Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 1 (2023) 681 - 687 ISSN Media Electronic: 2580-0760

Logistic Regression Using Hyperparameter Optimization on COVID-19 Patients' Vital Status

Vinna Rahmayanti Setyaning Nastiti[1], Yufis Azhar[2], Riska Septiana Putri[3]
[1,2,3]Informatics Engineering, Faculty of Engineering, Universitas Muhammadiyah Malang
[1]vinastiti@umm.ac.id, [2]yufis@umm.ac.id, [3]riskaseptianaputri@webmail.umm.ac.id

Abstract

This study aims to classify COVID-19 patients based on the results of their hematology tests. Hematology test results have been shown to be useful in identifying the severity and risk of COVID-19 patients. Specifically, this study focuses on classifying COVID-19 patients based on their vital status, namely Deceased and Alive. The dataset used in this study contains four variables: white blood cells (WBC), neutrophils (NEU), lymphocytes (LYM), and Neutrophil Lymphocyte Ratio (NLR). Logistic Regression algorithm was used to solve the problem, and hyperparameter optimization was implemented to obtain the best model performance. The objective of this study was to build the best parameter in classifying the patients' vital status. The proposed model achieved an accuracy score of 78%, which is the best performance among the tested models. The results of this study provide a key component for decision making in hospitals, as it provides a way to quickly and accurately identify the vital status of COVID-19 patients. This study has important implications for managing the COVID-19 pandemic and should be of interest to researchers and practitioners in the field.

Keywords: logistic regression; hyperparameter optimization; covid-19; patients status

1. Introduction

Corona virus disease 2019 (COVID-19) is a syndrome of corona virus-2 (SARS-COV-2) which affects the respiratory system. This disease has spread in more than 200 countries around the world [1]. The World Health Organization (WHO) stated that this disease has infected more than 63 million people and caused more than 1,466,000 deaths [2]. COVID-19, SARS and MERS are the third highly pathogenic coronavirus to emerge in the last two decades [3]. The symptoms of COVID-19 include asymptomatic infections, upper respiratory tract infections, and gastrointestinal infections. A blood test is a diagnostic tool to measure the severity and predict the patient's risk of COVID-19 infections [4]. Wang [5] found that peripheral blood neutrophils affected by cytokines increase the severity of infection in COVID-19 patients. A study by Potempa [6] mentioned the number of White Blood Cells (WBC) and Neutrophil Lymph Ratio (NLR) are the indicators of systemic inflammatory response. NLR is the calculation of the total ratio of neutrophils and lymphocytes on a hematology test. NLR is a simple inflammatory biomarker in hematology test [7]. NLR can be used as an early warning signal for patients with severe COVID-19 as well as a marker of poor clinical outcome [8] and can be used as an indicator of severity risk factors and predict the inflammatory status of COVID-19 patients. In addition, several studies found an increase in the number of WBC and lymphocytes in COVID-19 patients. Lymphocytes play an important role in maintaining immune homeostasis in the body infected by virus [9]. WBC differential may indicate any change in the body first immunological mechanism. In COVID-19 cases, the WBC counts higher than 6.16 x 109/L should receive a serious treatment [10].

WBC and NLR differential counts have been studied by many researchers around the world. Studies [11]- [16] found that NLR and WBC can be used as severity parameters of COVID-19. A study [11] in Wuhan, China observed 70% of COVID-19 patients were having drastic reduction in the number of lymphocytes. A study in Pennsylvania [13] concluded that NLR can be used as the biomarker of multiple-organ failure or death in COVID-19 cases. An increase in WBC and neutrophils and a decrease in lymphocytes in a patient indicate high severity [17], [18]. According to the Ministry of Health, the confirmed COVID-19 cases in Indonesia has reached 6,653,469 cases. Many studies on COVID-19 in Indonesia have been conducted by institutions and researchers. Mus [4] conducted a literature study on laboratory tests for COVID-19 patients. The results showed that NLR could be used as

Accepted: 16-01-2023 | Received in revised: 14-05-2023 | Published:02-06-2023

a risk marker for COVID-19. Several studies in Indonesia have classified the variables that influence COVID-19 using logistic regression. Logistic regression was used to model the relationship between the dependent variable and the independent variable, both continuous and categorical. Romadhon [19] compared Naïve Bayes, Logistic Regression, and kNN method. These three models were used to identify the age, gender, and province of the patient which were closely related to the spread of COVID-19. Wibowo [20] showed that logistic regression can be used to identify the model of the COVID-19 patients in Java island and prediction model for the cure rate with high accuracy. Some studies [21]-[23] used NLR as the variable in a logistic regression. They found that NLR significantly influenced the mortality rate of COVID-19 patients. Another study mentioned WBC can be used as the variable during early indication of COVID-19 cases [24]. Logistic regression with binary classification in this study obtained an accuracy score of 88% with 10 different features. Logistic regression was employed as the method of this study. Cai [25] implemented logistic regression by using NLR, LDH, D-Dimer, and CT for initial prediction of COVID-19 patients in Hubei, China. The evaluation used was the significance test of the logistic regression with p-value. The result of logistic regression with four tested indicators was statistically significant. Therefore, the statements of the problem in this study were: Can logistic regression classify patients' vital status based on the mortality? How accurate is the classification of COVID-19 patient's vital status using logistic regression?

Based on conclusion of the researchers, the aims of the study is classifiying patient's vital status using logistic regression. The logistic regression in this study was developed using SMOTE technique and hyperparameter grid search to increase the accuracy of the model. The patient dataset used were derived from UMM Hospital in Malang. This study was conducted to predict the emergency initial indications of patients infected with COVID-19 based on the results of the hematology test. The significance of this study can help doctors to improve the management through early prediction the mortality risk for COVID-19 patients in countries with low resources based on the hematology test of patients.

2. Research Methods

Figure 1 shows the overall implementation of logistic regression that was being used. The discusssion comprised of the dataset, data preprocessing, model training, and evaluation.

2.1 Dataset

This study used the data of COVID-19 patients derived from UMM Hospital. This dataset was the observation data on the patient's initial hematology test. The dataset had four variables with a total of 623 patients being observed. Dataset was obtained from the results of the hematology COVID-19 patients in January 2021 – June 2021. Dataset was collected using blood sampling test for COVID-19 patient. The observation show that patients infected COVID-19 with the delta variants.

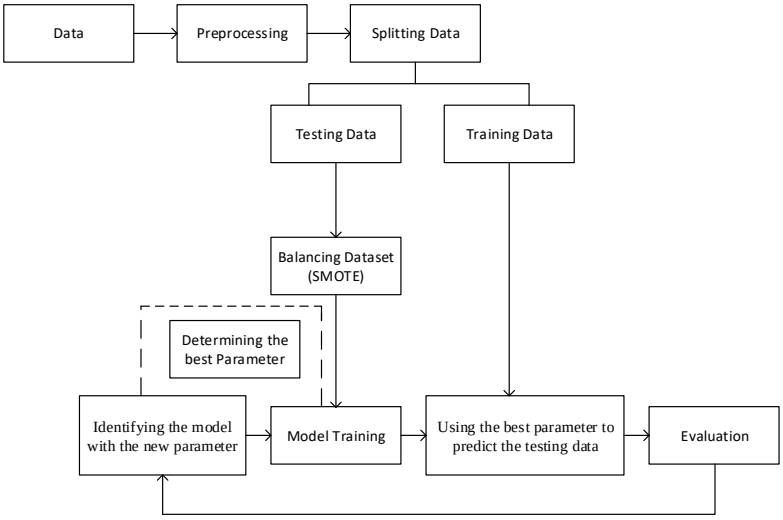


Figure 1. The implementation of Logistic Regression

A study reported that the delta variant grew more rapidly and at higher levels inside people's lungs and throats than did earlier versions of the virus. The variables identified were patients' vital status, WBC, neutrophil, and lymphocyte count in patients infected with COVID-19. All data variables related to patient and hospital identities had been deleted to maintain data privacy.

2.2 Data Preprocessing

Data preprocessing is a step to transform the data before it is processed into a dataset. This stage is required in order to obtain more balanced, precise, and good data

DOI: https://doi.org/10.29207/resti.v7i34868.xxx

quality. In this study, this stage included handling the missing data, splitting datasets, and data normalization. The missing data were identified by exploring the missing data in COVID-19 patient data. The missing data in lymphocytes, neutrophils, and WBC counts were completed by inserting median score into the data. The dataset was split into two parts, namely training data 80% and testing data 20%. Training data was used to train the model to obtain quality data, and testing data was used to evaluate and verify data from models that have been trained. Data splitting was performed by using the library from Sklearn in Python. Data normalization in this study used the Standard Scaler where this stage standardized variables by removing the mean score and scaling unit variances. It was implemented for each feature in the sample. This preprocessing was required to prevent data whose scores were too higher compared to other scores that can result in undesirable outputs [26].

The implementation of the model used two model scenarios, namely Logistic Regression and Random Forest. Modeling was started after going through the imbalanced data analysis stage. The imbalanced dataset in this study occurs due to the significant gap between classes. Imbalanced dataset had data with rare class and data with abundant class. Imbalanced datasets occurred because of the significant gap between the number of labels in COVID-19 patients'. The imbalanced dataset can affect the performance of the findings [27]. Therefore, oversampling by using SMOTE technique was employed to overcome this problem. Oversampling generated rare classes so that the number of rare classes equals the number of abundant classes. SMOTE technique is an oversampling approach for minority labels [28]. The equation for SMOTE technique is presented in formula 1.

$$Xnew = Xi + (Xl - Xi)\, x\, \delta \quad (1)$$

Xi is the vector from the minority class feature, X1 is the k-Nearest Neighbors for, and δ is a random number between 0 and 1.

2.3. Data Modelling

Logistic regression is a method used for one type of regression model that connects one or more independent variables with a categorical dependent variable with score of 0 and 1, true or false, large or small [19]. The type of these variables distinguished logistic regression from multiple regression or other types of regression. The formula for logistic regression is presented in Formula 2.

$$Ln\left(\frac{p}{1-p}\right) = B_0 + B_1 X \quad (2)$$

B_0 is the Constant, $B_1$is the Coefficient from every variable, The p-value or probability (Y = 1) can be found using Formula 3.

$$p = \frac{e^{(B_0+B_1X)}}{(1+e^{(B_0+B_1X)})} \quad (3)$$

Formula 3 was used to calculate the probability of the observed data that has been defined in the equation. The p-value ranged from 0 to 1. In this study, logistic regression was developed using Grid Search. The Grid Search method is an alternative method used to decide the best parameter of a model, so that the classification can accurately predict the data without label [25][29]. Logistic regression with Grid Search in this study was better because it obtained the best parameter from hyperparameter optimization. Hyperparameter worked by running several trials in one training process. The more the parameters, the more time required for model training. The combination of parameters resulting from the hyperparameter in the optimization process can represent the values that determine the training process of the logistic regression.

Random Forest was used in this study as the verificator to validate the proposed model performance[30]. Random Forest is a combined tree method derived from the Classification and Regression Tree (CART) method and is based on decision tree techniques so that it can be applied to nonlinear data Random Forest classification was implemented by splitting the data randomly and conducting voting in every class. Random Forest combined the votes from every class, and the highest vote was selected[31].

2.4 Data Evaluation

The evaluation stage in this study used a general evaluation matrix. The evaluation matrix was used to measure the model performance that has been built. The evaluation matrix contained the accuracy, precision, recall, and f1-score. Accuracy is the ratio of the total data that has been classified correctly in testing data. Precision is the ratio of true predictions to the overall false predicted results. Recall indicated the percentage of the data that were correctly classified. F1-score showed the comparison between the mean scores of precision and recall. Formula 4, Formula 5, Formula 6, and Formula 7 were derived from the four evaluation scores. Figure 2 shows the model evaluation with confusion matrix. Confusion matrix was used to visualize the true and false scores in the proposed classification model.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}\, x\, 100\% \quad (4)$$

$$Precision = \frac{TP}{TP+FP}\, x\, 100\% \quad (5)$$

$$Recall = \frac{TP}{TP+FN}\, x\, 100\% \quad (6)$$

$$F1 - Score = 2x\frac{Recall\ x\ precision}{recall+precision} \quad (7)$$

DOI: https://doi.org/10.29207/resti.v7i34868.xxx

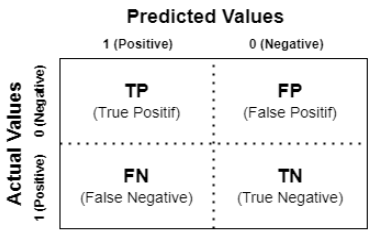


Figure 1. Model evaluation with confusion matrix

True Positives (TP) is the number of cases that were predicted positive to be canceled and actually canceled. True Negative (TN) is the number of cases that were predicted not to be canceled and actually not to be cancelled. False Negative (FN) is the total cases that is predicted not to be cancelled but is actually cancelled. False Positive (FP) is the number of cases that are predicted to be cancelled, when in fact they are not.

2.5 Cross Validation

Cross validation is a method for evaluating robustness of a model and avoid overfitting. In this study, Cross Validation method using k-fold cross validation. In K-fold cross validation, dataset divided into k-subset with the same size. The model is trained on k-1 of these folds and tested on the remaining one. This process is repeated k times, with each fold used as a test set once. The average performance across all k iterations is then used to evaluate the model's overall performance.

3. Results and Discussions

3.1 Results

The initial stage of this study was to collect data from UMM Hospital. The summary of the data is presented in Figure 3. Based on the data, there were 623 patients infected with COVID-19 at UMM Hospital. The distribution presented patients' vital status, the group of deceased and alive cases. The visualization of COVID-19 patient classes is presented in Figure 3.

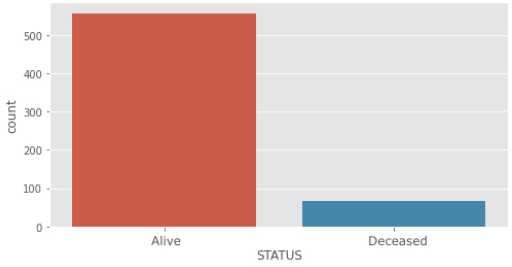


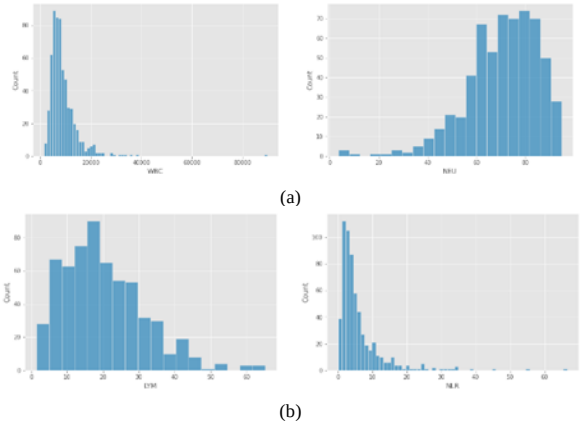
Figure 2. Dataset Summary of COVID-19 Patients Classes

Figure 3 shows that the classes have an imbalance class. The group of alive cases has 557 patients and deceases cases has 66 patients. Imbalancing data in this study were resolved by using oversampling approach, SMOTE method. The screening results are presented under patient status, white blood cells (WBC), neutrophils (NEU), lymphocytes (LYM), and Neutrophil Lymphocyte Ratio (NLR) in patients.

Table 1. Dataset Summary of COVID-19 Patients

	Status	WBC	NEU	LYM	NLR
0	Alive	8460.0	69.6	22.0	3.163636
1	Alive	8780.0	83.7	11.2	7.473214
2	Alive	39.80.0	68.0	25.6	2.656250
3	Alive	13320.0	84.8	10.4	8.153846
4	Alive	7530.0	71.8	20.3	3.536946

Dataset Summary of COVID-19 Patients data is presented in Table 1. The visualization of COVID-19 patient data is presented in Figure 4. The visualization used histogram to identify the data distribution from each variable; WBC, NEU, LYM, and NLR.



(a)

(b)

Figure 4. Data Visualization of WBC, NEU, LYM, and NLR

Figure 4 shows that the four variables have an uneven distribution. The skewness of the four variables is not symmetrical and not normally distributed. The threshold of healthy WBC count is considered to be between 4000 and 11000 WBCs per microliter of blood. NLR was calculated as a simple ratio of absolute neutrophil count and absolute lymphocyte count. Normally, it should be below 3, but a ratio of above 3 signifies acute stress, and a ratio of more than 9 signifies sepsis. Therefore, datasets need to be pre-processed to improve data quality. The preprocessing stage was conducted by processing the missing data. From the four variables, there were 18 null values. Missing data processing was carried out by replacing the missing data with the median score of each variable. Replacing the missing data with median score is presented in Table 2.

Table 2. Filling Median for the Missing Data

```
tes_darah['WBC'] = tes_darah['WBC'].fillna(tes_darah['WBC'].median())
tes_darah['NEU'] = tes_darah['NEU'].fillna(tes_darah['NEU'].median())
tes_darah['LYM'] = tes_darah['LYM'].fillna(tes_darah['LYM'].median())
tes_darah['NLR'] = tes_darah['NLR'].fillna(tes_darah['NLR'].median())
```

The second preprocessing stage was data transformation. In data transformation, data normalization was carried out by using StandardScaler. Data normalization aimed to eliminate data redundancy and overcome data distribution irregularities. The results of data normalization are presented in Figure 5.

```
array([[-0.07054601, -0.42428304],
       [-0.01593491,  0.22389587],
       [-0.83510152, -0.50059611],
       [ 0.75886018,  0.32626581],
       [-0.22925954, -0.3681357 ]])
```

Figure 5. Data Normalization

DOI: https://doi.org/10.29207/resti.v7i34868.xxx

The next data pre-processing was dataset splitting. The data were divided into two parts, namely training data 80% and testing data 20%. From the data splitting, it resulted 498 training data and 125 testing data. The training data were then used in the logistic regression modelling process. The initial stage of implementing logistic regression modeling was balancing the data. The imbalanced data were resolved by using oversampling approach in the minority class. SMOTE was implemented for the imbalanced data. The modelling was optimized by using hyperparameter Grid Search. This method was used to obtain the parameter combination with the best model. The combination of parameters generated in the optimization process can represent the values that determine the logistic regression model training. Figure 6 shows the results of the confusion matrix in the logistic regression model.

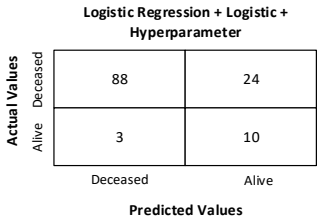


Figure 6. Confusion Matrix of Logistic Regression

Figure 6 shows 88 deceased labels were predicted correctly by the model, while 24 models on the alive labels were detected as deceased. Alive labels were correctly predicted by the model by 10 labels, and 3 deceased labels were detected alive. The results of the evaluation score were calculated with the accuracy, precision, recall, and F1-score in Table 3.

Table 3. Evaluation results on logistic regression

Label	Precision	Recall	F1-Score	Accuracy
Alive (0)	0.97	0.79	0.87	0.78
Deceased (1)	0.29	0.77	0.43	

The results of cross validation using k-fold cross validation method to avoid overtting and handle robustness. The results of the cross validation are shown in Table 4.

Table 4. Cross Validation Results

k	Accuracy
10	70%

Table 4 shows the accuracy of k-10 cross validation of 70%. The accuracy value shows logistic regression can deal with overfitting and avoid robustness around 70% from the dataset. In this study, Random Forest was used as a comparison of modeling performance with logistic regression. The results of the confusion matrix and evaluation from Random Forest are shown in Table 5.

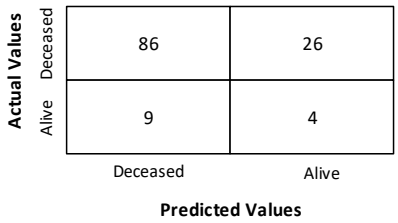


Figure 7. Confusion Matrix of Random Forest

From Figure 7, it can be seen that 86 deceased labels were predicted correctly by the model, while 26 models in alive labels were detected deceased. Subsequently, 4 alive labels were predicted correctly by the model, and 9 deceased labels were detected alive.

Table 5. Evaluation Score of Random Forest

Label	Precision	Recall	F1-Score	Accuracy
Alive (0)	0.91	0.77	0.83	0.72
Deceased (1)	0.13	0.31	0.19	

3.2 Discussions

After carrying out two models, various results were obtained from each model. Table 6 is the result of the comparison models of logistic regression and random forest. In two models, it can be seen that the highest score of precision, recall, and f1-score from alive and deceased label were obtained in logistic regression model. High score in deceased means that the model has described mortality of COVID-19 patients. Logistic regression was able to properly determine the mortality rate of COVID-19 patients based on the recall score of 0.77. The number was higher than the random forest recall score of 0.31. The mortality rate must have been accurately calculated since variations in mortality between populations and countries serve as a crucial proxy indication of relative risk of death that informs policy choices regarding the distribution of limited medical resources during the ongoing COVID-19 pandemic [32].

Table 6. The Comparison between logistic regression and random forest

Model	Label	Precision	Recall	F1-Score
Logistic Regression	Alive	0.29	0.79	0.87
	Deceased	0.91	0.77	0.43
Random Forest	Alive	0.13	0.77	0.83
	Deceased	0.29	0.31	0.19

Table 7 is the result of comparison accuracy between logistic regression and random forest.

Table 7. The Comparison accuracy between logistic regression and random forest

Model	Accuracy
Logistic Regression	0.78
Random Forest	0.72

DOI: https://doi.org/10.29207/resti.v7i34868.xxx

The accuracy in logistic regression was higher than the random forest which is equal to 0.78. Overall the logistic regression has better performance than the random forest method. This is because logistic regression can resolve imbalanced data with undersampling, oversampling, and combine sampling schemes by producing a better mean score than random forest[33]. In random forest, the undersampling scheme can increase the recall mean score. In this study, imbalanced data was caused by oversampling. Therefore, the more appropriate model to use for this data is logistic regression. The results of comparing the logistic regression model to the previous research show that logistic regression with alive and deceased labels has a higher accuracy value of 0.78, while logistic regression in previous study showed an accuracy of 0.703 with age and gender labels [19]. These results indicate that logistic regression is well-suited for classifying COVID-19 patients under various labels. The focus of this study is on the labels of alive and deceased, as the mortality rate is the main indicator used by the WHO to predict COVID-19 in various countries during the pandemic.

4. Conclusion

In this study, the classification was carried out by using several model scenarios, namely logistic regression and random forest. From the findings, it can be seen that the logistic regression method has better performance than random forest. The addition of data balancing techniques, namely SMOTE and hyperparameter grid search, aimed to obtain model performance and accuracy score of 0.78. The results of this study could be implemented in other COVID-19 cases for predicting mortality based on the result of hematology test patient. This method can also be applied in the form of a website-based application with restricted access, limited to the staff and management of the hospital, to predict mortality from COVID-19 cases.

Acknowledgment

We are grateful to Universitas Muhammadiyah Malang which have provided support for this research. Their support in this study has made it possible to carry out the necessary experiments, and data analysis. Furthermore, we would like to express my appreciation to the research participants who volunteered their time and expertise. Their willingness to participate and contribute to the data collection process has been instrumental in obtaining meaningful results and insights. We would also like to thank the editorial team at the journal for their professionalism and assistance during the submission process. We extend gratitude to all those mentioned above for their valuable support and involvement.

References

[1] S. Bhandari, A. Shaktawat, A. Tak, and B. Patel, "Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters," *Ibnosina J. Med. Biomed. Sci.*, vol. 12, no. 02, pp. 123–129, Jun. 2020, doi: 10.4103/ijmbs.ijmbs_58_20.

[2] F. Mohammadi *et al.*, "Artificial neural network and logistic regression modelling to characterize COVID-19 infected patients in local areas of Iran," *Biomed. J.*, vol. 44, no. 3, pp. 304–316, Jun. 2021, doi: 10.1016/J.BJ.2021.02.006.

[3] H. Rothan and S. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," *J. Autoimmun.*, 2020.

[4] R. Mus, M. Abbas, Y. Sunaidi, P. Studi DIII Teknologi Laboratorium Medis, and F. Teknologi Kesehatan, "Studi Literatur: Tinjauan Pemeriksaan Laboratorium pada Pasien COVID-19," *J. Kesehat. Vokasional*, vol. 5, no. 4, pp. 242–252, Jan. 2021, doi: 10.22146/JKESVO.58741.

[5] Q. Fu *et al.*, "Relationship between changes in the course of COVID-19 and ratio of neutrophils-to-lymphocytes and related parameters in patients with severe vs. common disease," *Epidemiol. Infect.*, vol. 149, 2021, doi: 10.1017/S0950268821000674.

[6] L. A. Potempa, I. M. Rajab, P. C. Hart, J. Bordon, and R. Fernandez-Botran, "Insights into the Use of C-Reactive Protein as a Diagnostic Index of Disease Severity in COVID-19 Infections," *Am. J. Trop. Med. Hyg.*, vol. 103, no. 2, p. 561, Aug. 2020, doi: 10.4269/AJTMH.20-0473.

[7] A. Ma, J. Cheng, J. Yang, M. Dong, X. Liao, and Y. Kang, "Neutrophil-to-lymphocyte ratio as a predictive biomarker for moderate-severe ARDS in severe COVID-19 patients," *Crit. Care*, vol. 24, no. 1, pp. 1–4, Jun. 2020, doi: 10.1186/S13054-020-03007-0/TABLES/1.

[8] M. M. Imran, U. Ahmad, U. Usman, M. Ali, A. Shaukat, and N. Gul, "Neutrophil/lymphocyte ratio—A marker of COVID-19 pneumonia severity," *Int. J. Clin. Pract.*, vol. 75, no. 4, Apr. 2021, doi: 10.1111/IJCP.13698.

[9] R. Channappanavar, J. Zhao, and S. Perlman, "T cell-mediated immune response to respiratory coronaviruses," *Immunol. Res.*, vol. 59, no. 1–3, pp. 118–128, May 2014, doi: 10.1007/S12026-014-8534-Z/FIGURES/1.

[10] B. Zhu *et al.*, "Correlation between white blood cell count at admission and mortality in COVID-19 patients: a retrospective study," *BMC Infect. Dis.*, vol. 21, no. 1, pp. 1–5, Dec. 2021, doi: 10.1186/S12879-021-06277-3/FIGURES/2.

[11] F. Zhou *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *Lancet (London, England)*, vol. 395, no. 10229, p. 1054, Mar. 2020, doi: 10.1016/S0140-6736(20)30566-3.

[12] C. Qin *et al.*, "Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China.," *Clin. Infect. Dis.*, vol. 71, no. 15, pp. 762–768, Aug. 2020, doi: 10.1093/CID/CIAA248.

[13] S. Shi, M. Qin, B. Shen, Y. Cai, T. Liu, and F. Yang, "Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China," *JAMA Cardiol.*, vol. 5, no. 7, pp. 802–810, 2020.

[14] L. Kuri-Cervantes *et al*., "Comprehensive mapping of immune perturbations associated with severe COVID-19," *Sci. Immunol.*, vol. 5, no. 49, Jul. 2020, doi: 10.1126/SCIIMMUNOL.ABD7114/SUPPL_FILE/ABD7114_SM.PDF.

[15] J. Liu *et al*., "Neutrophil-to-lymphocyte ratio predicts critical illness patients with 2019 coronavirus disease in the early stage," *J. Transl. Med.*, vol. 18, no. 1, pp. 1–12, May 2020, doi: 10.1186/S12967-020-02374-0/FIGURES/7.

[16] J. Zhou, Y. Sun, W. Huang, and K. Ye, "Altered Blood Cell Traits Underlie a Major Genetic Locus of Severe COVID-19," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 76, no. 8, pp. E147–E154, Aug. 2021, doi: 10.1093/GERONA/GLAB035.

DOI: https://doi.org/10.29207/resti.v7i34868.xxx

[17] R. He *et al.*, "The clinical course and its correlated immune status in COVID-19 pneumonia," *J. Clin. Virol.*, vol. 127, p. 104361, Jun. 2020, doi: 10.1016/J.JCV.2020.104361.

[18] G. Lu and J. Wang, "Dynamic changes in routine blood parameters of a severe COVID-19 case," *Clin. Chim. Acta.*, vol. 508, p. 98, Sep. 2020, doi: 10.1016/J.CCA.2020.04.034.

[19] M. Romadhon and F. Kurniawan, "A comparison of naive Bayes methods, logistic regression and KNN for predicting healing of Covid-19 patients in Indonesia," *3rd East Indones. Conf. Comput. Inf. Technol.*, pp. 41–44, 2021, doi: 10.1109/EIConCIT50028.2021.9431845.

[20] J. A. Behar, C. Liu, K. Kotzen, and F. W. Wibowo, "Prediction Modelling of COVID-19 Outbreak in Indonesia using a Logistic Regression Model," *J. Phys. Conf. Ser.*, vol. 1803, no. 1, p. 012015, Feb. 2021, doi: 10.1088/1742-6596/1803/1/012015.

[21] S. Bhandari *et al.*, "Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters," *Ibnosina J. Med. Biomed. Sci.*, vol. 12, no. 02, pp. 123–129, Jun. 2020, doi: 10.4103/IJMBS.IJMBS_58_20.

[22] C. Citu *et al*., "The Predictive Role of NLR, d-NLR, MLR, and SIRI in COVID-19 Mortality," *Diagnostics*, vol. 12, no. 1, p. 122, Jan. 2022, doi: 10.3390/DIAGNOSTICS12010122.

[23] W. Shang *et al.*, "The value of clinical parameters in predicting the severity of COVID-19," *J. Med. Virol.*, vol. 92, no. 10, p. 2188, Oct. 2020, doi: 10.1002/JMV.26031.

[24] T. Rahman, A. Khandakar, M. Hoque, … N. I.-I., and U. 2021, "Development and Validation of an Early Scoring System for Prediction of Disease Severity in COVID-19 Using Complete Blood Count Parameters," *IEEE Access*, vol. 9, 2021.

[25] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 6, no. 3, pp. 280–284, Dec. 2020, doi: 10.26418/JP.V6I3.40718.

[26] A. Ambarwari and Q. Adrian, "Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman," *J. Rekayasa Sist. dan Teknol. Inf.*, vol. 4, no. 1, pp. 112–117, 2020.

[27] Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, Jul. 2022, doi: 10.30812/MATRIK.V21I3.1726.

[28] N. Suryana and R. Tri Prasetio, "Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 6, no. 1, pp. 31–37, 2020.

[29] M. Lutz, "Learning Python," *Icarus*, vol. 78, no. 1, p. 700, 2007, doi: 10.1016/0019-1035(89)90077-8.

[30] A. Armonica, "Klasifikasi Jenis Persalinan pada Ibu Hamil dengan Metode Random Forest," *PHP Rosa - Pros. Semin. Nas.*, pp. 184–188, 2022.

[31] P. Subarkah, P. Pambudi, S. Oktaviani, and N. Hidayah, "Perbandingan Metode Klasifikasi Data Mining untuk Nasabah Bank Telemarketing," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 1, pp. 139–148, Sep. 2020, doi: 10.30812/MATRIK.V20I1.826.

[32] A. H. Jahromi and H. Mahmoudi, "Estimates of mortality following COVID-19 Infection; comparison between Europe and the United States," *Immunopathol. Persa*, vol. 7, no. 1, pp. e05–e05, Jul. 2020, doi: 10.34172/IPP.2021.05.

[33] T. Purwa, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten," *J. Mat. Stat. dan Komputasi*, vol. 16, no. 1, pp. 58–73, 2019, doi: 10.20956/jmsk.v16i1.6494.

DOI: https://doi.org/10.29207/resti.v7i34868.xxx