# Classification of Death Risk for COVID-19 Patients Using Bayesian Logistic Regression and Naive Bayes Classifier

Ferra Yanuar, Muhammad Qolbi Shobri, Ridha Fauza Mabrur, Isma Hasanati Putri, Aidinil Zetra

*Abstract*— **The COVID-19 virus, which was first confirmed in Wuhan, China, in 2019, has spread worldwide, and as of the beginning of 2022, more than 240 million cases and 4.77 million deaths have been confirmed. This present study aims to classify the significant risk of death for COVID-19 patients. Two statistical classification techniques, i.e., Bayesian logistic regression and Naive Bayes classifier, are implemented in this study. The performance of both methods is measured using several indicators. This study found that the significant explanatory variable involved in classifying the death risk for COVID-19 patients is Comorbidity, based on the output of both methods. Thus, more care is needed for those with comorbidities if they have been exposed to the virus.**

*Index Terms*—**Bayesian Logistics Regression, COVID-19, Naive Bayes Classifier, *Press's Q*.**

## I. Introduction

At the beginning of 2021, the outbreak of the COVID-19 pandemic has increased sharply worldwide. The overall mortality of COVID-19 based on evidence is 3.77-5.4% [1], [2]. Identifying the risk factors related to morbidity and mortality in COVID-19 patients is urgently required to reduce the overall mortality rate. Currently, Indonesia is amidst a second wave of infections, and more than 3 million confirmed cases have been reported, including more than 85 thousand deaths [3].

Studies investigating the risk factors of death for COVID-19 patients have been extensively published [4]. Several studies have considered that older age, male, and comorbidity were closely associated with the risk of mortality for patients with COVID-19 [1], [2], [5]–[7]. Most studies employed multivariate linear regression or logistic regression to determine the risk of mortality for Covid-19 patients, such as the study by [8], [9]. Limited studies have been available to classify the mortality risk factor for COVID-19 patients.

The logistic regression method and the Naïve Bayes Classifier method are popular methods used in classification algorithm [10]–[12]. Both methods are used to determine if sample belongs to a certain class, whether death or recovered in this case. This study combines logistic regression with the Bayesian method, known as the logistic Bayesian regression method. The Naïve Bayes Classifier method works well when it has a high dimensional space [13].

Sumatera Barat, Indonesia, is the province that experienced the highest spike in the number of COVID-19 patients outside Java. It was reported that 70.869 confirmed cases of COVID-19 in West Sumatra, with 1,496 reported dead [3, 13]. Therefore, this study aims to classify the risk factors of death for COVID-19 patients in West Sumatra using the logistic Bayesian and Naïve Bayes Classifier methods. Results based on both ways are then compared. No previous research has been studied on this topic yet.

## II. Materials and Methods

### A. *Data*

This study used COVID-19 patient data treated at Dr. M. Djamil hospital and Universitas Andalas hospital in Padang, West Sumatra, Indonesia, from March to December 2020. The patients treated at both hospitals were referral patients from other cities/subdistricts in West Sumatra. The object of observation is a COVID-19 patient who died or recovered after hospitalization, identified as a response variable in the study. Age, Gender, and Comorbidity are assumed as predictors of the response.

This study employs multinomial Bayesian logistic regression and the Naïve Bayes Classifier method since age and Comorbidity do not have a multivariate Gaussian distribution [15]. Thus, all three predictors are in categorical type. All 457 COVID-19 patients with complete information are involved in the analysis. By using splitting methods, this study used a data set randomly divided into two folds, i.e., 80% of all data (365 data) is training data, and 20% (92 data)

Manuscript received December 17, 2022; revised May 14, 2023. This research was funded by Andalas University Research (PGB) Grant No.T/3939/UN.16.17/PP.OK-KRPGB/LPPM/2019.

Ferra Yanuar (Corresponding Author) is a lecturer at the Department of Mathematics & Data Sciences, Faculty of Mathematics and Natural Sciences, Andalas University, Kampus Limau Manis, Padang, Indonesia, 25163. (e-mail: ferrayanuar@sci.unand.ac.id).

Muhammad Qolbi Shobri is postgraduate student at the Mathematics Magister Programme, Department of Mathematics & Data Sciences, Faculty of Mathematics and Natural Sciences, Andalas University, Kampus Limau Manis, Padang, Indonesia, 25163. (e-mail: qolbisyana@gmail.com).

Ridha Fauza Majbur is postgraduate student at the Mathematics Magister Programme, Department of Mathematics & Data Sciences, Faculty of Mathematics and Natural Sciences, Andalas University, Kampus Limau Manis, Padang, Indonesia, 25163. (e-mail: 2020432001_ridha@student.unand.ac.id).

Isma Hasanati Putri is undergraduate student at the Department of Mathematics & Data Sciences, Faculty of Mathematics and Natural Sciences, Andalas University, Kampus Limau Manis, Padang, Indonesia, 25163. (e-mail: ismahasanatiputri@gmail.com).

Aidinil Zetra is a lecturer at the Department of Political Science, Faculty of Social Sciences and Political Sciences, Andalas University, Kampus Limau Manis, Padang, Indonesia, 25163. (e-mail: aidinil@soc.unand.ac.id).

rest is testing data. Table I summarizes the features of predictors and responses in this study for training data, while testing data is provided in Table II.

*B. Methods*

Bayesian Logistic Regression

It is assumed the response variable $\boldsymbol{Y}$ has two values that are $\boldsymbol{Y} = 1$ for "Yes" and $\boldsymbol{Y} = \boldsymbol{0}$ for "No". Let in group *i*, the random variable $Y_i$ takes one with probability $\pi_i$ and zero with probability $1 - \pi_i$. The random variable $Y_i$ has Bernoulli distribution with parameter $\pi_i$ defined as [16],

$$P(Y_i = y_i) = \pi_i{}^{y_i}(1 - \pi_i)^{1-y_i} \quad (1)$$

Let $n_i$ expresses the number of observations in group $i$, $y_i$ is a number of "Yes" cases in $n_i$. It is assumed $n_i$ are independent in each group with probability $\pi_i$, then $Y_i$ has a binomial distribution with parameter $\pi_i$ and $n_i$. The probability distribution function of $Y_i$ is given by

$$P(Y_i = y_i) = \binom{n_i}{y_i}\pi_i{}^{y_i}(1 - \pi_i)^{1-y_i}$$

Here, $\boldsymbol{X_i}$ is a vector of predictors and $\boldsymbol{\beta}$ is a vector of the regression coefficient. The odds of *i*th observation are given by:

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\boldsymbol{X_i^T\beta}\right) \quad \text{or} \quad \pi_i = \frac{\exp\left(\boldsymbol{X_i^T\beta}\right)}{1+\exp\left(\boldsymbol{X_i^T\beta}\right)}$$

The likelihood function for the Bernoulli distribution in Eq. (1) is presented as,

$$L(\pi_i|y_i) = \prod_{i=1}^{n}\pi_i{}^{y_i}(1 - \pi_i)^{1-y_i} \quad (2)$$

Eq. (2) could be provided in

$$L(\boldsymbol{\beta}, \sigma|\boldsymbol{X_i}, Y_i) = \prod_{i=1}^{n}\left(\frac{\exp\left(\boldsymbol{X_i^T\beta}\right)}{1+\exp\left(\boldsymbol{X_i^T\beta}\right)}\right)^{y_i}\left(1 - \frac{\exp\left(\boldsymbol{X_i^T\beta}\right)}{1+\exp\left(\boldsymbol{X_i^T\beta}\right)}\right)^{1-y_i} \quad (3)$$

The estimation process of parameters in the Bayesian approach considers the updation of knowledge regarding unknown parameters $\boldsymbol{\beta}$ and $\sigma$ which is known as the prior distribution. Next, we assign normal distribution as a prior distribution to the unknown regression parameters $\boldsymbol{\beta}$ and $\sigma$ given as,

$$f(\boldsymbol{\beta}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}exp\left\{-\frac{1}{2}\left(\frac{\boldsymbol{\beta}-\mu}{\sigma}\right)^2\right\}, \quad (4)$$

for $-\infty < \boldsymbol{\beta}, \mu < \infty, \sigma > 0$.

Utilizing the well-known Bayes theorem, the posterior distribution of parameters $\boldsymbol{\beta}$ and $\sigma$ is proportional to the product of Eq. (3) and (4) as follows,

$$f(\boldsymbol{\beta}, \sigma|\boldsymbol{X_i}, Y_i) \propto L(\boldsymbol{\beta}, \sigma|\boldsymbol{X_i}, Y_i)\, f(\boldsymbol{\beta}, \sigma) \quad (5)$$

For Bayesian estimates, it has been shown that under mild conditions and for a sufficiently large number of iterations (*m*), the joint distribution of $\left(\boldsymbol{\beta}^{(m)}, \sigma^{(m)}\right)$ converges at an exponential rate to the desired posterior distribution $(\boldsymbol{\beta}, \sigma)$ [17]–[19]. Hence, $(\boldsymbol{\beta}, \sigma)$ can be approximated by the empirical distribution of a sufficiently large number of simulated observations collected after convergence of the algorithm. The convergence of the algorithm is identified by trace plot and density plot of the generated parameters values from different starting points [20].

Naïve Bayes Classifier

In the Naïve Bayes Classifier method, a COVID-19 patient is predicted to have a risk of death with k-diagnostic measurements as explanatory variables if the maximum posterior probability in the "$Y = 1$ or Yes" and "$Y = 0$ or No" group is defined as following [21]

$$P(Y = y|X_1, X_2, ..., , X_k) = \frac{P(Y=y)P(X_1,X_2,...,,X_k|Y=y)}{P(X_1,X_2,...,,X_k)}, for\ y = 0,1. \quad (6)$$

Since the denominator in Eq. (6) is a constant, this posterior equation becomes proportional to a multiplication of the likelihood function and prior distribution, as presented in the following equation.

$$P(Y = y|X_1, X_2, ..., , X_k) \propto P(Y = y)P(X_1, X_2, ..., , X_k|Y = y). \quad (7)$$

While the likelihood function in Eq. (7) is constructed by using the product rule of *k* explanatory variables given any variables. The rule is presented in Eq. (8)

$$\begin{aligned} &P(X_1, X_2, ..., , X_k|Y = y) \\ &= P(X_1|Y = y)P(X_2, ..., X_k|Y = y, X_1) \\ &= P(X_1|Y = y)P(X_2|Y = y, X_1)P(X_3, ..., X_k|Y = y, X_1, X_2) \\ &= P(X_1|Y = y)P(X_2|Y = y, X_1)P(X_3|Y = y, X_1, X_2) \\ &\quad P(X_{k-1}|Y = y, X_1, ..., X_{k-2})P(X_k|Y = y, X_1, ..., X_{k-1}) \end{aligned} \quad (8)$$

The definition of the likelihood function as presented in Eq. (8) can be made simpler and more efficacious since there is the strong independence assumption in this Naïve Bayes Classifier method [22]. Eq. (8) changes into,

$$P(X_1, X_2, ..., , X_k|Y = y) = \prod_{j=1}^{k}P\left(X_j|Y = y\right) \quad (9)$$

This work assumes that the probability $P\left(X_j|Y = y\right)$ has a Multinomial Gaussian distribution. In certain cases, it will be found that certain explanatory variables are not found in a category. So that the conditional probability is zero, this will cause the likelihood value to be zero. To overcome this, the Laplace Smoothing method can be used by adding a fake sample. Therefore, each of probability $P\left(X_j|Y = y\right)$ is formulated as,

$$P\left(X_j = x_j|Y = y\right) = \frac{N\left(x_j, y\right) + 1}{N(y) + v}, \quad (10)$$

where $N\left(x_j, y\right)$ is total observations in explanatory variable $X_j = x_j$ in category $y$, $N(y)$ is total observations in category $y$ and $v$ is total categories in explanatory variable $X_j$. Then, the posterior distribution in Eq. (7) can be formulated into

$$\begin{aligned} P(Y = y|X_1, X_2, ..., , X_k) &\propto P(Y = y)\prod_{j=1}^{k}P\left(X_j|Y = y\right) \quad (11) \\ &\propto P(Y = y)\prod_{j=1}^{k}\frac{N\left(x_j, y\right) + 1}{N(y) + v}. \end{aligned}$$

The rule to make decision in Naïve Bayes Classifier method is using Maximum A Posteriori (MAP) [15]. MAP is used to determine estimate group for each data, as defined following,

$$\widehat{Y} = \arg max_{y\epsilon 0,1}\, P(Y = y|X_1, X_2, ..., , X_k). \quad (12)$$

If $P(Y = 1|X_1, X_2, ..., , X_k) > P(Y = 0|X_1, X_2, ..., , X_k)$ so

classification for data is $y = 1$ or "Yes". Vice versa, if $P(Y = 1|X_1, X_2, ..., X_k) > P(Y = 0|X_1, X_2, ..., X_k)$ then classification for data is $y = 0$ or "No".

In Naïve Bayes Classifier, information gain is used as attribute selection measures for selecting the splitting criterion that best separates a given data partition [23]. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A), which is formulated as,

$$Gain\ (A) = Info\ (H) - Info_A\ (H), \quad (13)$$

where

$$Info\ (H) = -\sum_{i=1}^{m} p_i\, log_2(p_i),$$

and

$$Info_A\ (H) = \sum_{j=1}^{v} \frac{|H_j|}{|H|} \times Info\ (H_j),$$

where $p_i$ is the nonzero probability that an arbitrary tuple in A belongs to category $C_i$ and is estimated by $|C_{i,H}|/|H|$. Gain (A) informs us how much would be gained by branching on A.

The next analysis is to test the performance of both methods in classifying the determinants of the mortality risk for COVID-19 patients. This study uses six indicators to measure the performance of each technique based on the confusion matrix [24], as provided in Table III. The indicators are accuracy, sensitivity, precision, specificity, negative predicted value (NPV), and *Press's Q*.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (14)$$

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (15)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (16)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (17)$$

$$NPV = \frac{TN}{(TN+FN)} \quad (18)$$

$$Press's\ Q = \frac{(N-(nK))^2}{N(K-1)}, \quad (19)$$

where $N = TP + FN + FP + TN$, $n = (TP + TN)$, $K$ represents number categories of response. *Press's Q* measures the accuracy and stability in the classification method. Classification is said to be accurate and stable if the value of *Press's Q* is greater than $\chi^2_{\alpha,1}$ [25].

## III. Results and Discussion

This paper proposed two methods to model the mortality risk for patients with COVID-19 in West Sumatra. In the first step, we estimate the parameter model using Bayesian logistic regression by including all three predictors in the hypothesis model using data training. By using WinBugs version 14, the results of the estimation process are provided in Table IV.

Table IV shows that several categories are not significant to give effect to the outcome. We then modify the model by excluding Gender $(X_1)$ and Age $(X_2)$. The result of reduced model is presented in Table V. Table V informs us that Constanta and Comorbidity as the predictor variables are statistically significant of the reduced model. The predictor variables are also convergence since the MC error value is less than 5% of standard deviation for corresponding predictor. Thus, this final model based on Bayesian logistic regression method could be accepted.

We then classify the risk of death for patients with COVID-19 using the Naïve Bayes Classifier method. The Naïve Bayes Classifier is estimated using maximum likelihood estimation. Similar to Bayesian logistic regression, Naïve Bayes also does two steps of the estimation process. In the first step, all predictor variables are included in the analysis. The estimated values for information gain (IG) are used for determining the significant predictors, as presented in Table VI.

Table VI informs us that Gender $(X_1)$ and Age $(X_2)$ has a lower IG value than the average IG. Gender and Age are not significant in classifying the mortality risk of patients with COVID-19 in this study. Therefore, Gender and Age are excluded from the analysis, and the final model only includes comorbidities. This result is similar to the final model as the output of Bayesian Logistic Regression. Thus, both models could be compared. The performance indicators of accuracy, sensitivity, precision, specificity, negative predicted value (NPV), and *Press's Q* for this final model are provided in Table VII.

TABLE VI.
INFORMATION GAIN VALUE FOR TRAINING DATA BASED ON NAÏVE BAYES

| Predictor | IG Value |
|---|---|
| Gender $(X_1)$ | 0.0467 |
| Age $(X_2)$ | 0.0551 |
| Comorbidity $(X_3)$ | 0.1064 |
| Average IG | 0.0694 |

We can see in Table VII that death risk classification with Naïve Bayes Classification tends to produce higher values than Bayesian logistic regression for all indicators. The value of Press's Q for both methods is more significant than $\chi^2_{0.05,1}(= 3.841)$. This result indicates that classification using both methods has been accurate and stable. It also could be concluded that NBC is more precise and sensitive and has higher values for NPV and *Press's Q*. It implies that the NBC method performs better than BLR.

The performance of both models was then also checked by using testing data; the values of each indicator are provided in Table VIII. We found quite similar values as training data as well. The values for *Press's Q* are greater than $\chi^2_{0.05,1}(= 3.841)$. Therefore, the classification-based BLR and NBC are said to be stable and accurate in this data set. NBC tends to have higher values for all criteria than the BLR method. Thus, it could be concluded that NBC's results are a better model than BLR's.

## IV. Conclusions

In this paper, we have investigated the performance of two statistical classification techniques i.e., Bayesian logistic regression and Naive Bayes classifier in order to classify the

model of Death Risk for COVID-19 Patients. The performance of both methods is measured using tests of accuracy, sensitivity, specificity, precision, negative predicted value (NPV), and *Press's Q*. This study found that the death risk for patients of COVID-19 classification with Naive Bayes outperforms Bayesian logistic regression based on all performance indicators used in this study. The significant explanatory variable involved in the classification of the death risk for COVID-19 patients is Comorbidity only. Therefore, our recommendation is to take more attention to individuals who have comorbidity (such as cardiovascular disease, diabetes, chronic respiratory disease, and cancer) so they are not exposed to the virus. More care is needed for those who have Comorbidity if they have been exposed to the virus. This study also suggests using Naive Bayes for classifying the death risk for COVID-19 patients as well as medical informatics data.

TABLE I.
DESCRIPTIVE STATISTICS OF TRAINING DATA

| Variables | Category | Recovered | | Died | |
|---|---|---|---|---|---|
| | | Frequency | Percentage | Frequency | Percentage |
| Gender ($\boldsymbol{X_1}$) | Female | 177 | 59.2 | 27 | 40.9 |
| | Male | 122 | 40.8 | 39 | 59.1 |
| Age ($\boldsymbol{X_2}$) | $\leq$ 20 years old | 17 | 5.7 | 3 | 4.5 |
| | 21 – 40 years old | 116 | 38.8 | 3 | 4.5 |
| | 41 – 60 years old | 101 | 33.8 | 36 | 54.4 |
| | $>$ 60 years old | 65 | 21.7 | 24 | 36.6 |
| Comorbidity ($\boldsymbol{X_3}$) | No Comorbidity | 131 | 43.8 | 12 | 18.8 |
| | 1 – 3 Comorbidity | 168 | 56.2 | 38 | 57.7 |
| | $>$ 3 Comorbidity | 0 | 0 | 16 | 24.4 |

TABLE II.
DESCRIPTIVE STATISTICS OF TESTING DATA

| Variables | Category | Recovered | | Died | |
|---|---|---|---|---|---|
| | | Frequency | Percentage | Frequency | Percentage |
| Gender ($\boldsymbol{X_1}$) | Female | 42 | 57.5 | 7 | 36.8 |
| | Male | 31 | 42.5 | 12 | 63.2 |
| Age ($\boldsymbol{X_2}$) | $\leq$ 20 years old | 5 | 6.9 | 2 | 10.5 |
| | 21 – 40 years old | 31 | 42.5 | 2 | 10.5 |
| | 41 – 60 years old | 25 | 34.2 | 10 | 52.6 |
| | $>$ 60 years old | 12 | 16.4 | 5 | 26.,4 |
| Comorbidity ($\boldsymbol{X_3}$) | No Comorbidity | 33 | 45.2 | 9 | 47.4 |
| | 1 – 3 Comorbidity | 33 | 45.2 | 10 | 52.6 |
| | $>$ 3 Comorbidity | 7 | 9.6 | 0 | 0 |

TABLE III.
CONFUSION MATRIX

| Confusion Matrix | | Actual (True Group) | |
|---|---|---|---|
| | | Died | Recovered |
| Diagnosis (Hypothesized group) | Died (Yes) | True Positive (TP) | False Positive (FP) |
| | Recovered (No) | False Negative (FN) | True Negative (TN) |

TABLE IV.
ESTIMATED PARAMETER MODEL USING BAYESIAN LOGISTIC REGRESSION

| Predictor | Parameter | Mean | Standard Deviation | MC Error | Decision |
|---|---|---|---|---|---|
| Constanta | $\beta_0$ | 1.6171* | 0.5453 | 0.01389 | Converge |
| Gender ($X_1$) | | | | | |
| Female ($\boldsymbol{X_{1D1}}$) | $\beta_{1D1}$ | -0.2096 | 0.3277 | 0.00323 | Converge |
| Age ( $X_2$ ) | | | | | |
| $\leq$ **20** years old ($\boldsymbol{X_{2D1}}$) | $\beta_{2D1}$ | 0.1571 | 0.7352 | 0.00729 | Converge |
| $\boldsymbol{21-40}$ years old ($\boldsymbol{X_{2D2}}$) | $\beta_{2D2}$ | -1.957* | 0.6589 | 0.00654 | Converge |
| $\boldsymbol{41-60}$ years old ($\boldsymbol{X_{2D3}}$) | $\beta_{2D3}$ | 0.0095 | 0.359 | 0.00499 | Converge |
| Comorbidity ($X_3$) | | | | | |
| No Comorbidity ($\boldsymbol{X_{3D1}}$) | $\beta_{3D1}$ | -3.8912* | 0.6193 | 0.01173 | Converge |
| 1 - 3 Comordities ($\boldsymbol{X_{3D2}}$) | $\beta_{3D2}$ | -2.7802* | 0.5049 | 0.01219 | Converge |

*Significant at level 0.05

TABLE V.
REDUCED MODEL USING BAYESIAN LOGISTIC REGRESSION

| Predictor | Parameter | Mean | Standard Deviation | MC Error | Decision |
|---|---|---|---|---|---|
| Constanta | $\beta_0$ | 1.228* | 0.4545 | 0.01123 | Converge |
| Comorbidity ($\boldsymbol{X_3}$) | | | | | |
| No Comorbidity ($\boldsymbol{X_{3D1}}$) | $\beta_{3D1}$ | -4.049* | 0.5898 | 0.01215 | Converge |
| $\boldsymbol{1-3}$ Comordities ($\boldsymbol{X_{3D2}}$) | $\beta_{3D2}$ | -2.684* | 0.4917 | 0.01201 | Converge |

*Significant at level 0.05

TABLE VII.
INDICATOR PERFORMANCE OF BLR AND NBC

| Classification Method | Accuracy | Sensitivity | Precision | Specificity | NPV | *Press's Q* |
|---|---|---|---|---|---|---|
| Training Data | | | | | | |
| BLR | 86.30% | 24.44% | 100% | 100% | 85.67% | 192.39 |
| NBC | 86.85% | 27.78% | 100% | 100% | 86.17% | 198.24 |

TABLE VIII.
INDICATOR PERFORMANCE OF BLR AND NBC

| Classification Method | Accuracy | Sensitivity | Precision | Specificity | NPV | *Press's Q* |
|---|---|---|---|---|---|---|
| Testing Data | | | | | | |
| BLR | 82.61% | 52.63% | 58.82% | 88.00% | 90.14% | 39.13 |
| NBC | 84.78% | 52.63% | 66.67% | 88.31% | 93.15% | 44.52 |

## REFERENCES


[1] J. Zhang *et al.*, "Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China," *Clinical Microbiology and Infection*, vol. 26, no. 6, pp. 767–772, Jun. 2020, doi: 10.1016/j.cmi.2020.04.012.

[2] X.-B. Zhang *et al.*, "Risk factors for mortality of coronavirus disease-2019 (COVID-19) patients in two centers of Hubei province, China: A retrospective analysis," *PLoS ONE*, vol. 16, no. 1, p. e0246030, Jan. 2021, doi: 10.1371/journal.pone.0246030.

[3] R. Kominfo, "Situasi COVID-19 di Indonesia," 2021. Accessed: Jul. 28, 2021. [Online]. Available: https://covid19.go.id/berita/data-vaksinasi-covid-19-update-28-juli-2021

[4] G. Nijman *et al.*, "Risk factors for in-hospital mortality in laboratory-confirmed COVID-19 patients in the Netherlands: A competing risk survival analysis," *PLoS ONE*, vol. 16, no. 3, p. e0249231, 2021, doi: 10.1371/journal.pone.0249231.

[5] F. Zhou *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *The Lancet*, vol. 395, no. 10229, pp. 1054–1062, Mar. 2020, doi: 10.1016/S0140-6736(20)30566-3.

[6] F. Yanuar and A. Zetra, "Length-of-Stay of Hospitalized COVID-19 Patients Using Bootstrap Quantile Regression,"*IAENG International Journal of Applied Mathematics*, vol. 51, no. 3, pp. 799–810, 2021.

[7] F. Yanuar, A. S. Deva, A. Zetra, and M. Maiyastri, "Length of hospital stay model of COVID-19 patients with quantile Bayesian with penalty LASSO," *Commun. Math. Biol. Neurosci.*, vol. 23, no. 1–18, 2023, doi: 10.28919/cmbn/7881.

[8] M. Biswas, S. Rahaman, T. K. Biswas, Z. Haque, and B. Ibrahim, "Association of Sex, Age, and Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-Analysis," *Intervirology*, vol. 64, no. 1, pp. 36–47, 2021, doi: 10.1159/000512592.

[9] N. Rosenthal, Z. Cao, J. Gundrum, J. Sianis, and S. Safo, "Risk Factors Associated With In-Hospital Mortality in a US National Sample of Patients With COVID-19," *JAMA Netw Open*, vol. 3, no. 12, p. e2029058, Dec. 2020, doi: 10.1001/jamanetworkopen.2020.29058.

[10] S. Agarwal, N. Jain, and S. Dholay, "Adaptive Testing and Performance Analysis Using Naive Bayes Classifier," *Procedia Computer Science*, vol. 45, pp. 70–75, 2015, doi: 10.1016/j.procs.2015.03.088.

[11] N. Dureh, C. Choonpradub, and P. Tongkumchum, "An alternative method for logistic regression on contingency tables with zero cell counts," *Songklanakarin Journal of Science and Technology*, vol. 38, no. 2, pp. 171–176, 2016.

[12] L. M. Gladence, M. Karthi, and V. M. Anu, "A Statistical Comparison Of Logistic Regression And Different Bayes Classification Methods For Machine Learning," *Journal of Engineering and Applied Sciences*, vol. 10, no. 14, pp. 5947–5953, 2015.

[13] A. B. Adetunji, J. P. Oguntoye, O. D. Fenwa, and N. O. Akande, "Web Document Classification Using Naïve Bayes," *JAMCS*, vol. 29, no. 6, pp. 1–11, Dec. 2018, doi: 10.9734/JAMCS/2018/34128.

[14] I. D. Sudirman and D. Y. Nugraha, "Naive Bayes Classifier For Predicting The Factors That Influence Death Due To COVID-19 in China," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 10, pp. 1686–1696, 2005.

[15] J. J. Park, S.-C. Chen, and K.-K. Raymond Choo, Eds., "Bayesian Multinomial Naïve Bayes Classifier to Text Classification," in *Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech 2017*, in Lecture Notes in Electrical Engineering, vol. 448. Singapore: Springer Singapore, 2017. doi: 10.1007/978-981-10-5041-1.

[16] A. Shrivastava, K. Kumar, and N. Kumar, "Business Distress Prediction Using Bayesian Logistic Model for Indian Firms," *Risks*, vol. 6, no. 113, pp. 1–15, 2018, doi: 10.3390/risks6040113.

[17] S.-Y. Lee, *Structural Equation Modeling: A Bayesian Approach*. John Wiley & Sons, Ltd, 2007.

[18] F. Yanuar and A. Zetra, "The Performance of Bayesian Analysis in Structural Equation Modelling to Construct The Health Behaviour During Pandemic COVID-19," *Pak.j.stat.oper.res.*, vol. 18, no. 3, pp. 575–587, Sep. 2022, doi: 10.18187/pjsor.v18i3.4096.

[19] F. Yanuar, A. Zetra, C. Muharisa, D. Devianto, A. R. Putri, and Y. Asdi, "Bayesian Quantile Regression Method to Construct the Low Birth Weight Model," *J. Phys.: Conf. Ser.*, vol. 1245, p. 012044, Aug. 2019, doi: 10.1088/1742-6596/1245/1/012044.

[20] F. Yanuar, "The Use of Uniformative and Informative Prior Distribution in Bayesian SEM," *Global Journal of Pure and Applied Mathematics*, vol. 11, no. 5, pp. 3259–64, 2015.

[21] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web

Pages," *International Journal of Computer Science*, vol. 4, no. 1, pp. 16–23, 2009.
[22] E. Alpaydin, *Introduction to Machine Learning*. London, English: The Massachusetts Institute of Technology Press, 2004.
[23] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Technigues*, Third. Morgan Kaufman Publish: Elsevier, 2006.
[24] H. Yun, "Prediction model of algal blooms using logistic regression and confusion matrix," *IJECE*, vol. 11, no. 3, pp. 2407–2413, 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.
[25] J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis*, Seven. Prentice Hall, 2006.