JAMBURA JOURNAL OF PROBABILITY AND STATISTICS
Volume 4 Nomor 1, May 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jps
DOI: https://doi.org/10.34312/jjps.v4i1.19246

COMPARING LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINE IN BREAST CANCER PROBLEM

Caecilia Bintang Girik Allo[1*], Leonardus Sandy Ade Putra[2], Nicea Roona Paranoan[3], Vincentius Abdi Gunawan[4]
[1] Program Studi Statistika, FMIPA, Universitas Cenderawasih
[2] Teknik Elektro, Fakultas Teknik, Universitas Tanjungpura
[3] Program Studi Statistika, FMIPA, Universitas Cenderawasih
[4] Teknik Informatika, Fakultas Teknik, Universitas Palangka Raya

***e-mail**: caecilia.bintang@fmipa.uncen.ac.id

Abstract
There are several methods used for the classification problems. There are many different kinds of fields that can be used. Nowadays, Support Vector Machine (SVM) is a popular classification method that has been proposed by many researchers. Using the same method but different distribution methods for creating training and testing data in the same dataset can yield varying results in terms of prediction accuracy, which is crucial in classification. In this paper, we compare the prediction accuracy between SVM results and Logistic Regression results to determine the better method to classify the current condition of the patient after undergoing some treatment. Several treatments are used in this paper, including feature selection, feature extraction, separating the train and testing data using Holdout and K-Fold CV. Stepwise selection is done to reduce the features. Training and testing dataset is obtained using the five stratified and non-stratified holdout and five fold stratified and non-stratified cross validation. The result shows that the best method to classify the cancer dataset is five fold stratified cross validation SVM with radial kernel. The obtained accuracy is 81,816% with variance as much as 0,94%.
Keywords: *Support Vector Machine, Logistic Regression, Accuracy, Breast Cancer*

1. INTRODUCTION

Cancer is the main cause of death in the world. According to the World Health Organization (WHO), there were 9,6 million deaths caused by cancer in 2018. Breast cancer is one of the most common causes of cancer-related deaths. The World Health Organization (WHO) also reported that there were 2,3 million women diagnosed with breast cancer and 685000 deaths in 2020. There are several ways to treat breast cancer. They are surgery, drugs, and chemotherapy. After that treatment, patients of breast cancer hope that they can recover from breast cancer.

Classification is a common problems that is often encountered. The purpose of classification is to predict class (categorical) based on predictor variables. Sultana and Jilani (2018) used several classification methods, such as Logistic Regression, Decision Tree, and Multi Layer Perceptron, to predict breast cancer. Support Vector Machine and Logistic Regression were used for calibrating cellular automata land (Mustafa et al., 2018). Han et al. (2019) measured the performance of Logistic Regression and Support Vector Machine for Seismic Vulnerability Assessment and Mapping. Handayani (2021) compared Support Vector Machine (SVM), Logistic Regression and Artificial Neural Network (ANN) to predict cardiac disease. She also used several combinations to split the train and testing data. Nurlaily et al. (2022) compared Support Vector Machine (SVM) and Logistic Regression to classify hepatitis patients. The result showed that holdout stratified SVM using kernel radians is the best model. Khandezamin, Naderan, and Rasthi (2020) used logistic regression to eliminate the less important feature and then using Group Method Data Handling (GMDH)

Received: 2023-03-26 | Accepted: 2023-05-28 | Available Online: 2023-05-31

neural network to diagnose benign and malignant breast cancer. Allo et al. (2023) used several methods, such as Logistic Regression, Naïve Bayes, Support Vector Machine and Random Forest, to classify climate models. Nadh & Saraswathi (2023) compares Logistic Regression and Support Vector Machine to improve the accuracy in stroke prediction.

The purpose of this paper is to compare classification methods using Support Vector Machine (SVM) and Logistic Regression for the condition of breast cancer classification. This paper uses preprocessing data, data cleaning and data transformation, to fill in missing value and high range between variables. Data training and data testing are chosen using Five Stratified and Non-stratified Holdout and Five Stratified and Non-stratified Folds Cross Validation. The best model is chosen based on mean accuracy and variance of accuracy.

2. RESEACRH METHOD

This paper uses Wisconsin Prognostic Breast Cancer taken from UCI Machine Learning Repository. The researcher uses the R programming language. The data contains 33 predictor variables and one response variable. There are 198 observed patients with breast cancer. The response variable indicates whether a patient is recurrent or nonrecurrent based on the 33 predictor variables after receiving breast cancer treatment. Due to the large number of predictor variables, feature selection and feature extraction techniques are used to manage the dataset. Data preprocessing is applied to the dataset as shown in Figure 1. After obtaining the new data, the steps are illustrated in Figure 2.

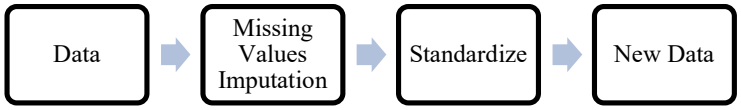

Figure 1. Systematic of Data Preprocessing

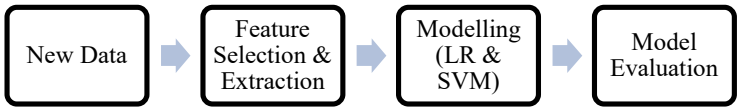

Figure 1. Step After Preprocessing

2.1 Data Cleaning

One of the utilities of data cleaning is to fill in missing value. Missing value can be caused by respondents in a survey, researcher in a survey, or data that is incomplete observed. There are several techniques to fill in missing values. Respondents may ignore answers to answer some questions such as salary and age. Animals or plants may die before all variables have been measured. Mean, median, modus, or regression can be applied to fill in missing values. If we use regression, first we should find correlation between variables that contain missing value and other variables and then we build the model regression between variables that contain missing value and variables that have high correlation. Then we can use the model to find the value of missing value. If the data has normal distribution then mean can be used to fill in the missing values. Median is better than mean if the data distribution for a given

class is skewed (Han, Kamber and Pei, 2012).

2.2 Data Transformation

Data transformation is a method to transform the data into forms appropriate for mining. Smoothing, attribute construction, aggregation, normalization, and discretization are strategies for data transformation. Data transformation by normalization is normalizing the data attempts to give all variables an equal weight. There are several methods for data normalization such as min-max, *normalization*, *z-score normalization*, and *normalization by decimal scaling*. In *z-score normalization*, a value v_i of A is normalized to v_i' by computing

$$v_i' = \frac{v_i - \overline{\mathrm{A}}}{\sigma_{\mathrm{A}}} \tag{1}$$

where $\overline{\mathrm{A}}$ is the mean of A and σ_{A} is the standard deviation of A (Han, Kamber and Pei, 2012).

2.3 Feature Selection and Feature Extraction

Feature selection and feature extraction actually use high dimension data. The purpose of feature selection is to find a subset of features that produce a better model. Benefit of feature selection is improving accuracy and reducing training time. There are three feature selection methods. They are filter method, wrapper method, and embedded method. Filter method is to find high correlation between variable predictors and ignore them. So, the final model is variable response and variables predictor that have low correlation between variables predictor. Backward selection, forward selection, and stepwise selection are ways to do feature selection in the wrapper method. Backward selection starts with all variables and drops at each step. Forward selection starts with empty variables and adds variables at each step. Stepwise starts with empty variables and adds or drops variables at each step. The purpose of feature extraction is to find new features that are linear combinations between variable predictors. Principal component analysis is a way to do feature extraction. Select the new variables based on the eigen vector.

2.4 Model Evaluation and Selection

Model evaluation is done after the classification model has been built. The main question is how well the model can classify the problem (Han, Kamber and Pei, 2012). If we obtain some classifier from several different methods, then which one has accurately predicted the problems. Therefore, this evaluation is done by dividing the data as training set and testing set. The training set will build the model which will be tested by the testing set. Holdout and cross validation are the most common methods for dividing the set. The result will be used as the evaluation measures, such as accuracy, sensitivity, and specificity. Confusion matrices can make the calculation for the evaluation measures easier, as shown in Figure 3.

TP (true positive) refers to the observation that belongs to the ‘yes’ group and is classified as the ‘yes’ group as well. FP (false positive) refers to the observation that belongs to the ‘no’ group but is classified as the ‘yes’ group. FN (false negative) refers to the observation that belongs to the ‘yes’ group but is classified as the ‘no’ group. TN (true negative) refers to the observation that belongs to the ‘no’ group and is classified as the ‘*no*’ group as well. Formula to calculate the evaluation measures is shown in equation (2).

$$accuracy = \frac{TP + TN}{P + N} \tag{2}$$

		Prediction		Total
		yes	no	
Actual Class	yes	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	total	*P'*	*N'*	*P+N*

Figure 3. Confusion Matrix

2.5 Logistic Regression

Regression is one of the methods used for describing the relation between response variable and predictor variable. Logistic regression is used when the response variable is categorical. So that is why logistic regression can be used as classification methods. There are three kinds of logistic regression based on the response variable (Agresti, 2007). Those are binary logistic regression for 2 category nominal response variable, multinomial logistic regression for more than 2 category nominal response variable and ordinal logistic regression for 2 or more category ordinal response variable. The logistic regression model is

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{3}$$

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \tag{4}$$

2.6 Support Vector Machine (SVM)

Support vector machine is used as a technique for classification, both linear and nonlinear data. Using a nonlinear mapping, the full data set is transformed to a higher dimensional space. Support Vector Machine (SVM) select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible (Han, Kamber and Pei, 2012). The SVM finds an optimal linear classifier or hyper plane in that higher dimension. The best hyper plane has the maximum marginal called maximum marginal hyperplane (*MMH*). *MMH* that optimally maximum the margin satisfied the condition:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \sum_{i=1}^{l} \xi \\ subject\ to \quad & y_i\left(w\phi\left(x_i\right)+b\right)+\xi_i \geq 1 \\ & \xi_i > 0 \ ; i = 1,\ldots,l \end{aligned} \tag{5}$$

If the data can be separate linearly, then the *MMH* can be rewritten as decision boundary

$$d\left(\boldsymbol{X}^T\right) = \sum_{i=1}^{l} y_i \alpha_i \boldsymbol{X}_i \boldsymbol{X}_i^T + b_0 \tag{6}$$

Where y_i is the class label of the $\boldsymbol{X}$, $\boldsymbol{X}_i$ is the training set, α_i, b_0 are the numeric parameters that can be found by optimization by Lagrangian. If the data can be separate nonlinearly, we can use a kernel function to get MMH, such as

1. Polynomial Kernel

$$K\left(x_i, x_j\right) = \left(\gamma x_i^T x_j + r\right)^d, \gamma > 0 \tag{7}$$

2. Radial Basis Function kernel

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \gamma > 0 \tag{8}$$

3. Sigmoid Kernel

$$K\left(x_i, x_j\right) = \tanh\left(\gamma x_i^T x_j + r\right)^d \tag{9}$$

where γ, r, and d are kernel parameters.

3. RESULT AND DISCUSSION

3.1 Best Feature Selection and Feature Extraction

In this part, researchers conduct feature selection and feature extraction using SVM and logistic regression with five stratified and non-stratified holdouts. We also do feature selection and feature extraction with five fold stratified and non-stratified cross validation using SVM and logistic regression. The dataset contains missing values, specifically in the lymph node status variable, where there are four missing values. Prior to feature selection and extraction, the researchers impute the missing values with the median, as the variable is an integer. After missing value imputation, the researchers standardize the data using z-score normalization to ensure equal weight among variables. The process of missing value imputation and standardization is illustrated in Figure 1. The researchers select the best method based on the mean and variance of accuracy. The dataset contains 33 predictor variables, including time, mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry,

mean fractal dimension, standard error of radius, standard error of texture, standard error of perimeter, standard error of area, standard error of smoothness, standard error of compactness, standard error of concavity, standard error of concave points, standard error of symmetry, standard error of fractal dimension, worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst compactness, worst concavity, worst concave points, worst symmetry, worst fractal dimension, tumor size, and lymph node status. The result can be seen in Table 1 where the best feature selection and feature extraction method is stepwise, which selects 12 predictor variables. Stepwise has a mean accuracy of 81.231% with a variance of accuracy of 0.379%. The boundary of this result is to choose the best method for feature selection and feature method.

Table 1. The Six Highest Mean and Its Variance Accuracy based on Feature Selection and Extraction

Feature Selection & Extraction	Mean Accuracy (%)	Variance of Accuracy (%)
General Model	77,231	2,272
Filter Method	78,289	3,713
Backward	80	1,893
Forward	81,538	0,947
Stepwise	81,231	0,379
PCA	77,231	0,379

3.2 Best Estimated Model

Using 12 variables selected from stepwise, we analyze the case with five stratified and non-stratified holdouts using SVM and logistic regression. We also analyze the case with five fold stratified and non-stratified cross validation using SVM and logistic regression. The selected variables are time, lymph node status, worst radius, mean texture, standard error of fractal dimension, mean smoothness, worst area, mean radius, mean area, standard error of concave points, mean fractal dimension, and standard error of texture. For the SVM, we use four kinds of kernel. They are linear kernel, polynomial kernel, radial kernel, and sigmoid kernel. We compare mean accuracy and variance of accuracy to get the best method. There are 20 mean accuracy and variance of accuracy that are compared. We choose the best method based on the lowest variance of accuracy. The result can be seen in Table 2. The best method for this case is five fold stratified cross validation using SVMSVM with radial kernel. This method has a mean accuracy of 81,816% and variance of accuracy 0,94%.

Table 2. Result Best Mean and Variance Accuracy Based On Classification Method

Method	Mean Accuracy (%)	Variance of Accuracy (%)
Holdout Stratified SVM Linear	75,38462	11,3609467
Holdout Stratified SVM Polynomial	75,07692	15,5266272
Holdout Stratified SVM Radial	80	5,6804734
Holdout Stratified SVM Sigmoid	73,84615	8,5207101
Holdout Stratified Logistic Regression	77,53846	9,0887574
Holdout Non-Stratified SVM Linear	78,18182	29,0174472
Holdout Non-Stratified SVM Polynomial	75,75758	28,466483
Holdout Non-Stratified SVM Radial	79,39394	27,1808999
Holdout Non-Stratified SVM Sigmoid	68,48485	29,7520661
Holdout Non-Stratified Logistic Regression	78,78788	22,9568411
5-Fold Stratified SVM Linear	78,81488	7,9335466
5-Fold Stratified SVM Polynomial	76,73859	28,2306288
5-Fold Stratified SVM Radial	81,81614	0,9395741
5-Fold Stratified SVM Sigmoid	73,27455	9,2003343
5-Fold Stratified Logistic Regression	82,89243	24,1896417
5-Fold Non-Stratified SVM Linear	76,73077	67,9881657
5-Fold Non-Stratified SVM Polynomial	76,76923	64,5177515
5-Fold Non-Stratified SVM Radial	81,32051	21,7087442
5-Fold Non-Stratified SVM Sigmoid	72,24359	16,4464168
5-Fold Non-Stratified Logistic Regression	80,29487	55,8458251

4. CONCLUSION

The researcher compared classification methods for classifying the condition of breast cancer, which can be either recurrent or nonrecurrent. As the predictor variable containing missing values was an integer variable, the median method was used for imputing missing values. Z-score normalization was used to ensure that all predictor variables had the same weight. The accuracy of five stratified and non-stratified holdout SVM and logistic regression, as well as the accuracy of five-fold stratified and non-stratified cross-validation SVM and logistic regression, were used to select the best method for feature selection and feature extraction. The stepwise method was found to be the best method for feature selection and feature extraction. Therefore, a new dataset was created that contained 12 chosen variables and the class. The condition of breast cancer was classified based on these 12 variables chosen from the stepwise method. To select the best method for classifying the condition of breast cancer, the new dataset was used. Five stratified and non-stratified holdout and cross-validation were used to obtain training and testing data. The accuracy of

SVM and logistic regression in classifying the condition of breast cancer was compared. Linear, polynomial, radial, and sigmoid kernels were used in SVM. The best method for this case was found to be five-fold stratified cross-validation using SVM with a radial kernel. This method was able to accurately classify the condition of breast cancer with an accuracy of 81.816% and a variance of accuracy of 0.94%.

REFERENCES

Agresti, A. (2007) *An Introduction to Categorical Data Analysis*, New Jersey: John Wiley & Sons, Inc.

Allo, Caecilia B. G., et al. (2023) 'Perbandingan Metode Klasifikasi Kegagalan Simulasi Model Iklim', Koloni: Jurnal Multidisplin Ilmu, 2(1), pp. 242 – 249. https://doi.org/10.31004/koloni.v2i1.438.

Handayani, Fitri. (2021) 'Komparasi Suppor Vector Machine, Logistic Regression dan Artificial Neural Network dalam Prediksi Penyakit Jantung', *JEPIN: Jurnal Edukasi & Penelitian Informatika*, 7(3). http://dx.doi.org/10.26418/jp.v7i3.48053.

Han, J., Kamber, M., and Pei, J. (2012) *Data Mining: Concepts and Techniqu*, United States of America: Elsevier Inc.

Han, J., et al. (2019) 'Performance of Logistic Regression and Support Vector Machine for Seismic Vulnerability Assessment and Mapping: A Case Study of The 12 September 2016 ML5.8 Gyeongju Earthquake, South Korea', *Sustainabilty*, 11(24). https://doi.org/10.3390/su11247038.

Mustafa, Ahmed., et al., (2018) 'Comparing Support Vector Machines with Logistic Regression for Calibrating Cellular Automata Land Use Change Models', *European Journal of Remote Sensing*, 51(1). https://doi.org/10.1080/22797254.2018.1442179.

Nadh, Kamal., & Saraswathi, S. (2023) 'Using Logistic Regression vs The Support Vector Machine Algorithm for Stroke Prediction', *Journal of Survey in Fisheries Sciences*, 10(1S), pp. 2675 – 2682. https://doi.org/10.17762/sfs.v10i1S.497.

Nurlaily, Diana., *et al*. (2022) 'Classification of Hepatitis Patients Using Logistic Regression and Support Vector Machine Methods', *Jurnal Pendidikan Matematika (Kudus)*, 5(2), pp. 237 – 254. http://dx.doi.org/10.21043/jpmk.v5i2.17052.

Khandezamin, Ziba., Naderan, Marjan., Rasthi, M. J. (2020) 'Detection and Classification of Breast Cancer Using Logistic Regression Feature Selection and GMDH Calssifier', *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2020.103591.

Sultana, Jabeen and Jilani, Abdul Khader. (2018) 'Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers', *International Journal of Engineering & Technology*, 7(4.20), pp. 22 – 26. https://doi.org/10.14419/ijet.v7i4.20.22115.

UCI Machine Learning (2023) *Breast Cancer Wisconsin (Prognostic) Data Set*. Available at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29.

World Health Organization (2023) *Cancer*. Available at https://www.who.int/health-topics/cancer.

World Health Organization (2023) *Breast Cancer*. Available at https://www.who.int/news-room/fact-sheets/detail/breast-cancer.