Codeverse: The Journal of Emerging DigitalRealities
Vol. 01 No.
ISSN: xxxx-xxxx, DOI: xxxxx-xxxxx

Support Vector Machine (SVM) For Classification of Disadvantaged

Nur Fadila Palisoa[1], L. J. Sinay[2], Yudistira[3], M. Y. Matdoan[4]

[1,2,3,4] Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Pattimura
M. J. Leimena Street, Kampus Unpatti-Poka, Ambon, 97233, Indonesia

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Classification;
Disadvantaged;
Support vector machine.

ABSTRACT

Equitable development and development of regions is very important to ensure regional socio-economic equality and balance, for this reason it is necessary to classify regions in order to determine priorities in equitable develop ent that is fast and on target. The classification method used is Support Vector Machine (SVM). This research aims to analyze the accuracy of the classification of disadvantaged areas in Maluku Province with the data source used is secondary data sourced from several Statistich Maluku Province publications. Based on the results of classification of disadvantaged areas using SVM with the RBF kernel, it has the best results with parameters cost = 0.1 and gamma = 1 and the resulting classification accuracy level is 95.4% and the AUC value = 0.9285 which is classified as very good classification results.

Corresponding Author:

Nur Fadila Palisoa
Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Pattimura
97233, Ambon, Indonesia
Email: nurfadilapalisoa@gmail.com

1. INTRODUCTION

Underdeveloped regions are districts whose regions and communities are less developed than other regions on a national scale. According to the Presidential Regulation (Perpes) N0.63 of 2020, the determination of underdeveloped regions is seen with a composite index based on 6 criteria of underdevelopment. The criteria used are community economy, human resources, facilities and infrastructure, regional financial capacity, accessibility, and regional characteristics.

The government is trying to overcome the existence of underdeveloped regions, one of which is the National Strategy for the Acceleration of Regional Development 2020-2024. Maluku Province is one of the regions located in Eastern Indonesia with 6 districts classified as underdeveloped regions, namely Tanimbar Islands District, Aru Islands District, West Seram District, East Seram District, South Buru District, and Southwest Maluku District.

Equitable development and regional development are essential to ensure regional socio-economic equality and balance to prevent underdeveloped regions. For this reason, it is necessary to classify regions in order to determine priorities in equitable development that is fast and targeted. One of the statistical methods that can be used in classification is Support Vector Machine (SVM). SVM is a classification method that can separate two data sets from two different classes by maximizing the boundary of the separating function (hyperplane). The advantage of SVM compared to other methods is that it can produce a good classification model with higher accuracy[1].

Research on the classification of districts in East Java Province using the SVM method, by looking at indicators of backwardness in East Java Province, obtained an average accuracy of 79.46% [2]. In addition, similar research on the classification of underdeveloped regions in Indonesia, resulted in an accuracy rate of 92.2% using the SVM method in this study[3]. Based on this description, this research examines the characteristics of underdeveloped and not underdeveloped regions in Maluku Province measuring the accuracy of classification provisions with Support Vector Machine (SVM).

.

2. METHOD

2.1 Data Sources and Research Variables

This research is a type of quantitative research by applying the SVM method in classifying disadvantaged areas in Maluku Province. The data used in the study is secondary data obtained from the Central Statistics Agency (BPS). There are 12 variables used to measure the classification of disadvantaged regions in Maluku Province, namely.

Table 1. Research Variables

Variable	Description	Scale
Class	1 (Non-Least Developed Region) 0 (Disadvantaged Areas)	Nominal
PPM	Percentage of Poor Population	Rasio
PRRP	Average expenditure per capita	Rasio
AHH	Life Expectancy	Rasio
RLM	Average Years of Schooling	Rasio
AMH	Literacy Rate	Rasio
PJJPJTA	Length of road with the widest road surface type asphalt/concrete	Rasio
PRTPL	Percentage of households using electricity	Rasio
PRTMTS	Percentage of Households with a Mobile Phone	Rasio
PRTPAB	Percentage of Households Using Clean Water	Rasio
JPRS	Number of health centres and hospitals	Rasio
JD	Number of doctors	Rasio
JSDSMP	Number of primary schools - junior secondary schools	Rasio

2.2 Data Mining

Data mining (nowledge Discovery in Databaseis an activity related to data collection, utilizing historical data to find knowledge, information, regularities, patterns or relationships in large data[4].
The tasks and functions of data mining can be grouped into parts, namely[5].

1. Classification
   Generalize known structures to apply to new data.
2. Clustering
   Grouping data that has no known class labels into a certain number of groups according to a measure of similarity.
3. Regression
   Find a function that models the data with the minimum possible prediction error.
4. Anomaly detection
   Detect unusual data, which can be outliers, changes or deviations that may be very important and need further investigation.
5. Association rule learning
   Look for relationships between variables.
6. Summarization
   Provides simpler data representation, including visualization and report generation.

2.2.1 Standarisasi Data

Before conducting the analysis process, one thing that needs to be considered is whether the data units have a large difference or not. For this reason, it is necessary to equalize the variables called the data standardization technique[6].

$$X_s = \frac{x_i - \min(x)}{\max - \min} \quad (1)$$

2.2.2 Data Training and Data Testing

Training data is part of the dataset used to build a classification model. The resulting model is the measure of the extent to which the classification succeeds in making predictions correctly on the testing data. The proportion between training data and testing data is not inherently binding, but in order to avoid large variations in the model, it is recommended that the training data has a larger proportion than the testing data [7]. The proportions commonly used by researchers are 60:40, 65:35, 70:30, 75:25 or 80:20 [8].

2.2.3 Support Vector Machine (SVM)

Support vector machine (SVM) is a relatively new technique for prediction, both in classification and regression cases. In the SVM method, the classifier function is found by separating two data sets from two different classes such as class (1) and (-1). The convex optimization problem in this method is based on quadratic programming, so the classifier sought so far is a linear function. The basic principle of SVM is a linear classifier which is further developed on non-linear problems by incorporating the concept of "kernel trick" in a high-dimensional workspace (feature space) for classification processing [9].

In linear SVM classification, suppose there are m research data $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ with $x_i \in \mathbb{R}^n$ is the sample data and $y_i \in \{1,0\}$ is the target or class of the data sample. Suppose that the data for the two classes are linearly separable then the hyperplane for the two classes can be written in the following equation:

$$f(x) = xw + b = 0 \quad (2)$$

with $w \in \mathbb{R}^{n\times 1}$ are the weight parameters and $b \in \mathbb{R}$ is the bias parameter, can be calculated with equation (3).

$$w = \sum_{i=1}^{ns} \alpha_i y_i x_i \quad \text{dan} \quad b = \frac{1}{sv} \sum_{i=1}^{ns} (\alpha_i - x_i\, w) \quad (3)$$

and applies:

$$xw + b > 0 \text{ for } y_i = 1$$
$$xw + b < 0 \text{ untuk } y_i = 0 \quad (4)$$

For example H: $xw + b = 0$ is the separator you want to find while H_1: $xw + b = 1$ and H_2: $xw + b = 0$ is the separator of class 1 and class 0. To get the optimal value of H, the distance of H_1 and H_2 to H must be

the same under the condition that there are no data samples between H_1 and H_2 and the distance of H_1 and H_2 is the maximum distance.

To maximize the distance between H_1 and H_2, positive data samples located at H_1 and negative data samples located at H_2 are used. This data sample is called a support vector because its function is to determine the optimal separator. Other data samples can be discarded or moved towards H_1 and H_2 as long as they do not pass through each separator.

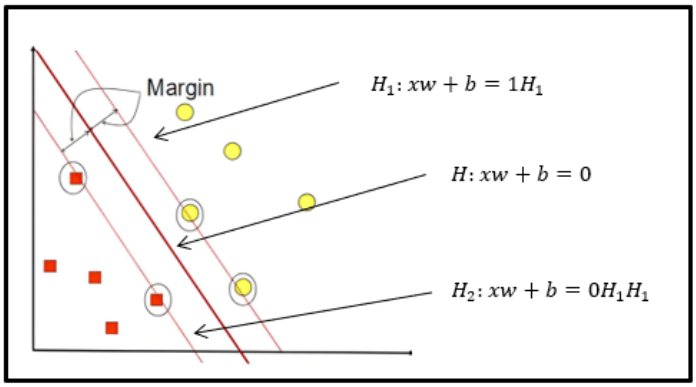

Figure 1. Optimal separation of two classes [11].

For example $(x_0, y_0) \in \mathbb{R}^2$ any point then the distance of this point to the line $Ax + By + C = 0$ is

$$\frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}} \tag{5}$$

So that the data sample distance x located on H_1 to H are

$$\frac{|x\,w + b|}{\sqrt{w^t w}} = \frac{1}{\|w\|} \tag{6}$$

Since the distance between H_1 and H_2 is the same, the distance between H_1 and H_2 is $\frac{2}{\|w\|}$. [9]

The problem of maximizing $\frac{2}{\|w\|}$ is equivalent to the problem of minimizing $\frac{\|w\|^2}{2}$, with the condition that there is no sample data between H_1 and H_2 that is $x_i w + b \geq 1$ for $y_i = 1$ and $x_i w + b \leq 1$ for $y_i = 0$ if you combine the two conditions, you get $y_i(x_i w + b) \geq 1$. Thus, the problem of finding the parameters w and b to obtain the optimal separation is a quadratic programming problem.

$$\min_{w,b} \frac{1}{2} w^t w$$

with constraints (7)

$$y_i(x_i w + b) \geq 1 \qquad , i = 1, \dots, m$$

It is usually difficult to solve the above prime form, so the primal form is converted to its dual form by introducing the Lagrange multiplier[10].

For exmple $\alpha \in \mathbb{R}^{m \times 1}$ is the Lagrange multiplier then the quadratic programming problem (7) changes to

$$L(w, b, \alpha) = \frac{1}{2} w^t w - \sum_{i=1}^{m} \alpha_i \big(y_i(x_i w + b)\big) + \sum_{i=1}^{m} \alpha_i \tag{8}$$

The solution of this problem must fulfill the Karush-Kuhm-Tucker conditions, namely

1. $$\frac{\partial L}{\partial w} = 0 \rightarrow w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0$$
$$\rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{9}$$

2. $$\frac{\partial L}{\partial w} = 0 \rightarrow 0 - \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{10}$$

$$\rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$$

3. $\alpha_i(y_i(x_i w + b) - 1) = 0$ (11)

4. $\alpha_i \geq 0$ (12)

Thus the dual form obtained is

$$\max L(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j^T \quad (13)$$

with constraints

$\sum_{i=1}^{m} \alpha_i y_i = 0, \quad \alpha_i \geq 0$ dimana $i = 1,2,\ldots,m$

The weight and bias parameters can be calculated with the equation

$$w = \sum_{i=1}^{Nsv} \alpha_i y_i x_i \qquad b = \frac{1}{Nsv}\sum_{i=1}^{Nsv} (y_i - x_i w) \quad (14)$$

SV is the set of support vectors and $i \in SV$ if $\alpha_i \neq 0$. NSV is the number of support vectors. By using the equation.

$$f(x) = x\,w + b$$

Then the input data $x \in \mathbb{R}^n$ which are newly classified into

$$\begin{cases} \text{Class } 1, & \text{if } f(x) > 0 \\ \text{Class } 0, & \text{if } f(x) < 0 \end{cases} \quad (15)$$

Furthermore, if there is a case of imperfect separation, to solve the problem, a variable is given slack μ which is non-negative ($\mu \geq 0$) is substituted into equation (7) so that it becomes

$y_i(x_i w + b) \geq 1 - \mu_i, \quad i = 1,\ldots,m$

Meanwhile, the objective function adds a positive parameter C so that it becomes

$$\frac{1}{2} w^t w - C\sum_{i=1}^{m} \mu_i \quad (16)$$

with constraints

$$y_i(x_i w + b) \geq 1 - \mu_i, \quad \mu_i \geq 0, \quad i = 1,\ldots,m$$

By using the lagrange multiplier $\alpha \in \mathbb{R}^{m\times 1}$ then the primal form of equation (16) can be converted into the dual form as follows:

$$\max L(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j^T \quad (17)$$

with constraints

$\sum_{i=1}^{m} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$ dimana $i = 1,2,\ldots,m$

In general, the dominant real world is rarely linear, mostly non-linear. To solve non-linear problems, SVM is modified by including a kernel function (kernel trick). Kernel trick provides convenience because in the learning process of SVM, to know the support vector, it is only enough to know the kernel function used, and there is no need to know the form of the non-linear function Φ **[11]**. This kernel method works by mapping the input data to a higher dimensional feature space using a function such as $\boldsymbol{\phi}$, for example suppose $u = u_1, u_2$ is the input data in $\mathbb{R}^2$ and $\phi(u) = \left(1, \sqrt{2}u_1, \sqrt{2}u_2, {u_1}^2, {u_2}^2, \sqrt{u_1}u_2\right)$ is the higher dimensional feature space input data $\mathbb{R}^5$. It is expected that the input data mapped to the feature space will be linearly separated so that the optimal separation can be found.

For example $x \rightarrow \phi(x)$ then equations (16) and (17) can be written as

$$\max \psi(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j \phi_i \phi_j^T \quad (18)$$

with constraints

$$\sum_{i=1}^{m} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \text{ dimana } i = 1,2,\ldots,m$$

The weight and bias parameters can be calculated with the equation

$$w = \sum_{i=1}^{Nsv} \alpha_i y_i \phi(x_i) \qquad b = \frac{1}{Nsv} \sum_{i=1}^{Nsv} (y_i - w^T \phi(x_i)) \tag{19}$$

While the optimal separator converts to

$$f(x) = w^T \phi(x) + b = 0 \tag{20}$$

The kernel functions that are usually used in SVM are as follows[12]:

Table 2. Kernel functions in SVM

Kernel Type	Definition
Linear	$K(x_i, x_j) = x_i . x_j$
Polynomial	$K(x_i, x_j) = (x_i . x_j + c)^d$
RBF	$K(x_i, x_j) = \exp\left(-\gamma \left\|\|x_i - x_j\right\|\right)^2$
Sigmoid	$K(x_i, x_j) = \tan(\gamma(x_i . x_j) + c)$

by using the concept of kernel function above, **Equations (18) - (20)** change into

$$\max \psi(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{21}$$

with constraints

$\sum_{i=1}^{m} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$ dimana $i = 1,2, \ldots, m$

The bias parameter can be calculated with the equation

$$b = \frac{1}{Nsv} \sum_{i=1}^{Nsv} \left(y_i - \sum_{i=1}^{Nsv} \alpha_i y_i K(x_i, x_j) \right) \tag{22}$$

Meanwhile, the optimal separator changes to

$$f(x) = \sum_{i=1}^{Nsv} \alpha_i y_i K(x_i, x_j) + b = 0 \tag{23}$$

Table 3. SVM parameters

Kernel Type	Parameters	Value
Polynomial	Cost (c), degree (d), gamma (γ)	$c = 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3$ $d = 2,3$
RBF	Cost (c), gamma (γ)	$c = 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3$ $\gamma = 10^{-1}, 10^0, 10^1, 10^2$
Sigmoid	Cost (c), gamma (γ)	$c = 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3$ $\gamma = 10^{-1}, 10^0, 10^1, 10^2$

2.2.4 Classification Performance Evaluation

A classification model is said to be the best model based on the highest accuracy, specificity, sensitivity, AUC criteria. A classification model is said to be the best model based on the highest accuracy, specificity, sensitivity, AUC criteria [13]. In addition, the model classification category based on the value is summarized in **Table 4**[14]:

Table 4. Category of AUC value

AUC value	Classification Result Categories
0.91-1.0	Very good
0.81-0.9	Good
0.71-0.8	Good enough
0.61-0.7	Bad
≤ 0.6	Failed

Classification performance measurement is carried out to determine the performance of the classification when predicting the class of data. The results of the number of correct or incorrect observations from the SVM classified by the model can be organized in a confusion matrix as follows[15].

Table 5. Coufusion matrix

Prediction	Reference	
	0	1
0	TP	FP
1	FN	TN

Based on Table 5, the following values of accuracy, sentivity, specificity, and AUC can be calculated.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \quad (24)$$

$$\text{Sensivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (25)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (26)$$

$$\text{AUC} = \frac{1}{2}(\text{sensivity} + \text{specificity}) \times 100\% \quad (27)$$

2.2.5 Work Procedure

The procedure of this research is as follows.

1 Conduct descriptive analysis to see an overview of the research data.
2 Standardization of attribute variables using Min-Max Standardization.
3 Divide data into training data and testing data.
4 Determining the Best Parameters along with the Best Kernel Function.
5 Build SVM classification model on training data.
6 Evaluate SVM classification performance on testing data.
7 Conclusion.

3. RESULTS AND DISCUSSION

3.1 Data Standardization

Before building the classification model of disadvantaged regions using SVM, the data will first be standardised in value, the standardisation process is done with the Min-Max standardisation method for each variable.

3.2 Determination of Training and Testing Data

After the standardisation process is complete, the next step is the division of training data and testing data, the first proportion of training and testing data division is the proportion of 60:40. Of the 110 observations of research data, 66 observations were used to build the SVM classification model while the remaining 44 observations were used as testing data, the second proportion of 65: 35 data was divided into 71 training data and 39 testing data. third proportion 70: 30 resulted in 77 observations as training data while 33 observations as testing data. The fourth proportion of 75:25 with 83 observations as training data and 27 observations as testing data and the proportion of 80:20 produces 88 observations of research data as training data while 22 observations as testing data.

3.3 Support Vector Machine (SVM)

Based on Figure 2, which is a training data distribution plot, it can be seen that the red coloured points are the data distribution for class 0 (Disadvantaged Areas) and the black coloured points are the data distribution for class 1 (Non-Disadvantaged Areas). The pattern of data distribution in Figure 3 cannot be

separated by 1 straight line or linear, thus kernel trick assistance is needed in classifying disadvantaged regions in Maluku Province using Support Vector Machine (SVM). The kernels that will be used in this research are polynomial kernel, RBF kernel, and sigmoid kernel.

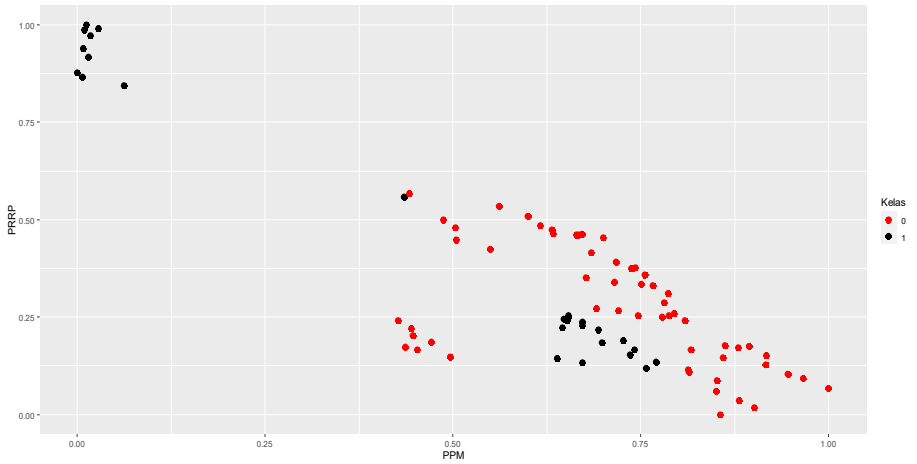

Figure 2. Data distribution pattern

The next step is to find the best parameters based on Table 2 for each kernel, the results can be seen in **Table 6** below:

Table 6. Best Parameter Results

Rasio data	Kenel	gamma	Cost	degree	akurasi
80:20	**RBF**	**0.1**	**2**	**-**	**1**
	Polynomial	-	8	2	0.9432
	Sigmoid	0.1	1	-	0.9432
75:25	**RBF**	**0.1**	**1**	**-**	**1**
	Polynomial	-	8	2	0.9518
	Sigmoid	0.1	1	-	0.9639
70:30	**RBF**	**0.1**	**1**	**-**	**1**
	Polynomial	-	8	2	0.9351
	Sigmoid	0.1	1	-	0.961
65:35	**RBF**	**0.1**	**1**	**-**	**1**
	Polynomial	-	8	2	0.9437
	Sigmoid	0.1	1	-	0.9577
60:40	**RBF**	**0.1**	**1**	**-**	**1**
	Polynomial	-	8	2	0.9394
	Sigmoid	0.1	1	-	0.394

Based on **Table 6**, from each proportion of data division, the best parameters are tested for each kernel with a repetition of 70 times for one proportion of data division. The highest average accuracy is produced by the RBF kernel in each data division proportion with a training data accuracy rate of 100%. So it can be said that the best SVM model is to use the RBF kernel.

After obtaining the SVM model form, the next step is implemented to the testing data, this process is completed with the help of software whose results are described in the classification performance evaluation.

3.4 Classification Performance Evaluation

Classification performance evaluation is continued based on the highest accuracy results in Table 7. The results of its implementation on testing data can be seen in the following explanation:

1. Kernel RBF 80:20

The results of the RBF Kernel Confusion Matrix are as follows.

Table 7. Coufusion Matrix Kernel RBF 80:20

	0	1
0	15	2
1	0	5

Based on **Table 7**, it explains that of the 22 testing data there are 17 data classified in the class of disadvantaged areas (0) and 5 data classified in class 1, with 2 data that are actually class 1 after SVM classification into class 0. Based on the coufusion matrix table, the accuracy, sentivity, specificity, and AUC values can be calculated as follows:

$$\text{Accuracy} = \frac{5+15}{5+15+0+2} = 0{,}90$$

$$\text{Sensivity} = \frac{15}{15+0} = 1$$

$$\text{Specificity} = \frac{5}{5+2} = 0.7143$$

$$\text{AUC} = \frac{1}{2}(1+0.7143) \times 100\% = 0.857$$

Based on the calculation of the RBF 80:20 kernel classification performance, the level of accuracy or similarity of the SVM classification results with the actual data is 90% with the sensivity number or the model's ability to correctly identify positive class data is 100% and the specificity number or the model's ability to correctly identify negative class data is 71.43%, and the AUC value is 0.857.

2. Kernel RBF 75:25

The results of the RBF Kernel Confusion Matrix are as follows.

Table 8. Confusion Matrix Kernel RBF 75:25

	0	1
0	19	2
1	0	6

Seen in **Table 8**, using the 75:25 RBF kernel model on 27 testing data, 21 data are classified in class 0 (disadvantaged areas) and 6 data are classified in class 1 (areas not disadvantaged) with 19 data correctly predicted in class 0 (disadvantaged areas) with 2 actual data in class 1 after SVM classification including class 0 (disadvantaged areas) then 6 data correctly predicted in class 1 (areas not disadvantaged). Based on the coufusion matrix table, the accuracy, sentivity, specificity, and AUC values can be calculated as follows:

$$\text{Accuracy} = \frac{6+19}{6+19+0+2} = 0.925$$

$$\text{Sensivity} = \frac{19}{19+0} = 1$$

$$\text{Specificity} = \frac{6}{6+2} = 0.75$$

$$\text{AUC} = \frac{1}{2}(1+0.75) \times 100\% = 0.875$$

Based on the calculation of the RBF 75:25 kernel classification performance, the resulting SVM classification result is 92.50% with a sensivity rate or the model's ability to correctly identify positive class data is 100% and the specificity rate or the model's ability to correctly identify negative class data is 75%, and the AUC value is 0.875.

3. Kernel RBF 70:30

The results of the RBF Kernel Confusion Matrix are as follows.

Table 9. Confusion Matrix Kernel RBF 70:30

	0	1
0	23	2
1	0	8

Based on **Table 9**, it can be seen from the 33 training data, 25 data are classified as class 0 (disadvantaged areas) and 8 data are class 1 (not disadvantaged areas), with 23 data correctly classified in class 0 (disadvantaged areas) and 2 data incorrectly classified in class 0 (disadvantaged areas) and 8 data correctly predicted in class 1 (not disadvantaged areas). Based on the coufusion matrix table, the accuracy, sentivity, specificity, and AUC values can be calculated as follows:

$$\text{Accuracy} = \frac{8+23}{8+23+0+2} = 0.93$$

$$\text{Sensivity} = \frac{23}{23+0} = 1$$

$$\text{Specificity} = \frac{8}{8+2} = 0.8$$

$$\text{AUC} = \frac{1}{2}(1+0.8) \times 100\% = 0.9$$

Based on the calculation of the 70:30 RBF kernel classification performance, the resulting level of accuracy or similarity of SVM classification results with actual data is 93% with a sensivity number or the ability of the model to correctly identify positive class data is 100% and the specificity number or the ability of the model to correctly identify negative class data is 80%, and the AUC value is 0.9.

4. Kernel RBF 65:35

The results of the RBF Kernel Confusion Matrix are as follows.

Table 10. Confusion Matrix Kernel RBF 65:35

	0	**1**
0	27	2
1	0	10

Based on **Table 10**, it can be seen from the 39 training data, 29 data are classified as class 0 (underdeveloped regions) and 10 data are class 1 (non-developed regions), with 27 data correctly classified in class 0 (underdeveloped regions) and 2 data incorrectly classified in class 0 (underdeveloped regions) and 10 data correctly predicted in class 1 (non-developed regions). Based on the coufusion matrix table, the accuracy, sentivity, specificity, and AUC values can be calculated as follows:

$$\text{Accuracy} = \frac{10+27}{10+27+0+2} = 0.9487$$

$$\text{Sensivity} = \frac{27}{27+0} = 1$$

$$\text{Specificity} = \frac{10}{10+2} = 0.83$$

$$\text{AUC} = \frac{1}{2}(1+0.83) \times 100\% = 0.915$$

Based on the calculation of the 65:35 RBF kernel classification performance, the level of accuracy or similarity of the SVM classification results with the actual data is 94.87%. The sensivity rate or the ability of the model to correctly identify positive class data is 100% and the specificity rate or the ability of the model to correctly identify negative class data is 83%, and the AUC value is 0.915.

5. Kernel RBF 60:40

The results of the RBF Kernel Confusion Matrix are as follows.

Table 11. Confusion Matrix Kernel RBF 60:40

	0	**1**
0	30	2
1	0	12

Based on Table 11, it shows that out of 44 training data, 32 data are classified as class 0 (disadvantaged areas) and 12 data are class 1 (not disadvantaged areas), with 30 data correctly classified in class 0 (disadvantaged areas) and 2 data incorrectly classified in class 0 (disadvantaged areas) and 12 data correctly

predicted in class 1 (not disadvantaged areas). Furthermore, from the coufusion matrix table, the values of accuracy, sentivity, specificity, and AUC can be calculated.

$$\text{Accuracy} = \frac{12+30}{12+30+0+2} = 0.954$$
$$\text{Sensivity} = \frac{30}{30+0} = 1$$
$$\text{Specificity} = \frac{12}{12+2} = 0.857$$
$$\text{AUC} = \frac{1}{2}(1+0.857) \times 100\% = 0.9285$$

Based on the calculation of the RBF 60:40 kernel classification performance, the level of accuracy or similarity of the SVM classification results with the actual data is 95.4% with the sensivity number or the ability of the model to correctly identify positive class data is 100%, the specificity number or the ability of the model to correctly identify negative class data is 85.7%, and the AUC value is 0.9285.

From the evaluation of classification performance for each data division ratio to RBF kernel. RBF kernel 60:40 produces the highest accuracy rate and AUC value, namely 95.4% accuracy with an AUC value of 0.9285, which means it is classified as a very good classification result category. SVM with kernel RBF 60:40 parameter cost = 1, gamma = 0.1 is very good at classifying the most remote areas in Maluku Province. Next is to display the actual and predicted data from the implementation of the SVM kernel RBF classification model on the testing data can be seen in Table 12.

Table 12. Comparison of Actual Data and Classification Results

No	Data to -	Actual	SVM Classification Results	No	Data to -	Actual	SVM Classification Results
1	2	0	0	**23**	47	0	0
2	4	0	0	**24**	48	0	0
3	5	0	0	**25**	51	0	0
4	8	0	0	**26**	64	0	0
5	12	1	1	**27**	67	0	0
6	16	1	1	**28**	72	0	0
7	18	1	1	**29**	73	0	0
8	**21**	**1**	**0**	**30**	75	0	0
9	**22**	**1**	**0**	**31**	79	0	0
10	23	0	0	**32**	81	0	0
11	25	0	0	**33**	82	0	0
12	28	0	0	**34**	83	0	0
13	31	1	1	**35**	85	0	0
14	34	0	0	**36**	87	0	0
15	35	0	0	**37**	92	1	1
16	36	0	0	**38**	94	1	1
17	37	0	0	**39**	97	1	1
18	38	0	0	**40**	101	1	1
19	39	0	0	**41**	104	1	1
20	40	0	0	**42**	106	1	1
21	45	0	0	**43**	107	1	1
22	46	0	0	**44**	108	1	1

Based on Table 12, it shows that the results of the SVM model implementation using the 60:40 RBF kernel from 44 testing data with 32 class 0 data (underdeveloped regions) and 12 class 1 data (regions not underdeveloped), 41 data are correctly classified according to the actual data. While the two wrongly classified

observations are the 21st and 22nd data observations, namely the Central Maluku Regency in 2019-2020, which is actually class 1 (not underdeveloped regions) after SVM classification of Central Maluku Regency including class 0 (underdeveloped regions). Then the weight value of each research attribute (variable) is shown in **Table 13**.

Table 14. Weight of Research Attributes

Variable	Weighted Absolute Value
RLM	15.2229
AHH	12.75332
PPM	10.50248
PRRP	8.909589
PRTMTS	8.13954
PJPPJTA	7.28002
JD	6.947509
AMH	5.207653
PRTPL	4.292884
JSDSMP	3.267386
JPRS	1.030281
PRTPAB	0.3425791

Based on **Table 13**, it shows that what influences the classification accuracy level based on the weight value (w) of each attribute above is obtained that the average length of schooling gives the greatest value to the classification accuracy level of disadvantaged regions in Maluku Province. The second is life expectancy, so it can be said that the main factor that contributes most to whether regions are left behind or not is the average years of schooling.

4. CONCLUSION

The best classification results of underdeveloped regions in Maluku Province using Support Vector Machine (SVM) with a proportion of research data 66 observations of training data and 44 observations of testing data using the RBF kernel with parameter values gamma = 0.1 and cost = 1. The accuracy of SVM classification results is 95.4% with an AUC value (0.9285) classified as a "very good" classification result category so that it can be said that Support Vector Machine with RBF kernel is very good at classifying the most remote areas in Maluku Province. The two most dominant research attributes are average years of schooling and life expectancy.

REFERENCES

[1] Ichwan, M., Dewi, I. A., & S, Z. M. (2018). Klasifikasi Support Vector Machine (SVM) Untuk Menentukan Tingkat Kemanisan Mangga Berdasarkan Fitur Warna. MIND JOURNAL, 3(2), 16-24.

[2] Maulana, J. P., & Irhamah. (2018). Klasifikasi Kabupaten di Provinsi Jawa Timur Berdasarkan Indikator Daerah Tertinggal dengan Metode Support Vector Machine (SVM) dan Entropy Based Fuzzy Support Vector Machine (EFSVM). Jurnal INFERENSI, 1(1), 9-15.

[3] Azies, H. A., & Anuraga, G. (2021). Klasifikasi Daerah Tertinggal di Indonesia Menggunakan Alogaritma SVM dan k-NN. Jurnal ILMU DASAR, 22(1), 31-38.

[4] Buulolo, E. (2020). Data Mining Untuk Perguruan Tinggi. Sleman: CV Budi Utama.

[5] Suyanto, D. (2017). Data Mining Untuk Klasifikasi dan Klasterisasi Data Edisi Revisi. Bandung: Informatika .

[6] Athifaturrofifah, Goejantoro, R., & Yuniarti, D. (2019). Perbandingan Pengelompokkan K-Means dan K-Medoids Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Pesebaran Titik Panas. Jurnal EKSPONENSIAL, 10(2), 143-151.

[7] Yahya, S. A. (2018). Klasifikasi Ketetapan Lama Studi Mahasiswa Menggunakan Metode Support Vector Machine dan Random Forest. Universitas Islam Indonesia, Yogyakarta.
[8] Fitriani, R. D., Yasin, H., & Tarno. (2021). Penanganan Klasifikasi Kelas Data Tidak Seimbang dengan Random Oversampling Pada Naive Bayes. Jurnal Gaussan, 10(1), 11-20.
[9] Leleury, Z. A., & Tomasouw, B. P. (2015). Diagnosa Penyakit Saluran Pernapasan dengan Menggunanakan Support Vector Machine (SVM). Jurnal Barekeng, 9(2), 109-119.
[10] Enjeligue, L. (2023). Penerapan Metode Twin Support Vector Machine Unuk Deteksi Dini Penyakit Stroke (Studi Kasus : RSUD Dr. H. Ishak Umarella Maluku Tengah RS Sumber Hidup-GPM). 8-14.
[11] Susilowati, E., Sabariah, M. K., & Gozali, A. A. (2015). Implementasi Metode Support Vector Machine Untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter. e-proceeding of Engineering, 2, hal. 1478-1484.
[12] Raehanun, M. (2021). Analisis Support Vector Machine (SVM) Dalam Prediksi Permintaan Emas Perhiasan. Universitas Islam Indonesia, Yogyakarta.
[13] Wisudawati, D. T. (2020). Analisis Sentimen Terhadap Dampak Covid-19 Pada Peforma E-Commerce di Indonesia Menggunakan Support Vector Machine (Review Aplikasi Tokopedia Pada Google Play). Universitas Muhammadiyah Semarang, Semarang.
[14] Sari, E. A., Saragih, M. T., Shariati, I. A., Sofyan, S., Baihaqi, R. A., & Nooraeni, R. (2020). Klasifikasi Kabupaten Tertinggal Di Kawasan Timur Indonesia dengan Support Vector Machine. Jurnal Informatika dan Komputer, 3(3), 188-195.
[15] Maulana, J. P., & Irhamah. (2018). Klasifikasi Kabupaten di Provinsi Jawa Timur Berdasarkan Indikator Daerah Tertinggal dengan Metode Support Vector Machine (SVM) dan Entropy Based Fuzzy Support Vector Machine (EFSVM). Jurnal INFERENSI, 1(1), 9-15.