Comparisons of Logistic Regression and Support Vector Machines in Classification of Echocardiogram Dataset

Neni Alya Firdausanti[1*], Ratih Ardiati Ningrum[2], and Siti Qomariyah[3]
[1]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Data Science Technology, Airlangga University, Surabaya, Indonesia
[3]Institut Agama Islam Negeri Kudus, Kudus, Indonesia
*Corresponding author: neni.alya@its.ac.id

Received: 16 August 2022
Revised: 19 September 2022
Accepted: 22 September 2020

ABSTRAK – **Echocardiography is a test that uses sound waves to produce an image of our heart. This image is called an echocardiogram. This paper uses Echocardiogram Dataset, in which the problem is to classify from 7 features whether the patient will survive or not. In this study, the classification method is used to solve this problem. Some classification methods can be applied to classify category response variables, such as Logistic regression and Support Vector Machines (SVM). The method for predicting best accuracy used holdout and cross-validation. Before doing classification, some preprocessing procedures were applied to this dataset. The preprocessing procedures include missing value imputation using median imputation, outliers' detection in univariate and multivariate procedures, and feature selection using the backward method. The result of classification in the analysis showed that SVM with unstratified holdout gave the best accuracy, that is 91.54%.**

Keywords– **Classification, Cross Validation, Echocardiogram, Holdout, Logistic Regression, SVM**

I. INTRODUCTION

Many machine learning techniques have been developed in the past few decades [1] [2] [3]. This machine learning came with various kinds of algorithms that are suitable for so many kinds of big datasets [4]. Machine learning is used to capture data patterns, assess links between data, validate conclusions by using patterns found, and predict new findings on new datasets consistently or methodically [5]. One of the key techniques for supervising learning in machine learning is classification. A multivariate approach called classification works with grouping together various sets of objects (or observations) and assigning newly discovered objects (observations) to pre-determined categories [6]. The classification method has been used in many research areas, including health care [7], economics [8], meteorology [9], astronomy [10], biology [11], and many more, in which the dependent variable is categorical. Several problems that have categorical as the dependent variable are the response to the treatment of a patient, the presence or absence of myocardial infarction, the categories of patients presenting a particular symptom, and many more. Hence, these problems should be solved by appropriate methods for categorical classification problems.

The main theoretical advances are about classification problems that have grown algorithms such as Logistic Regression and Support Vector Machines (SVMs). Support Vector Machines (SVMs) are a method for classification problems of both linear and nonlinear data [12]. This method uses a nonlinear mapping to transform the original training data into a higher dimension. SVMs have been chosen for classification problems because they classify in highly accurate values, are much less prone to overfitting, provide a compact description of the learned model, and can be used for numeric prediction. Research in classification problems that used SVMs and logistic regression has been done by Rahman et al. 2015, who compared Adaboost, KNN, SVM, and logistic regression classifiers for the survival of cardiac surgery patients data. This research's result stated that SVM-RBF gave high accuracy for the random oversampling dataset [13]. Although SVMs are a suitable classification method, they have poor generalization ability when the training dataset is small or contains noisy data [14] and lack interpretability. On the other hand, one statistical technique for classification is logistic regression. This technique builds a linear model based on a transformed target variable [15]. By using logistic distribution, logistic regression is flexible and easily used function, and it lends itself to a clinically meaningful interpretation [16]. Accuracy and interpretability are essential in classification, especially in health care. Therefore in this study, we compare the performance of SVM and Logistic Regression.

Various evaluation method for the predictive accuracy of a classifier, such as a holdout, cross-validation, and bootstrap, has usually been used to assess accuracy [17]. The result of the classification model can usually be used for prediction. Therefore, the accuracy of prediction by the classifiers must be chosen as well as the classification method. So, this paper not only compares logistic regression and the SVM method but also compares for training and testing data split method.

Myocardial infarction is one example of a dependent variable in categorical. As mentioned above, logistic regression is a method for classification that can handle categorical classification problems and the SVMs method. Therefore, this paper uses both methods to classify the myocardial infarction problem. The goal of simultaneously using logistic regression and SVMs method is to know which method gives the best accuracy in classifying myocardial infarction

problems. Hence, the training and testing data split used the holdout and cross-validation method. Although the primary goal of this research is on classification, the procedure of examining data or preprocessing data at the beginning of analysis should be done.

II. LITERATURE REVIEW

A. Logistic Regression

Logistic regression is the most important model for categorical response data [18]. The goal of an analysis using this method is to find the best fitting and most parsimonious yet biologically reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables [19]. A logistic regression model is distinguished from linear regression because the outcome variable in logistic regression is binary or dichotomous.

For a binary response variable Y and an explanatory variable X, let

$$\pi(x) = P(Y=1 \mid X=x) = 1 - P(Y=0 \mid X=x) \quad (1)$$

The logistic regression model is

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \quad (2)$$

Equivalently, the log odds, called the logit, has the linear relationship

$$\text{logit}[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha+\beta x \quad (3)$$

The importance of this transformation is that logit [π(x)] has many of the desirable properties of a linear regression model. It is linear in its parameters, may be continuous, and may range from -∞ to +∞, depending on the range of x.

B. Support Vector Machine

Support Vector Machine (SVM) is a learning system that uses a hypothetical linear function in a high-dimensional. SVM is trained with algorithms based on optimization theory by applying learning bias derived from statistical theory [20]. The primary purpose of this method is to build OSH (Optimal Separating Hyperplane), which makes an optimum separation function that can be used for classification.

Linearly separable data is data that can be separated linearly. Suppose that $\mathbf{x}_i = \{\mathbf{x}_i,\ldots,\mathbf{x}_n\}, \mathbf{x}_i \in \Re^n$ is the data set and $y_i \in \{+1,-1\}$ is the class label of the x_i data. The best dividing fields is the one that not only separate the data but also have the largest margins. Data which is located in the boundary field is called the support vector. In Figure 1, two classes can be separated by a pair of parallel bounding plane. The first delimiter field limits the first class while the second bounding field limits the second class,

$$\begin{aligned}\mathbf{x_i w}+b &\geq +1, y_i = +1\\ \mathbf{x_i w}+b &\leq -1, y_i = -1\end{aligned} \quad (4)$$

W is the normal field and b is the position of the alternate field to the coordinate center. The margin (distance) value between the bounding plane (based on the formula of spacing to the center) the data is $(1-b-(1-b))/\|b\|$. The value of this margin is maximized by satisfying equation (4). By multiplying b and w by a constant, a margin value multiplied by the same constellation will be generated. Therefore, the constraint in equation (2.4) is a scaling constraint that can be satisfied by rescaling b and w. Maximize $1/\|w\|$ equal to minimize $\|w\|^2$ so that the two bounding plot of equation (4) are represented in the inequality.

$$y_i(\mathbf{x_i w}+b)-1 \geq 0 \quad (5)$$

So, the search for the best dividing field with the largest margin value can be formulated into a constraint optimization problem, that is

$$\begin{aligned}&\min \frac{1}{2}\|\mathbf{w}\|^2\\ &y_i(\mathbf{x_i w}+b)-1 \geq 0\end{aligned} \quad (6)$$

This problem (6) will be more easily resolved if it is converted into a Lagrange formula using a Lagrange multiplier. Thus, the constraint optimization problem can be changed to:

$$\min_{w,b} L_p(\mathbf{w},b,\alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i y_i(x_i\mathbf{w}+b) + \sum_{i=1}^{n}\alpha_i \quad (7)$$

By adding constraint, $\alpha_i \geq 0$ (value of the lagrange coefficient). Minimizing Lp to w and b, then from $\frac{\partial L_p(\mathbf{w},b,\alpha)}{\partial \mathbf{w}} = 0$ and $\frac{\partial L_p(\mathbf{w},b,\alpha)}{\partial b} = 0$, the equation (7) is obtained.

$$w = \sum_{i=}^{n}\alpha_i y_i \mathbf{x}_i \text{ and } \sum_{i=1}^{n}\alpha_i y_i = 0 \quad (8)$$

The value of vector w is often a great value (infinity), but the value of αi is finite. Therefore, LP lagrangian formula (primal problem) is converted into LD (dual problem). Substituting equation (8) to equation (7), so that LD become

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j \quad (9)$$

So the question of finding the best dividing field can be formulated in the following equation:

$$\max_{\alpha} L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j$$

$$\text{where, } \sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0 \quad (10)$$

Thus, can be obtained from the value αi which will be used to find w. there is αi value for each training data, training data having αi > 0 is support vector while the rest have αi = 0. Thus the decision function generated is only influenced by the support vector. The best dividing search formula is quadratic programming problem so that the global maximum value αi can always be found. After the quadratic programming problem solution is found (αi value), then the class of an x testing can be determined by the following equation.

$$f(x_d) = \sum_{i=1}^{ns} \alpha_i y_i \mathbf{x_i} \mathbf{x_d} + b \quad (11)$$

Where x_i is support vector, ns is the number of support vector, and x_d is the data to be classified. While the value of b is obtained by equation (12)

$$b = \frac{1}{2} \mathbf{w} \left[\mathbf{x_r} + \mathbf{x_s} \right] \quad (12)$$

$\mathbf{X}_r$ and $\mathbf{X}_s$ are support vector for each class on conditional of equation $\alpha_r, \alpha_s > y_r = 1, y_s = -1$ [16].

C. Training and Testing

For a classification problem, it is natural to measure a classifier's performance in terms of the error rate. The classifier predicts the class of each instance. If it is correct, it is counted as a success. On the opposite, it is counted as an error. To predict the performance of a classifier on new data, it needs to assess its error rate on a dataset that played no part in the formation of the classifier. This independent dataset is called the test set. Usually, the datasets are divided into the training data, the validation data, and the test data.

The training data is used by one or more learning schemes to come up with classifiers, and the test data is used to calculate the error rate of the final, optimized method. Generally, the larger the training sample, the better the classifier. Then, the larger the test sample, the more accurate the error estimate. Several procedures to split the training and testing data include holdout, cross-validation, leave one out cross-validation, and the bootstrap [22].

III. METHODOLOGY

The classification experiment was conducted on an Echocardiogram dataset. This data is taken from the website UCI Machine Learning Repository. In this experiment, the dependent variable is still alive. This variable consisted of two categories. The categories are 0 for dead at the end of the survival period and 1 for still alive. The total numbers of independent (feature) used in this experiment are eight variables shown in Table 1 below. In this study, there are 132 instances used in this experiment.

Table 1 Independent Variable

No	Variable	Description
1	Survival	The number of months patient survived.
2	Age at heart attack	Age in years when heart attack occurred.
3	Fractional shortening	A measure of contracility around the heart lower numbers are increasingly abnormal.
4	Epss	E-point septal separation. This is another measure of contractility. Larger numbers are increasingly abnormal.
5	Lvdd	Left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
6	Wall motion score	A measure of how the segments of the left ventricle are moving.
7	Wall motion index	A measure of how the segments of the left ventricle are moving divided by number of segments seen.

Before processing the data using classification methods using logistic regression and SVMs to get the best classification result, the first step is doing preprocessing. Preprocessing on this data set includes imputation on missing values, outliers detection and features selection.

IV. RESULTS AND DISCUSSIONS

The first part of the analysis is to identify the echocardiogram data set based on descriptive statistics simultaneously. The missing values, mean, standard deviation, minimum and maximum for each feature are shown in the table below.

Table 2 Descriptive Statistics Echocardiogram Dataset

Variable	N*	Mean	StDev	Min	Max
Survival	2	22.18	15.86	0.03	23.50
Age at heart attack	6	62.81	8.34	35.00	62.00
Fractional shortening	8	0.22	0.11	0.01	0.20
Epss	15	12.17	7.37	0.00	11.00
Lvdd	11	4.76	0.81	2.32	4.65
Wall motion score	4	14.44	5.02	2.00	14.00
Wall motion index	2	1.38	0.45	1.00	1.22

N* = the number of missing values

According to Table 2, the number of missing values is 2 for the variable Survival, and its value for the mean is 22.18, which means that, on average, patients have survived for 22.18 months. The standard deviation is 15.86, which means that the data is heterogeneous. This heterogeneity can be seen through the difference between the minimum and maximum values. Then, the minimum and maximum values indicate that the shortest time for the patient to survive is 0.03 months, and the longest is 23.50 months. Furthermore, the other descriptive statistics for each variable are shown in Table 2.

As mentioned in the methodology, preprocessing was done to examine the dataset. This procedure consisted of the imputation of missing values, outliers' detection, and feature selection. Median imputation was done for each category response variable which is shown in the table below.

Table 3 Median Imputation

Variable	Median Imputation			
	N*	Still alive = 0	N*	Still alive = 1
Survival	0	29	2	1
Age at heart attack	3	61	3	65
Fractional shortening	3	0.24	5	0.17
Epss	8	9.35	7	14.8
Lvdd	3	4.49	8	5.1
Wall motion score	1	13.5	3	15.5
Wall motion index	0	1.125	2	1.45

N*= the number of missing values

The imputation was done one by one for each category response variable. This procedure should be done so that the whole dataset can be analyzed.

The following procedure in examining data is outliers detection, done by univariate and multivariate methods. Univariate outliers detection using boxplot showed that almost all variables have outliers except the survival variable. Furthermore, the result in multivariate outliers detection using Mahalanobis distance stated that data number 36, 59, and 75 was an outlier for category response still alive = 0, and data number 33, 97, and 109 was an outlier for category response still alive = 1. These outliers were not deleted from the dataset.

The last procedure in preprocessing is feature selection. The backward selection was used to select the forming number of features, as shown in Table 1. Making backward selection showed that the variable LVDD should be deleted from the dataset because this variable is not significant. Then the subsequent analysis, classification, was done using six variables.

Table 4 Classification Accuracy Using Logistic Regression

No	Stratified Holdout	Stratified CV	Unstratified Holdout	Untratified CV
1	66.67	71.43	80	85.71
2	66.67	85.71	100	92.86
3	33.33	69.23	80	100
4	66.67	100	80	53.85
5	66.67	100	100	69.23
Average	60.00	85.27	88.00	80.33
StDev	14.91	14.86	10.95	18.69

Logistic regression classification is done using different testing and training data methods (i.e., hold out and cross-validation) for the stratified and unstratified datasets. The procedure for the holdout method was done by splitting the dataset into 80% for training and the rest for testing. This procedure was done five times to get the best accuracy for the stratified and unstratified holdout. Then for cross-validation method was done by using different values of the fold (fold = 1, 2, 3, 4, 5). This procedure was done for stratified and unstratified cross-validation. The results are shown in the table 4.

In this paper, the logistic regression model was not shown because this research aimed to compare holdout and cross-validation methods for stratified and unstratified datasets. Logistic regression classification is done using different testing and training data methods (i.e., hold out and cross-validation) for the stratified and unstratified datasets. In Table 4, the unstratified holdout method gave the best accuracy (88%), and its standard deviation is the most minor (10.95). The procedure for the holdout method was done by splitting the dataset into 80% for training and the rest for testing. This procedure was done five times to get the best accuracy for the stratified and unstratified holdout. Then for cross-validation method was done by using different values of the fold (fold = 1, 2, 3, 4, 5). This procedure was done for stratified and unstratified cross-validation. The results are shown in the table below.

In classification using SVM, first, we experiment using different parameter costs to get the best cost. The best cost for this experiment is. Besides using different costs, we also try different kernel types. The result is that the most suitable kernel type for this data is linear. We use a different method to partition the training and testing data to get the best model. Classification results using SVM are shown in Table 5.

Table 5 Classification Accuracy Using SVM

No	Stratified Holdout	Stratified CV	Unstratified Holdout	Untratified CV
1	80	88.89	92.31	88.89
2	80	88.89	92.31	88.89
3	40	84.61	96.15	76.92
4	80	76.92	80.77	88.46
5	80	84.61	96.15	88.61
Average	72	84.78	**91.54**	86.35
StDev	17.89	4.89	**6.32**	5.28

Based on the Table, we can see that the highest standard deviation is using stratified holdout. It means that if we run a classification using this method, the accuracy result between the first, second, and fifth runs is much different. On the other hand, the other method has only an insignificant difference in their standard deviation. It can be concluded that the best method to get the highest accuracy on SVM is unstratified holdout, with an accuracy of 91.54% and a standard deviation of 6.32.

V. CONCLUSIONS AND SUGGESTIONS

The main goal of this research is to compare the classification method between logistic regression and Support Vector Machines (SVMs) to the Echocardiogram dataset. Both methods are used for classifying Echocardiograms whose dependent variable is in categorical type. The results showed that the best method for classifying the Echocardiogram dataset is SVM with an unstratified holdout with an accuracy of 91.54%.

In this research, we used accuracy to evaluate the performance of the classification. Evaluating the classification performance based on various metrics will be a future study to get a more profound understanding of the classification algorithms.

REFERENCES

[1] F. Gorunescu, Data Mining Concepts, Models, and Techniques, Berlin Heidelberg: Springer-Verlag, 2011.

[2] D. Dall, R. Kaur and M. Juneja, "Machine Learning: A Review of the Algorithms and Its Application," 2020.

[3] L. Zhang, J. Wen, Y. Li, J. Chen, Y. Ye, Y. Fu and W. Livingood, "A review of machine learning in building load prediction," *Applied Energy,* vol. 285, 2021.

[4] I. Sarkeh, "Machine Learning: Algorithms, Real-World Application and Research Direction," *SN Computer Science,* vol. 2, no. 3, 2021.

[5] M. Tariq, S. Tayyaba, M. Ashraf and V. Balas, "Deep learning techniques for optimizing medical big data," *Deep Learning Techniques for Biomedical and Health Informatics,* pp. 187-211, 2020.

[6] F. Herrera, F. Charte, A. Rivera and M. del Jesus, Multilable Classification, Cham: Springer International Publishing, 2915.

[7] E. Garba and A. Amadu, "A Systematic Review of Data Mining in Health Care: A Case of Breast Cancer," *International Journal of Research anf Analysis in Science and Engineering,* vol. 2, no. 1, pp. 19-25, 2022.

[8] A. Pena, D. Cisgar and D. Unal, "Comparison of Data Mining Classification Algorithms Determining the Default Risk," *Scientific Programming* , vol. 2019, 2019.

[9] A. Niazalizadeh Moghadam and R. Ravanmehr, "Multi-agent distributed data mining approach for classifying meteorology data: case study on Iran's synoptic weather stations," *International Journal of Environmental Science and Technology,* vol. 15, no. 1, pp. 149-158, 2018.

[10] P. Yang, G. Yang , F. Zhang, B. Jiang and M. Wang, "Spectral Classification and Particular Spectra Identification Based on Data Mining," *Archives of Computational Methods in Engineering,* vol. 28, no. 3, pp. 917-935, 2021.

[11] D. Degadwala and D. Vyas, "Data Mining Approach for Amino Acid Sequence Classification," *International Journal of New Practices in Management and Engineering,* vol. 10, no. 04, pp. 01-08, 2021.

[12] S. Huang, N. Cai, P. Pacheco, S. Narrandes, Y. Wang and W. Xu, "Application of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics & Proteomics January,* vol. 15, no. 1, pp. 41-51, 2018.

[13] H. Rahman, Y. Wah, H. He and A. Bulgiba, "Comparison of ADABOOST, KNN, SVM, and Logistic Regression in Classification of Imbalanced Dataset," in *International Conference on Soft Computing in Data Science*, Singapore, 2015.

[14] X. Shen, L. Niu, Z. Qi and Y. Tian, "Support vector machine classifier with truncated pinball loss," *Pattern Recognition,* vol. 68, pp. 199-210, 2017.

[15] H. Best and C. Wolf, Logistic Regression, Los Angles: Sage, 2015.

[16] E. Choi, M. Bahadori, J. Kulas , A. Schuetz and W. Stewart, "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism," in *Advances in Neural Information Processing Systems*, 2016.

[17] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, USA: Morgan Kaufmann, 2012.

[18] A. Agresti, Categorical Data Analysis, Second Edition, New Jersey: Johm Wiley & Sons, 2002.

[19] D. Hosmer and S. Lemeshow, Applied Logistic Regression Second Edition, USA: John Wiley & Sons, 2000.

[20] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machine, Cambridge: Cambridge University Press, 2000.

[21] K. Sembiring, "Penerapan Teknik Support Vector Machine untuk Pendeteksian Intrusi pada Jaringan," Bandung, 2007.

[22] I. Witten , E. Frank and M. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, USA: Morkan Kaufmann Publisher, 2011.