$ was used as the shrinkage parameter in the LASSO logistic regression model. The value of the upper horizontal line from Figure 1 shows the number of LASSO coefficients with nonzero values for each log $\lambda$. By using the $\lambda_{\text{min}}$ value, there are four LASSO coefficients with nonzero values. The predictor variables with non-zero LASSO coefficient were shown in Table 1.

Table 1. LASSO logistic regression coefficient

| Variables                                    | Coefficient |
|----------------------------------------------|-------------|
| Leukocytes                                   | -0.9116     |
| Basophils                                    | -0.4482     |
| Red blood cell distribution width (RDW)      | -0.0015     |
| C-reactive Protein                           | 0.3936      |

The plot of the LASSO logistic regression coefficient spread for each log $\lambda$ can be seen in Figure 2. All values of the LASSO logistic regression coefficient go to zero as the increases of the shrinkage parameter.

Figure 2. Plot of LASSO logistic regression coefficients for each shrinkage parameter
3.1.2. **Elastic Net Logistic Regression Model**

The data of the patient was randomly partitioned. They were 61 patient data (80%) as modelling data and 14 patient data (20%) as validation data. The optimum \( \alpha \) in the modelling data was obtained by 10-fold cross-validation based on the minimum CV. Based on modelling data, the elastic net logistic regression analysis used \( \alpha = 0.3 \). Elastic net logistic regression required \( \lambda \) (shrinkage parameter) to shrink the regression coefficient.

![Cross validation plot for optimizing the elastic net shrinkage parameter](image)

Figure 3. Cross validation plot for optimizing the elastic net shrinkage parameter

Figure 3 shows the binomial deviance value for each log \( \lambda \). The vertical line showed the optimum log \( \lambda \). The first vertical line represented the value of \( \lambda_{\text{min}} \) which was 0.096 and the second vertical line represented the value of \( \lambda_{1\text{se}} \) which was 0.43. In this study, the minimum binomial deviance value was obtained when log \( \lambda_{\text{min}} \). So, the value of \( \lambda_{\text{min}} \) was used as a parameter shrinkage of elastic net logistic regression. The upper horizontal line in Figure 3 showed the number of elastic net coefficients with nonzero values for each log \( \lambda \). By using the log \( \lambda_{\text{min}} \), there are 11 elastic net logistic regression coefficients which have non-zero values. The predictor variables with non-zero elastic net coefficient were shown in Table 2.

| Variables                                      | Coefficients |
|------------------------------------------------|--------------|
| Leukocytes                                     | -0.5462      |
| Basophils                                      | -0.3504      |
| Platelets                                      | -0.1969      |
| Red blood cell distribution width (RDW)        | -0.1693      |
| Eosinophils                                    | -0.0892      |
| Potassium                                      | -0.0674      |
| Urea                                           | -0.0550      |
| Mean corpuscular volume (MCV)                  | -0.0093      |
| pH (venous blood gas analysis)                 | 0.0168       |
| Monocytes                                      | 0.1028       |
| C-reactive Protein                             | 0.3241       |

The plot of the elastic net logistic regression coefficient spread for each log \( \lambda \) can be seen in Figure 4. All values of the elastic net logistic regression coefficient go to zero as the increases of the shrinkage parameter.
3.1.3. Selection of the best model

The samples of modelling data and validation data for 30 repetitions was carried out with the set.seed setting during the sampling process. So, there are 30 sets of modelling data and validation data that would be used to predict the results of a RT-PCR test. The sampling process was randomly carried out. They were 61 patient data or 80% as modelling data and 14 patient data or 20% as validation data. LASSO logistic regression model used $\alpha = 1$, while the elastic net logistic regression used the optimum $0 < \alpha < 1$. The optimum $\alpha$ was obtained based on the modeling data taken for each repetition.

Furthermore, LASSO and elastic net logistic regression modelling was carried out for each repetition. Modelling performance of LASSO and elastic net logistic regression were carried out using the accuracy and AUC values based on the validation data. The comparison of accuracy and AUC values of 30 repetitions for each model was shown in Figure 5 and Figure 6.

![Figure 4](image4.png)

**Figure 4.** Plot of elastic net logistic regression coefficients for each shrinkage parameter

![Figure 5](image5.png)

**Figure 5.** Accuracy value of LASSO and elastic net logistic regression model
The accuracy average for logistic regression with LASSO and elastic net penalties were 0.88 and 0.86, respectively. The mean AUC values for the LASSO and elastic net logistic regression were 0.93 and 0.92, respectively. The criteria for selecting the best model was the model with the greatest accuracy and AUC value. Based on the average of the two goodness criteria of the model, the LASSO logistic regression model was better than the elastic net logistic regression model.

To state that the LASSO logistic regression is better than the Elastic Net logistic regression statistically, the mean difference test was carried out using paired sample t-test. The results of the paired sample t-test for the two models used could be seen in Table 3. The hypothesis was used to see the differences significantly in accuracy/AUC of the two models. The hypothesis was as follows:

\[ H_0: \text{the accuracy/AUC average of the LASSO logistic regression was less than or equal to the accuracy/AUC average of the Elastic Net logistic regression.} \]

\[ H_1: \text{the accuracy/AUC average of the LASSO logistic regression was more than the accuracy/AUC average of the Elastic Net logistic regression.} \]

Table 3. The results of the paired sample t-test of accuracy and AUC

|        | t-stat | p-value |
|--------|--------|---------|
| Accuracy | 2.8045 | 0.0045  |
| AUC     | 1.8810 | 0.0350  |

The t-test result in Table 3 showed the p-value for accuracy and AUC less than 0.05 at significance level 5%. It could be concluded that the accuracy/AUC average of LASSO logistic regression was more than the accuracy/AUC average of elastic net logistics regression. So, the LASSO logistic regression model had a better performance for predicting the results of the RT-PCR test.

3.2. Discussion

The logistic regression analysis with LASSO and elastic net penalties could select 28 predictor variables that were used to predict the results of the RT-PCR test. Figure 2 and Figure 4 showed that the LASSO logistic regression places more exact regression coefficients at zero than the elastic net logistic regression. The predictor variables selected by LASSO penalty in Table 1 showed that four predictor variables that affect the results of RT-PCR. They are leukocytes, basophils, RDW and C-reactive protein. C-reactive protein had a positive coefficient. It means that the higher the C-reactive Protein level, the more likely a person is to get positive RT-PCR test result. While leukocytes,
basophils and RDW had a negative coefficient. It can be interpreted that the less a number of leukocytes, basophils and RDW caused the tendency to get positive RT-PCR test results increases.

The results of the Elastic Net logistic regression analysis have the same four predictor variables as the LASSO penalty, the other seven predictor variables that influence the prediction of RT-PCR results are platelets, eosinophil, potassium, urea, MCV, pH (venous blood gas analysis) and monocytes. C-reactive protein, monocytes and pH (venous blood gas) had a positive coefficient. It means the higher the pH (venous blood gas analysis), the number of monocytes and the C-reactive protein level caused the more likely a person is to get a positive RT-test results. Conversely leukocyte, basophil, platelets, RDW, eosinophils, potassium, urea and MCV had a negative coefficient. It means that the smaller the number of leukocytes, basophil, platelets, RDW, eosinophils, potassium, urea and MCV caused the tendency to obtain a positive RT-PCR test results increases.

In the early stages of a patient infected with COVID-19, previous research has explained that the patient's C-reactive protein levels have a positive correlation with lung lesions and can reflect the severity of the disease [9]. The higher the level of C-reactive protein the patient has, the more lung lesions will be so that the patient's disease gets worse. Patients with COVID-19 pneumonia have a statistically different number of leukocytes and platelets from non-COVID-19 pneumonia patients [10]. The results of this study also showed that decreasing the number of leukocytes and platelets can increase the probability of a person getting a positive RT-PCR test. Venous blood gas analysis provides useful information regarding lung function and acid-base balance [11]. The higher the pH (venous blood gas analysis), the more likely it is to get a positive RT-PCR test. So, if there is an increase in pH it can be an indicator of impaired lung function. MCV is a measure of the average size of red blood cells, if too low, it can indicate that the patient has some type of anaemia. The lower the MCV value, the more likely it is to get a positive RT-PCR test result. Urea and potassium are mineral substances that are essential for organs such as the heart or kidneys to function normally. The lower the urea and potassium levels in a person's body, there will be interference with heart or kidney function.

The results of this research showed that prediction performance of a RT-PCR test using LASSO logistic regression was better than elastic net logistic regression. The LASSO logistic regression model had a good performance for predicting the RT-PCR test with 88% accuracy and 93% AUC. Based on the result of LASSO logistic regression model, the features of laboratory blood tests that affect a RT-PCR test were leukocytes, basophils, RDW and C-reactive protein. The odds ratio of leukocytes was 0.4. It meant that the odds of a positive RT-PCR test result given an increase of a unit leukocytes (10^3/μl) 0.4 times was lower than before the increase. The odds ratio of basophils was 0.64. It meant that the odds of a positive RT-PCR test result given an increase in 1% basophils 0.64 times was lower than before the increase. The odds ratio of RDW was 0.99. It meant that the odds of a positive RT-PCR test result given an increase of 1% RDW 0.99 times was lower than before the increase. The odds ratio of C-reactive protein was 1.48. It meant that the odds of a positive RT-PCR test result given an increase in one milligram of C-reactive protein per liter of blood (mg/L) 1.48 times was higher than before the increase.

4. Conclusion

Selection of variables to predict the results of the RT-PCR test using LASSO logistic regression resulted four predictor variables with non-zero regression coefficients, namely Leukocytes, Basophils, Red blood cell distribution width (RDW) and C-reactive protein. The selection of variables to predict the results of the RT-PCR test using elastic net logistic regression resulted eleven predictor variables with non-zero regression coefficients, namely platelets, eosinophils, mean corpuscular volume (MCV), monocytes, urea, potassium, pH (venous blood gas analysis) and four other predictor variables were equal to the predictor variables of the LASSO logistic regression model outcome.
The results of the model goodness measurement analysis showed that the LASSO logistic regression model is better than the elastic net logistic regression model for predicting the RT-PCR test results. The laboratory results of blood tests as predictor variables have high accuracy and are fast to predict the results of the RT-PCR test.

References

[1] James G, Witten D, Hastie T and Tibshirani R 2013 An Introduction to Statistical Learning with Applications in R New York: Springer
[2] Hastie T, Tibshirani R and Friedman J 2008 Springer Series in Statistics New York: Springer
[3] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net J. R. Stat. Soc. Ser. B Stat. Methodol. 67 301–320
[4] Tibshirani R 2011 Regression shrinkage and selection via the lasso: A retrospective J. R. Stat. Soc. Ser. B Stat. Methodol. 73 273–282
[5] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent J. Stat. Softw. 33 1–22
[6] Sirimongkolkasem T and Drikvandi R 2019 On Regularisation Methods for Analysis of High Dimensional Data Ann. Data Sci. 6 737–763
[7] Susanti A 2019 Perbandingan Metode Regresi Logistik Terpenalti pada Data Ekspresi Gen [thesis]. Institut Pertanian Bogor
[8] Yang H S, Hou Y, Vasovic L V, Steel P, Chadburn A, Racine-Brzostek S E, Velu P, Cushing M M, Loda M, Kaushal R, Zhao Z and Wang F 2020 Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning Clin. Chem. 66 (11) 1396-1404
[9] Wang L 2020 C-reactive protein levels in the early stage of COVID-19 Med. Mal. Infect. 50 332–334
[10] Cheng Z, Lu Y, Cao Q, Qin L, Pan Z, Yan F and Yang W 2020 Clinical features and chest CT manifestations of coronavirus disease 2019 (COVID-19) in a single-center study in Shanghai, China Am. J. Roentgenol. 215 121–126
[11] Rieser T M 2013 Arterial and Venous Blood Gas Analyses Top. Companion Anim. Med. 28 86–90