Scientific Journal of Informatics
Vol. 11, No. 1, Feb 2024
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Comparison of Discriminant Analysis and Support Vector Machine on Mixed Categorical and Continuous Independent Variables for COVID-19 Patients Data

Husnul Aris Haikal[1*], Aji Hamim Wigena[2], Kusman Sadik[3], Efriwati[4]
[1,2,3] Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Pertanian Bogor, Indonesia
[4]National Research and Innovation Agency, Indonesia

Abstract.
Purpose: Numerous factors can affect the duration of COVID-19 recovery. One method involves utilizing natural herbal medication. This study seeks to determine the variables influencing the duration of COVID-19 recovery and to compare discriminant analysis and support vector machine models using COVID-19 patient data from West Sumatra.
Methods: Two data mining methods, Discriminant Analysis and Support Vector Machine with different types of kernels (linear, polynomial, and radial basis function), were employed to categorize the time of COVID-19 recovery in this work. The study utilized 428 data points, with 75% allocated for training data and 25% for testing data. The independent factors were evaluated by determining the selection variables' information value (IV) to gauge their influence on the dependent variable. Data resampling techniques were employed to tackle the problem of data imbalance. This study employs data resampling techniques, including undersampling, oversampling, and SMOTE. The balancing accuracy of Discriminant Analysis and Support Vector Machine was examined.
Result: The Discriminant Analysis with SMOTE achieved a balanced accuracy of 66.50%, outperforming the linear kernel Support Vector Machine with SMOTE, which had a balanced accuracy of 63.20% in this dataset.
Novelty: This study assessed the novelty, originality, and value by comparing Discriminant Analysis and SVM algorithms with categorical and continuous independent variables. This research explores techniques for managing imbalanced data using undersampling, oversampling, and SMOTE, with variable selection based on information value assessment.

Keywords: Discriminant analysis, Support vector machine, Mixed independent variable, Resampling, COVID-19
Received November 2023 / **Revised** February 2024 / **Accepted** February 2024

INTRODUCTION

SARS-CoV-2, also known as COVID-19, can cause a variety of effects ranging from no symptoms to multi-organ failure and death. [1] During the initial stages of the pandemic in early 2020, approximately 80% of people who contracted SARS-CoV-2 showed no symptoms, while around 13% experienced severe illness necessitating respiratory assistance, and about 7% needed intensive care due to clinical manifestations such as acute respiratory infection (ARI), sepsis, and multi-organ failure [2].

Natural herbal treatments have been used globally for treating COVID-19. [3] Several COVID-19 individuals in different countries, such as China, have been treated using traditional herbal medicine prescriptions. [4] As a tropical country, Indonesia has abundant medicinal plants, with the West Sumatra region particularly abundant in natural medicinal flora. The inhabitants of West Sumatra have traditionally used indigenous botanicals to treat various illnesses, including COVID-19. The leaves of the sungkai tree (Peronema canescens) in West Sumatra are thought to provide medicinal potential for treating COVID-19. [5] The leaves of the sungkai tree are traditionally used to cure fever, colds, diarrhoea, hypertension, and malaria. They are also being explored as an alternative therapy for COVID-19. Yani [6] conducted research indicating that extracts from young sungkai leaves can enhance immunity by raising the white blood cell count in the blood, therefore strengthening the immune system against many infectious diseases.
COVID-19 has an incubation period, which is the duration between viral infection and the appearance of illness symptoms [7]. COVID-19's incubation time is reportedly 14 days [8], [9]. This study will focus on

*Corresponding author.
Email addresses: haikalhusnularis@apps.ipb.ac.id (Haikal)
DOI: 10.15294/sji.v11i1.48565

analysing the recovery time from COVID-19 as the response variable. Recovery time from COVID-19 is categorised into two groups: patients who recovered during the incubation period (≤ 14 days) and patients who recovered after the incubation period (> 14 days). This study utilises mixed independent variables, encompassing both categorical and continuous variables. The information value was computed for each independent variable, a technique for variable selection that is especially beneficial when the answer variable is binary [10].

The categorisation approach was selected to categorise the COVID-19 recovery time. Classification algorithms predict data groups based on existing class categories utilising independent factors. [11] Imbalanced class data can create classification issues, resulting in misclassification. [12] Imbalanced class data, with unequal distribution of data points among distinct classes, can impact the model's performance [13]. The study's response variable, the duration of COVID-19 recovery, exhibits uneven class characteristics and needs to be addressed. Resampling techniques can assist in addressing imbalanced data [14]. This study employed undersampling, oversampling, and the Synthetic Minority Oversampling Technique (SMOTE).

Undersampling is a resampling technique that randomly decreases the data in the majority class to match or come close to the number in the minority class [15]. Qian [16] discovered that undersampling enhanced classification accuracy in Support Vector Machine (SVM) and discriminant Analysis. Oversampling involves randomly adding data to the minority class to balance or approximate its number with the dominant class, addressing the issue of class imbalance. [17], [18], [19] SMOTE is a skilful resampling technique that has emerged as a suitable alternative for addressing issues associated with imbalanced data. [20] It is an oversampling technique that equalises the class distribution of a dataset by introducing artificial samples to the minority class. [21] Wang [22] observed that SMOTE is an excellent technique for addressing unbalanced data and enhancing accuracy metrics.

Discriminant Analysis is a statistical technique for categorising and assigning new objects to predetermined groups. [23] Ronald A. Fisher established it in 1936, and it is regarded as a classic data mining method. Discriminant Analysis initially had limitations as it exclusively operated with continuous independent variables. [24] Mbina [25] expanded Discriminant Analysis to accommodate mixed categorical-continuous independent variables, providing an alternative for discriminant models with categorical variables. Categorical independent variables are managed by constructing cells from a multinomial table of categorical values in each group rather than converting them into dummy variables [26]. This research employs the Support Vector Machine (SVM) approach alongside Discriminant Analysis.

In his study, Guhathakurata [27] evaluated the performance of a Support Vector Machine (SVM) against various classification algorithms like K-Nearest Neighbour (kNN), Classification Tree (CART), Random Forest, Naïve Bayes, and AdaBoost in categorising COVID-19 patient symptoms. The findings indicated that SVM outperformed the other methods regarding predictive accuracy—James [28], emphasised SVM's excellent performance in object classification. The SVM approach aims to identify the best hyperplane that maximally separates the classes. A *hyperplane* is a mathematical function that can distinguish between different classes.

Scholars have studied mixed independent variables in Discriminant Analysis and Support Vector Machines (SVM). Mahat [29] studied the process of selecting continuous variables in discriminant Analysis with mixed independent variables. Mbina [25] investigated variable selection in discriminant Analysis, including mixed categorical-continuous independent variables. Their research needs to address imbalanced class data management and the utilisation of Information Value for variable selection. Guhathakurata [27] said that SVM is the most effective method for categorising COVID-19 symptoms, but it has yet to address concerns about imbalanced classes and variable selection. Anggrawan [30], in his research, explains the use of SMOTE to overcome the problem of imbalanced data in SVM but needs to clarify the variable selection approach, notably the usage of Information Value. This study intends to investigate the classification outcomes of two techniques utilising data on the duration of COVID-19 recovery in West Sumatra, which includes a combination of independent factors and unbalanced class data.

METHODS

Data Collection

The study utilised secondary empirical data from the West Sumatra Regional Research and Development Agency. The participants in this study are persons living in West Sumatra who tested positive for COVID-19 (COVID-19 survivors) in 2021. This study uses the response variable of patient recovery duration after being cured of COVID-19, categorised as recovery during the incubation phase (≤ 14 days) and recovery beyond the incubation period (> 14 days).

The variables used in the study include mixed categorical and continuous independent variables, as shown in Table 1.

Table 1. Description of dataset

Variables	Features	Type	Description
Y	the duration of recovery for patients from COVID-19	Categorical	1 : ≤ 14 days, 0 : >14 days
X_1	Duration of COVID-19 Symptoms Disappearing	Continuous	Years
X_2	Age	Continuous	Days
X_3	Duration of Consumption	Continuous	Days
X_4	Amount of Sungkai Leaves Consumed in the Potion	Continuous	Leaves
X_5	Symptoms Experienced during COVID-19 Infection	Categorical	Mild, moderate, severe
X_6	Number of Glasses Sungkai Leaf Potion	Continuous	Glass per day
X_7	Daily Intensity of Drinking the Sungkai Leaf Potion	Continuous	Intensity per day
X_8	Gender	Categorical	Male, female

Information Value

Information value (IV) is a commonly used techniques for selecting independent variables in classification algorithms with binary answer variables. [31] The Information Value (IV) is computed by analysing data for each independent variable, which is segmented into certain intervals referred to as bins $B_1, B_3, B_3, \dots, B_B\}$. Next, calculate the information value using equation [32].

$$pos_b = \frac{|\{(x_i, y_i): x_i \in B_b \; and \; y_i = g_2\}|}{|\{(x_i, y_i): y_i = g_2\}|}$$

$$neg_b = \frac{|\{(x_i, y_i): x_i \in B_b \; and \; y_i = g_1\}|}{|\{(x_i, y_i): y_i = g_1\}|}$$

$$\text{IV} = \sum(pos_b - neg_b) \times \log\left(\frac{pos_b}{neg_b}\right)$$

x_i represents data on the independent variable, y_i represents data on the dependent variable, B_b represents the bth bin, and $g_{1,2}$ represents categories on the dependent variable. The IV value can be a practical or poor predictor of the independent variable's relationship before constructing the classification model. Stojanovic [33] classifies IV values according to many parameters, as displayed in Table 2.

Table 2. IV Value categories

No	Information Value	Description
1	< 0.02	*Unpredictive*
2	0.02 ≤ IV < 0.1	*Weak*
3	0.1 ≤ IV < 0.3	*Medium*
4	≥ 0.3	*Strong*

Resampling Data

Resampling is a technique utilised to address the issue of imbalanced data. Imbalanced data refers to a situation where the answer variable contains a majority class and a minority class [34]. The majority class contains more data than the minority class, resulting in an imbalance in the distribution of data points between the two classes [35]. Imbalanced data might result in models that primarily categorise observations into the most common class and show minor sensitivity to the less common class [36]. The study utilised resampling techniques such as undersampling, oversampling, and Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance.

Discriminant Analysis with Mixed Independent Variables

Let us consider a random V divided into $(Z'|X')'$', where Z represents a vector of d category variables, and X represents a vector of p continuous variables. Each unique set of components $Z_1, Z_2, \dots, Z_k$_k in vector Z represents a state of the multinomial random variable W. Let the maximum number of potential states for

W be denoted as k, where $k = 2^d$. W_m represents state number-m. Please assume that the probability of getting cell W_m from vector $W\_$ is denoted as P_{im}, where m ranges from 1 to k and i from 1 to 2. The discriminant rule $g = 2$ is defined by obtaining state W_m from Z and then categorising v into π_1 if this assumption is used.

$$(\mu_{1m} - \mu_{2m})^T \sum^{-1} \left\{x - \frac{1}{2}(\mu_{1m} - \mu_{2m})\right\} \geq log(P_{2m}/P_{1m}) \quad (2)$$

And into π_2 if it doesn't satisfy the above condition. μ_{im} represents the population mean for class i in cell m. $\sum$ is the covariance matrix for the entire set of observations. x is the vector of p continuous variables, and P_{im} is the probability of an observation falling into population i in cell m. The formula for $\sum$ is as follows:

$$\hat{\Sigma} = \frac{1}{(n_1+n_2-2k)} \sum_{i=1}^{2} \sum_{m=1}^{k} \sum_{r=1}^{n_{im}} (x_{rim} - \hat{\mu}_{im})(x_{rim} - \hat{\mu}_{im})^T \quad (3)$$

Where n_i is the number of observations in population-i, k is the number of cells. x_{rim} is the vector of continuous variables for the r observation on cell-m and population-i and n_{im} is the number of observations in cell m, and population-i.

$$\hat{\mu}_{im} = \bar{x}_{im} \quad (4)$$

$\bar{x}_{im}$ is the mean of observations in cell-m and population-i. The equation for obtaining the value of $\hat{P}_{im}$ is as follow:

$$\hat{P}_{im} = \frac{n_{im}}{n_i} \quad (5)$$

Where n_{im} is the number of observations in cell-m and population-i, while n_i is the number of observations in population-i.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification technique that creates a hyperplane to separate data into different classes. The SVM hyperplane is determined by computing the margin of the hyperplane and identifying its highest point. The margin refers to the distance between the hyperplane and the nearest instance of each class. The nearest occurrence of the hyperplane is referred to as the support vector. The SVM hyperplane utilised possesses the widest margin between the classes. If the data is perfectly separable, the objective function of the hyperplane with the most significant margin can be defined as:

$$min\left\{\frac{1}{2}\|w\|^2\right\} \quad (6)$$

With constraints

$$y_i(w^T x_i + b) \geq 1 \quad (7)$$

With w being a vector orthogonal to the hyperplane, x_i s the training data vector, and y_i as the class for the i-th training data where $y_i \in \{-1,1\}$ and b satiesfies the following equation :

$$b = y_i - \sum_{i=1}^{ns} \alpha_i y_i \; K(x_i, x_j) \quad (8)$$

With α_i satiesfies the following equation :

$$0 \leq \alpha_i \leq C \quad (9)$$

The kernel function $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ defines the geometry of the hyperplane. Linear, polynomial, and radial basis function (RBF) kernels are frequently utilised kernel functions, as stated by Choubey [37] and Hussain [38]. The following equations define three types of kernels.

a. kernel linear : $K(x_i, x_j) = x_i\, x_j$ (10)

b. kernel *polynomial* : $K(x_i, x_j) = (x_i^T x_j + 1)^2$ (11)

c. kernel RBF : $K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right)$ (12)

Classification based on the optimal hyperplane function in equation is

$$g(x) = sign\left(f(x) = \sum_{i=1}^{ns} \alpha_i y_i \; K(x_i, x_j) + b\right) = \begin{cases} 1, & f(x) > 0 \\ -1, & f(x) < 0 \end{cases} \quad (13)$$

The measure of model goodness

An assessment of model quality is typically conducted to evaluate the model obtained [39]. The data is imbalanced, so the model is evaluated using the balanced accuracy metric. One metric used to evaluate a model's effectiveness with unbalanced data is balanced accuracy. [40] The formula for determining balanced accuracy requires the data provided in Table 3.

Table 3. Confusion matrix

Prediction	Actual	
	π_1	π_2
π_1	TP	FP
π_2	FN	TN

Balance accuracy Formula:

$$Balance\ Accuracy = \frac{sensitivity+specificity}{2} \quad (10)$$

Sensitivity dan specificity is obtained using the following equation:,

$$Sensitivity = \frac{TP}{TP+FN} \quad (11)$$

$$Specivicity = \frac{TN}{TN+FP}$$

The confusion matrix displays four possible combinations of predicted and actual values. The symbols π_i $i = 1,2$ denote individual categories of the answer variable. TP (true positive) is the count of observations correctly predicted to be in the first category. [41] False positive (FP) refers to the number of observations anticipated to be in one category but belong to a different category. [42] False negative (FN) refers to the number of observations anticipated to be in the second category but belong to the first category [43]. True negative (TN) is the count of observations correctly predicted to be in the second category. [44] The confusion matrix calculates different parameters to evaluate the model's performance [45]. Sensitivity and Specificity are utilised to compute balanced accuracy, which is especially beneficial for addressing imbalanced response variables [46].

Analysis Flowchart

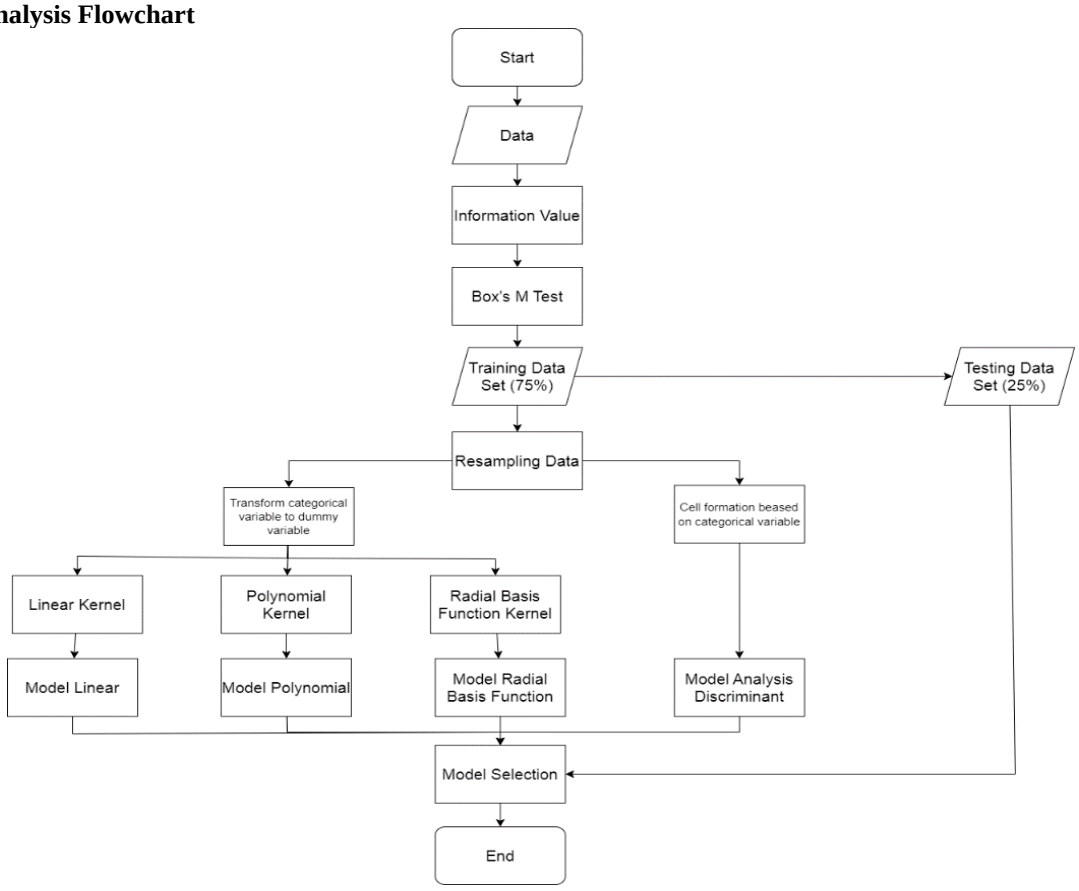

Figure 1. Flowchart comparing discriminant analysis and SVM analysis.

Analysis begins with examining the response and independent variables, and independent variables are chosen based on information values. A variance-covariance matrix equivalence test is conducted, with the data split into 75% training data and 25% testing data. Subsequently, data imbalance was rectified using undersampling, oversampling, and SMOTE methodologies. Modelling was conducted using Discriminant Analysis and SVM, and the balanced accuracy values were compared. The optimal model demonstrates the highest level of balanced accuracy [47].

RESULTS AND DISCUSSIONS

This study utilised data on the recovery time of COVID-19 patients in West Sumatra Province who took sungkai leaves during their rehabilitation. Data was gathered from 428 participants who had tested positive for COVID-19 and eaten sungkai leaves while recovering. Out of the participants, 338 recovered within (≤ 14 days) of the incubation period (>14 days)., whereas 90 recovered after this period. 78.97% of responders recovered within the incubation time, whereas 21.02% recovered after the incubation period. The data distribution shows an imbalance in the response variable, necessitating the employment of a resampling technique to address this issue. Resampling techniques employed are undersampling, oversampling, and SMOTE. These three methods on the dataset aim to standardise the number of observations in each category of response variables to address the data imbalance. The study examined independent variables such as gender, symptoms during COVID-19 infection, age, duration of symptom disappearance after confirmed infection, and the quantity of sungkai leaves utilised in preparing sungkai leaf mixture.

Data Exploration

This study uses mixed independent variables, including continuous and categorical variables. Figure 2 displays a summary of the continuous independent variables.

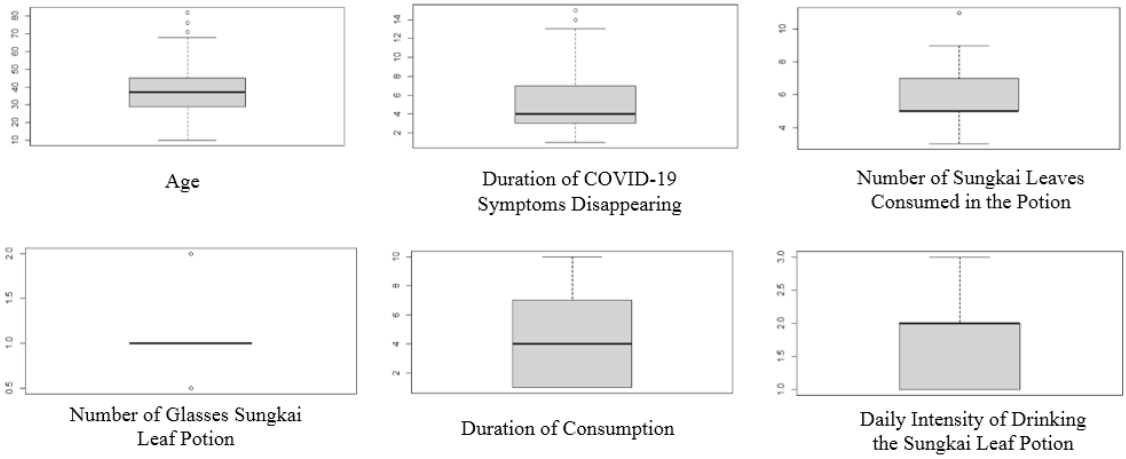

Figure 2. Boxplot for each continuous independent variable.

Figure 2 displays a boxplot for each continuous independent variable. Outliers were observed regarding age, duration of COVID-19 symptoms fading, number of sungkai leaves consumed in the concoction, and number of glasses of sungkai leaf concoction. Data exploration of categorical independent variables may be found in Figure 3.

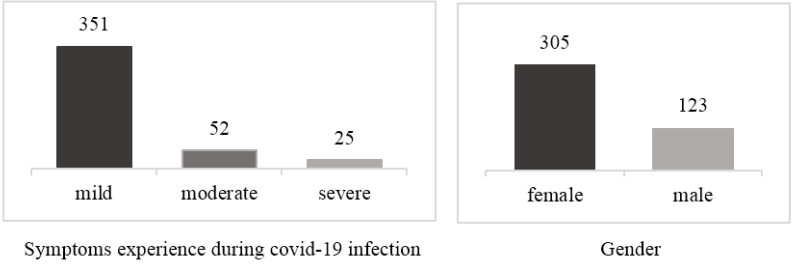

Figure 3. Bar chart of categorical independent variables

Figure 3 displays the descriptive Analysis of the category-independent variables. The dataset included 123 male respondents and 305 female respondents. There are 351 individuals with mild symptoms, 60 with moderate symptoms, and 28 with severe symptoms when exposed to COVID-19.

Preprocessing Data

The information value of each independent variable is utilised to select independent variables that impact the dependent variable. The information value indicates the impact of the independent variable on the dependent variable. The data values are displayed in Table 4.

Table 4. Information value of each independent variable

Independent Variables	IV	Description
Duration of COVID-19 Symptoms Disappearing	1.1009	*Strong Predictor*
Age	0.5048	*Strong Predictor*
Duration of Consumption	0.2548	*Medium Predictor*
Amount of Sungkai Leaves Consumed in the Potion	0.1801	*Medium Predictor*
Symptoms Experienced during COVID-19 Infection	0.1049	*Medium Predictor*
Number of Glasses Sungkai Leaf Potion	0.0459	*Weak Predictor*
Daily Intensity of Drinking the Sungkai Leaf Potion	0.0161	*Unpredictive*
Gender	0.0043	*Unpredictive*

Table 4 indicates that out of the eight independent variables, there are two variables with predictive solid power, three with moderate predictive power, one with weak predictive power, and two unpredictable factors. This study utilises independent variables categorised as strong and moderate predictors based on their information value. The study utilised the independent factors of age, duration of COVID-19 symptom resolution, duration of sungkai leaf intake, symptoms during COVID-19 infection, and the quantity of sungkai leaves consumed in the herbal remedy. The covariance homogeneity test was conducted on the continuous independent variables. Prior to performing discriminant Analysis, a covariance homogeneity test must be executed. Assessing covariance homogeneity with Box's M technique [48]. The Box's M test yielded a p-value of 0.333. If the p-value is more significant than α (0.05), it indicates that the data meets the condition of covariance homogeneity.

The study data necessitates a method to address unbalanced data due to the disproportionate distribution of the response variable data. Uneven data distribution on response variables can lead the model to exhibit bias towards categorising objects into the predominant class, diminishing prediction accuracy [49]. The dataset was divided into 75% for training data and 25% for testing data before modelling. Training data is utilised to construct the model, whereas testing data is employed to assess the model. The study utilised resampling to address the issue of data imbalance. The study involved resampling techniques such as undersampling, oversampling, and SMOTE. Table 5 displays the quantity of data following resampling.

Table 5. The number of data based on the response variable

Resampling	The number of data	
	Recover during the incubation period (≤14 days)	Recover after the incubation period (>14 days)
Without resampling	253 (79.06%)	67 (20.93%)
Undersampling	67 (50%)	67 (50%)
Oversampling	253 (50%)	253 (50%)
SMOTE	253 (50%)	253 (50%)

Based on Table 5, it can be seen that by using the undersampling, oversampling, and SMOTE technique, the data on the response variable has been balanced.

Discriminant analysis with mixed independent variables.

Categorical independent variables are handled differently in discriminant Analysis with mixed independent variables compared to other classification methods. Support Vector Machines often manage categorical independent variables by transforming them into dummy variables. In discriminant Analysis with mixed independent variables, categorical independent variables are handled by constructing cells according to the mixture of categories in the variable. One categorical independent variable, "Symptoms Experienced During COVID-19 Infection," will be employed in discriminant Analysis based on the variable selection method. This category-independent variable will create distinct categories in the discriminant Analysis. The Analysis categorises cells as "mild symptoms", "moderate symptoms", and "severe symptoms". The model

relies on these cells and the management of imbalanced data. Table 6 shows the cell arrangement derived from the training data.

Table 6. Proportion of training data

Analisis Diskriminan	Sel	Proportion of Training Data Based on Response Variable	
		Recovered within the incubation period (≤14 days)	Recovered after the incubation period (>14 days)
Without Resampling	Severe	0.0375	0.0187
	Moderate	0.1031	0.0343
	Mild	0.6500	0.1562
Undersampling	Severe	0.0223	0.0447
	Moderate	0.0597	0.0746
	Mild	0.4477	0.3507
Oversampling	Severe	0.0237	0.0573
	Moderate	0.0652	0.0928
	Mild	0.4110	0.3498
SMOTE	Severe	0.0237	0.0415
	Moderate	0.0652	0.0770
	Mild	0.4110	0.3814

Table 6 displays the percentage of each training data set using unbalanced data handling. A discriminant analysis model was created for each training data set. The discriminant analysis models were compared using their balanced accuracy scores. The optimal model is the one with the maximum balanced accuracy. Table 7 displays the balanced accuracy values for each model.

Table 7. Evaluation of model fit in discriminant analysis

No	Metode	Goodness-of-Fit Values		
		Balance Accuracy	Sensitivity	Specificity
1	Discriminant Analysis without resampling	0.6500	0.6520	0.6470
2	Discriminant Analysis with undersampling	0.6280	0.6090	0.6470
3	Discriminant Analysis with oversampling	0.6530	0.5650	0.7410
4	Discriminant Analysis with SMOTE	0.6654	0.6957	0.6353

Table 7 displays the overall accuracy value of four models in predicting the recovery duration of COVID-19 patients in West Sumatra using the testing data. The undersampling strategy in mixed independent variable discriminant Analysis reduces accuracy, which differs from earlier studies that found this method can improve accuracy [16]. Previous research has shown that oversampling and SMOTE are valuable methods for addressing unbalanced data and improving accuracy. [17] [22] The discriminant analysis model, utilising the SMOTE approach for unbalanced data handling, is considered the best due to its excellent balanced accuracy value of 66.54%.

Support Vector Machine (SVM)

Multiple kernel methods will be utilised to construct a hyperplane through the Support Vector Machine (SVM) technique. Kernel methods include linear, polynomial, and radial basis function (RBF) kernels. All three kernels will be utilised in the modelling procedure. This method also includes hyperparameter adjustment. SVM hyperparameter tuning involves adjusting the gamma (γ) and penalty (C) parameters. According to Hsu [50], a suitable range for gamma parameters (γ) is between 2^{-15}, 2^{-13},... $2^{3}$when the penalty parameter (C) falls within the range of 2^{-5}, 2^{-3},... 2^{15}. The study involves choosing the gamma value (γ) and penalty amount (C) within a specific range. Hyperparameter tuning will be conducted on all three kernel types using various datasets. The model's performance on the test data is presented in Table 8.

Table 8. Model performance metrics in discriminant analysis

No	Method	Goodness of fit value		
		Balance Accuracy	Sensitivity	Specificity
1	SVM linear kernel without resampling	0.5000	0.0000	1.0000
2	SVM linear kernel with undersampling	0.5767	0.5652	0.5882
3	SVM linear kernel with oversampling	0.6238	0.5652	0.6824
4	SVM linear kernel with SMOTE	0.6320	0.6522	0.6118
5	SVM polynomial kernel without resampling	0.5000	0.0000	1.0000
6	SVM polynomial kernel with undersampling	0.5158	0.0434	0.9882

7	SVM polynomial kernel with oversampling	0.5158	0.0434	0.9882
8	SVM polynomial kernel with SMOTE	0.5158	0.0434	0.9882
9	SVM RBF kernel without resampling	0.5000	0.0000	1.0000
10	SVM RBF kernel with undersampling	0.5409	0.4347	0.6470
11	SVM RBF kernel with oversampling	0.5468	0.4347	0.6558
12	SVM RBF kernel with SMOTE	0.5291	0.4347	0.6235

Table 8 displays the balanced accuracy values of 12 models used to predict test data. The utilisation of undersampling, oversampling, and SMOTE in SVM aligns with prior studies that have demonstrated the effectiveness of these methods in addressing data imbalance issues and enhancing accuracy. [16] [17] [22] The linear kernel SVM with SMOTE stands out as the top-performing model among the 12, boasting a balanced accuracy value of 63.20%.

Comparison of model performance

A study was conducted to compare Discriminant Analysis and Support Vector Machine (SVM) models to identify the optimal model for categorising the recovery duration of COVID-19 patients in West Sumatra. The models being compared are the top outcomes from each Analysis. The top-performing discriminant analysis model, utilising the SMOTE resampling technique, generates a confusion matrix in Table 9.

Table 9. Confusion matrix discriminant analysis with SMOTE

Prediction	Actual	
	>14 days	≤ 14 days
>14 days	16	31
≤ 14 days	7	54

Sensitivity, specificity, and balanced accuracy values from the confusion matrix are as follows,

$$Sensitivity = \frac{16}{16+7} = 0.6957$$

$$Specivicity = \frac{54}{54+31} = 0.6353$$

$$Balance\ Accuracy = \frac{sensitivity+specificity}{2} = \frac{0.6957+0.6353}{2} = 0.6654$$

The optimal SVM model, utilizing the SMOTE resampling method, yields the confusion matrix results presented in Table 10.

Table 10. Confusion matrix SVM with SMOTE

Prediction	Actual	
	>14 days	≤ 14 days
>14 days	15	33
≤ 14 days	8	52

Sensitivity, specificity, and balanced accuracy values from the confusion matrix are as follows,

$$Sensitivity = \frac{15}{15+8} = 0.6522$$

$$Specivicity = \frac{52}{52+33} = 0.6118$$

$$Balance\ Accuracy = \frac{sensitivity+specificity}{2} = \frac{0.6522+0.6118}{2} = 0.6320$$

Based on the evaluation results of the best model from both methods, you can see the comparison of these two models in Table 11.

Table 11. Comparative evaluation of model fit

No	Metode	Goodness of fit value		
		Balance Accuracy	Sensitivity	Specificity
1	Discriminant Analysis with SMOTE	0.6654	0.6957	0.6353
2	SVM Linear Kernel with SMOTE	0.6320	0.6522	0.6118

Table 11 displays the adequacy of fit for each Analysis. The Discriminant Analysis applied to imbalanced data with the SMOTE method yielded a sensitivity of 69.57%, Specificity of 63.53%, and balanced accuracy of 66.54%. The linear kernel Support Vector Machine with SMOTE achieved a sensitivity of

65.22%, Specificity of 61.18%, and balanced accuracy of 63.20%. The Discriminant Analysis model achieved the most excellent balanced accuracy rating of 66.54% across the two analyses. Discriminant Analysis outperforms the Support Vector Machine approach in predicting COVID-19 recovery time in West Sumatra. Discriminant Analysis is more effective than SVM in classifying observations in the recovery time data of COVID-19 patients in West Sumatra, as indicated by the highly balanced accuracy value. Furthermore, the high balanced accuracy value suggests that discriminant Analysis is better at categorising observations into major and minor data classes than SVM.

CONCLUSION

Analysis results indicate that addressing data imbalance using SMOTE yields the highest balanced accuracy for both approaches in this dataset. Discriminant Analysis with data balancing using SMOTE achieves a balanced accuracy of 66.54%. However, employing the support vector machine technique with a linear kernel and data balancing by SMOTE yielded a balancing accuracy of 63.20%. The results indicate that the discriminant analysis model outperforms the support vector machine on this dataset.

Recommendations for future research based on the study findings. Future research should investigate the impact of underlying disorders or comorbidities on the duration of COVID-19 recovery using COVID-19 data. Another recommendation is to perform research utilising discriminant analysis and support vector machine (SVM) approaches on a spatial level, incorporating mixed independent variables.

REFERENCES

[1] Z. Wu and J. M. McGoogan, "Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China," *Jama*, vol. 323, no. 13, p. 1239, 2020, doi: 10.1001/jama.2020.2648.

[2] P. K. Perera and A. C. B. Meedeniya, "Curcumin as a Potential Treatment for COVID-19," *Front. Pharmacol.*, vol. 12, no. September 2021, pp. 1–10, 2021, doi: 10.3389/fphar.2021.675287.

[3] Ö. Güngör and H. Baykal, "Attitudes toward herbal medicine for COVID-19 in healthcare workers: A cross-sectional observational study," *Med. (United States)*, vol. 102, no. 38, p. E35176, 2023, doi: 10.1097/MD.0000000000035176.

[4] J. Ren, A. Zhang, and X. Wang, "Traditional Chineses Medicine for Covid-19 Treatment," *Pharmacol. Res.*, p. 104743, 2020, doi: 10.1016/j.phrs.2020.104743.

[5] R. F. Noor'An, Karmilasanti, and C. B. Wiati, "Potential and distribution of Vitex sp and Peronema canescens jack as anti -COVID 19 plants in East Kalimantan Province, Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 886, no. 1, 2021, doi: 10.1088/1755-1315/886/1/012030.

[6] A. P. Yani, A. Ruyani, I. Ansyori, and R. Irwanto, "UJI POTENSI DAUN MUDA SUNGKAI (Peronema canescens) UNTUK KESEHATAN (IMUNITAS) PADA MENCIT (Mus.muculus) The Potential Test of Sungkai Young Leaves (Peronema canescens) to Maintain Goodhelth (Immunity)in Mice (Mus musculus)," *Semin. Nas. XI Pendidik. Biol. FKIP UNS 245*, pp. 245–250, 2014.

[7] M. Kakehashi and S. Kawano, *Fundamentals of Mathematical Models of Infectious Diseases and Their Application to Data Analyses*, 1st ed., vol. 36. Elsevier B.V., 2017. doi: 10.1016/bs.host.2017.06.002.

[8] S. A. Lauer *et al.*, "The incubation period of coronavirus disease 2019 (CoVID-19) from publicly reported confirmed cases: Estimation and application," *Ann. Intern. Med.*, vol. 172, no. 9, pp. 577–582, 2020, doi: 10.7326/M20-0504.

[9] C. Elias, A. Sekri, P. Leblanc, M. Cucherat, and P. Vanhems, "The incubation period of COVID-19: A meta-analysis," *Int. J. Infect. Dis.*, vol. 104, pp. 708–710, 2021, doi: 10.1016/j.ijid.2021.01.069.

[10] E. Zdravevski, P. Lameski, A. Kulakov, and D. Gjorgjevikj, "Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets," *2014 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2014*, vol. 2, pp. 387–394, 2014, doi: 10.15439/2014F500.

[11] A. J. Izenman, *Linear Discriminant Analysis 8.1*. 2013. doi: 10.1007/978-0-387-78189-1.

[12] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009, doi: 10.1142/S0218001409007326.

[13] R. Van Den Goorbergh, M. Van Smeden, D. Timmerman, and Ben Van Calster, "The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression," *J. Am. Med. Informatics Assoc.*, vol. 29, no. 9, pp. 1525–1534, 2022, doi:

10.1093/jamia/ocac093.
[14] R. Rofik, R. Aulia, K. Musaadah, S. S. F. Ardyani, and A. A. Hakim, "Optimization of Credit Scoring Model Using Stacking Ensemble Learning and Oversampling Techniques," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, pp. 11–20, 2023, doi: 10.52465/joiser.v2i1.203.
[15] J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00738-z.
[16] Q. Shi and H. Zhang, "Fault Diagnosis of an Autonomous Vehicle with an Improved SVM Algorithm Subject to Unbalanced Datasets," *IEEE Trans. Ind. Electron.*, vol. 68, no. 7, pp. 6248–6256, 2021, doi: 10.1109/TIE.2020.2994868.
[17] L. Qadrini, "Oversampling , Undersampling , Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017," vol. 3, no. 4, pp. 386–391, 2022, doi: 10.47065/josyc.v3i4.2154.
[18] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020, doi: 10.52465/joscex.v1i1.5.
[19] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Sci. J. Informatics*, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.
[20] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.
[21] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A Neural Network Ensemble with Feature Engineering for Improved Credit Card Fraud Detection," *IEEE Access*, vol. 10, pp. 16400–16407, 2022, doi: 10.1109/ACCESS.2022.3148298.
[22] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.
[23] A. H. Ali, Z. F. Hussain, and S. N. Abd, "Big Data Classification Efficiency Based on Linear Discriminant Analysis," *Iraqi J. Comput. Sci. Math.*, vol. 1, no. 1, pp. 9–14, 2020, doi: 10.52866/ijcsm.2019.01.01.001.
[24] M. A. Mukid and T. Widiharih, "Model Penilaian Kredit Menggunakan Analisis Diskriminan Dengan Variabel Bebas Campuran Biner Dan Kontinu," *Media Stat.*, vol. 9, no. 2, p. 107, 2017, doi: 10.14710/medstat.9.2.107-117.
[25] A. Mbina Mbina, G. M. Nkiet, and F. Eyi Obiang, "Variable selection in discriminant analysis for mixed continuous-binary variables and several groups," *Adv. Data Anal. Classif.*, vol. 13, no. 3, pp. 773–795, 2019, doi: 10.1007/s11634-018-0343-0.
[26] N. Walidaini, M. A. Mukid, A. Prahutama, and A. Rusgiyono, "Analisis Diskriminan Berganda Dengan Peubah Bebas Campuran Kategorik Dan Kontinu Pada Klasifikasi Indeks Prestasi Kumulatif Mahasiswa," *Media Stat.*, vol. 10, no. 2, p. 71, 2017, doi: 10.14710/medstat.10.2.71-83.
[27] S. Guhathakurata, S. Kundu, A. Chakraborty, and J. S. Banerjee, "A novel approach to predict COVID-19 using support vector machine," no. January, pp. 351–364, 2020.
[28] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An introduction to statistical learning (2nd ed.), website," *Springer texts*, vol. 102, p. 618, 2021.
[29] N. I. Mahat, W. J. Krzanowski, and A. Hernandez, "Variable selection in discriminant analysis based on the location model for mixed variables," pp. 105–122, 2007, doi: 10.1007/s11634-007-0009-9.
[30] S. Guhathakurata, K. Souvik, A. Chakraborty, and J. S. Banerjee, "A novel approach to predict COVID-19 using support vector machine," *Glob. Heal.*, vol. 167, no. 1, pp. 1–5, 2020.
[31] B. Lund and D. Brotherton, "Information Value Statistic," *Mark. Assoc. LLC*, no. 2010, pp. 1–18, 2013.
[32] C. Nguyen, X. Li, S. Blanton, and X. Li, "Efficient Classification via Partial Co-Training for Virtual Metrology," *IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA*, vol. 2020-Septe, pp. 753–760, 2020, doi: 10.1109/ETFA46521.2020.9212012.
[33] B. Stojanović *et al.*, "Follow the trail: Machine learning for fraud detection in fintech applications," *Sensors*, vol. 21, no. 5, pp. 1–43, 2021, doi: 10.3390/s21051594.
[34] J. Kim, "suffer from its expensive data acquisition process and the la- Jinwoo Shin Korea Advanced Institute of Science and Technology (KAIST) Daejeon , South Korea M2m : Imbalanced

Classification via Major-to-minor Translation," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020.
[35] B. Pes, "Learning from high-dimensional and class-imbalanced datasets using random forests," *Inf.*, vol. 12, no. 8, 2021, doi: 10.3390/info12080286.
[36] S. J. Yen and Y. S. Lee, "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset," *Lect. Notes Control Inf. Sci.*, vol. 344, pp. 731–740, 2006, doi: 10.1007/978-3-540-37256-1_89.
[37] D. K. Choubey, S. Tripathi, P. Kumar, V. Shukla, and V. K. Dhandhania, "Classification of Diabetes by Kernel Based SVM with PSO," *Recent Adv. Comput. Sci. Commun.*, vol. 14, no. 4, pp. 1242–1255, 2019, doi: 10.2174/2213275912666190716094836.
[38] M. Hussain, S. K. Wajid, A. Elzaart, and M. Berbar, "A comparison of SVM kernel functions for breast cancer detection," *Proc. - 2011 8th Int. Conf. Comput. Graph. Imaging Vis. CGIV 2011*, pp. 145–150, 2011, doi: 10.1109/CGIV.2011.31.
[39] A. Onan and M. A. Tocoglu, "A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021, doi: 10.1109/ACCESS.2021.3049734.
[40] M. Gösgens, A. Zhiyanov, A. Tikhonov, and L. Prokhorenkova, "Good Classification Measures and How to Find Them," *Adv. Neural Inf. Process. Syst.*, vol. 21, no. NeurIPS, pp. 17136–17147, 2021.
[41] D. Chicco, N. Tötsch, and G. Jurman, "The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Min.*, vol. 14, pp. 1–22, 2021, doi: 10.1186/s13040-021-00244-z.
[42] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/8387680.
[43] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.
[44] A. Singh and R. Kumar, "Heart Disease Prediction Using Different Machine Learning Algorithms," *Proc. - 2022 IEEE World Conf. Appl. Intell. Comput. AIC 2022*, pp. 60–65, 2022, doi: 10.1109/AIC55036.2022.9848885.
[45] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice," *Front. Public Heal.*, vol. 5, no. November, pp. 1–7, 2017, doi: 10.3389/fpubh.2017.00307.
[46] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," p. 3, 2022.
[47] J. S. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data," *SAS Glob. Forum*, vol. 942, pp. 1–12, 2017.
[48] T. Hungsapruek, "Compare Between Personal Factors & Healthcare Service Needs after Becoming Senior Citizens," vol. 20, no. 3, pp. 1419–1439, 2021.
[49] V. Kumar *et al.*, "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," *Healthc.*, vol. 10, no. 7, 2022, doi: 10.3390/healthcare10071293.
[50] C. W. Hsu, C. C. Chang, and C. J. Lin, "A Practical Guide to Support Vector Classification," *http://www.csie.ntu.edu.tw/~cjlin*, 2016.