Model Evaluation for Logistic Regression and Support Vector Machines in Diabetes Problem

Baiq Siska Febriani Astuti, Neni Alya Firdausanti , Santi Wulan Purnami
Statistics Department, Faculty of Mathematics, Computing, and Data Science
Institut Teknologi Sepuluh Nopember Surabaya
Email: santi_wp@statistika.its.ac.id

Abstract— Machine learning is a method or computational algorithm to solve problems based on data that already available from the database. Classification is one of the important methods of supervised learning in machine learning. Support Vector Machine and Logistic Regression are some supervised learning methods that can be used both for classification and regression. In datamining process, Preprocessing is an important part before doing further analysis. In preprocessing data, feature selection and deviding training and testing data are important part of preprocessing data. In this research will be compared some evaluation model of deviding method for training and testing data, namely Random Repeated Holdout, Stratified Repeated Holdout, Random Cross-Validation, and Startified Cross-Validation. Evaluation model would be implying in logistic regression and Support Vector Machines (SVMs). From the analysis, can be concluded that by selecting features can improve the accuracy of classification with logistic regression, but opposite of Support Vector Machines (SVMs). For training and testing data pertition method can not be sure what method is better, because each method of partition training and testing data using the concept of random selection. Model evaluation cannot sure influence to increase best perform for SVMs model in particular this case.

Keywords: Calssification, Cross-validation, Feature Selection, Logistic Regression, Preprocessing, Repeated Holdout, Support Vector Machine

I. INTRODUCTION

Along with the development of science and technology, recently has developed a variety of data analysis methods that can be applied to solve the problems especially to datamining problem. Machine learning is one method of datamining which is currently widely used in various fields of science. Machine learning is a method or computational algorithm to solve problems based on data that already available from the database [1]. The purpose of machine learning is to capture data patterns consistently or systematically analyze relationships among data, validate findings by applying detected patterns and predicting new findings on new datasets [2]. Classification is one of the important methods of supervisid learning in machine learning [3]. According to [4] classification is a multivariate technique that deals with the separation of different sets of objects (or observations) and by allocating new objects (Observation) against pre-defined groups.

There are several methods of classification in machine learning that can be used, one of them is the Support Vector Machine. SVM is a supervised learning method that can be used both for classification and regression. Currently there are many studies that discuss about SVM. Research conducted by [5] states that SVM provides better classification precision and lower error rates than other classification algorithms. However, SVM is very sensitive to the value of cost parameters and kernel parameters. There are several methods that can be used to obtain the best value consist of cost parameters and kernel parameters. In this research, used grid search method to find the best cost paremeters and kernel parameters. In addition to using SVM, in this study will also be done classification analysis by using classical methods that will serve as a performance comparison of SVM. Binary logistic regression is one of the classical classification methods where the dependent variable consists of two categories [6].

Preprocessing is an important part before doing further analysis. In preprocessing data, it is necessary to perform feature selection stage to select attributes that have a significant influence in determining the differences between classes. Purnami, Rahayu and Embong [7] mentioned that by doing feature selection can improve accuracy of classification. The feature selection method used in this research is backward elimination. In addition to feature selection, deviding training and testing data is also an important part before analyzing the classification data. Poor training data will result in poor models that will result in poor implementation [8]. In this research will be compared some evaluation model of deviding method for training and testing data, namely Random Repeated Holdout, Stratified Repeated Holdout, Random Cross-Validation, and Startified Cross-Validation. Evaluation model would be implies in logistic regression and Support Vector Machines (SVMs).

II. DATA SET

This paper will be solving classification problem which is used diabetes datasets. Diabetes dataset has showed diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurement included in dataset. All of The object of dataset is females at least 21 years old of pima Indian heritage. There are two variable that we used in this paper, it is dependent variable and independent variables. Dependent variable, which is type of patients detected diabetic or not diabetic. The independent variables are number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index ($\text{weight in kg} / (\text{height in m})^2$), Diabetes

pedigree function, and Age of Patients. The completeness of diabetes dataset given in table 1.

Table 1 Information of diabetes dataset

Attribute	Information
Y	Class
	0 = non-diabetes
	1 = diabetes
X_1	Number of times pregnant
X_2	Plasma glucose concentration a² hours in an oral glucose tolerance test
X_3	Diastolic blood pressure (mm Hg)
X_4	Triceps skin fold thickness (mm)
X_5	2-Hour serum insulin (mu U/ml)
X_6	Body mass index (weight in kg/(height in m)²)
X_7	Diabetes pedigree function
X_8	Age (years)

III. LITERATURE REVIEW

A. Binary Logistic Regression

Logistic regression is a method used to explain the relationship between dependent variable in the form of dichotomic / binary data with independent variables in the form of interval and or categorical data [6]. The dichotomic / binary variable is a variable that has only two categories, namely the category that states the success event (Y = 1) and the category that states the failed event (Y = 0). The general form of logistic regression probability model with explanatory variable p, is formulated as follows:

$$\pi(x) = \frac{}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)} \quad (1)$$

It needs to be changed in logit form to use linear function in order to be seen from independent variable and non-free variable. By doing the transformation of the logit π(x), then we get a simpler equation as follows:

$$g(x) = \ln \frac{\pi(x)}{[1-\pi(x)]} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2)$$

B. Support Vector Machine

Support Vector Machines (SVMs) is method for solving classification case which have high dimension. SVM is the one of supervised learning models with associated learning algorithm that used for classification and regression case. The main idea of SVM is to construct a hyperplane representing the decision surface such the margin of separator, also called decision boundary, between the two classes are maximized. Practically, this boundary is a strip with a maximum width between the classes. The width of the strip is determined based on the samples lying close to the boundary of the two classes to be separated. These samples are called support vectors where the name of the algorithm is derived from [9].

Considering Vapnik as a state of the art of machine learning algorithm, classical SVM is based on guarantees risk bounds of statistical learning theory which is known as structural risk minimization (SRM) principle [10]. Binary classification problem of classifying m objects belonging to two sets denoted by I_1 dan I_2 . Each object the i equal with 1,2,…,m has n features which are stored in the i-th row of an mxn matrix **X.** For each object i, y_i defines its label as follow: $y_i = 1$ if object i belongs to set I_1 or $y_i = -1$ if object *I* belongs to set I_2 [11].

SVMs method tries to build a decision boundary using hyperplane, that is to separating between two classes. Figure 1 shown hyperplane for two possible hyperplanes.

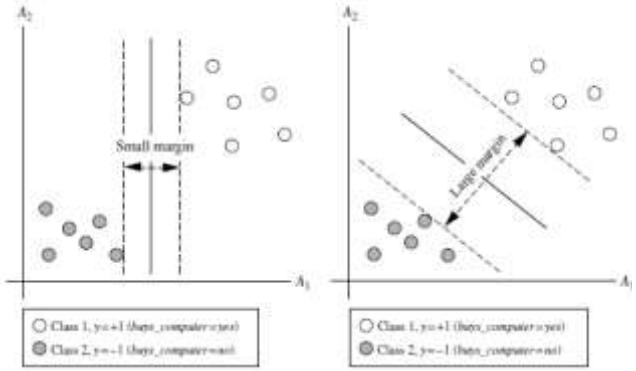

Figure 1 Two possible separating hyperplanes and their associated margin

Although we can use both of hyperplanes on figure 1, the one with large margin (right side) should have greater generalization accuracy [2].

SVMs can be solved non-linearly separable data or nonlinear data. There are two step to doing it. Firstly, we transform the original input data into a higher dimensional space using nonlinear mapping. Second step is searches for linear separating hyperplane in the new space. We again end up with a quadratic optimization problem that can be solved using the linear SVM formulation. The maximal marginal hyperplane found in the new space corresponds to a nonlinear separating hypersurface in the original space.

Therefore, non-linearly separable case is the solving problem in classify case. Instead of sloving that problem, practically, it is simpler to optimize its dual form. With a specific kernel function $k(,)$, the optimization problem in the dual form [9]. There are some kernel function that we used in this paper:

1) Linear kernel

$$K(x, x_i) = x^T x_i \quad (3)$$

2) Polynomial kernel

$$K(x, x_i) = \left(\gamma x^T x_i + r\right)^d, \gamma > 0 \quad (4)$$

3) Radial Base Function kernel

$$K(x, x_i) = \exp\left(-\gamma \left\| x - x_i \right\|^2\right)^d, \gamma > 0 \quad (5)$$

C. Feature Selection

Feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results. There are three general classes of feature selection consist of filter method, wrapper method and embedded methods.

Filter method is method for selection feature with ranked of dataset by the scoring. Wrapper method is selection problem with search problem, after that evaluated and compare another combination. The last, embedded method is selection the feature with best contributed to the accuracy while the model being created [12].

D. Model Evaluation and Selection

Before we have built a classification model, there are many process to get best result in classification model. We would like an estimate of how accurately our model can be predict classification problem. Therefore, in this subsection we explain about model evaluation and selection. There are three kind of method which it is metrics for evaluation classifier performance, holdout method and k-fold cross validation.

Firstly, a confusion matrix is one of the common approaches to measure performance for classification model. We defined metrics for evaluation classifier performance. In a confusion matrix, the two classes are identified as positive class (+1) and negative class (−1). As shown in Table 2, each predicted class is compared with its actual class for each instance to calculate four metrics:

Table 2 confusion matrix

		PREDITED POSSITIVE	
		Positive	Negative
ACTUAL CLASS	Positive	TP	TN
	negative	FP	TN

1) True Positives (TP) –the number of positive instances that is correctly classified as positive classes.
2) False Positives (FP) –the number of negative instances that is incorrectly classified as positive classes.
3) True Negatives (TN) –the number of negative instances that is correctly classified as negative classes.
4) False negatives (FN) –the number of positive instances that is incorrectly classified as negative classes

Furthermore, the confusion matrix can be derived as follow:

$$Accuracy : a = \frac{TN + TP}{TN + TP + FN + FP} \quad (6)$$

$$Sensitivity : r = \frac{TP}{TP + FN} \quad (7)$$

$$Specificity : s = \frac{TN}{FP + TN} \quad (8)$$

Accuracy is many observations have classification which have accurately their category. Sensitivity is many observationd have positively category which have accurately their category. Furthermore, Specificity is many observations have negatively category which have accurately their category.

Secondly, holdout method is what we have alluded to so far in our discussions about accuracy. In this method, the given data are randomly partitioned into two independent sets consist of a training set and a test set. Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated to the test set [2]. Holdout validation avoids the overlap between training data and test data, yielding a more accurate estimate for the generalization performance of the algorithm [13].

Finally, Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments consist of one used to learn or train a model which it defined training set and the other used to validate the model which it defined testing data. The basic form of cross-validation is k-fold cross-validation. In k-fold cross-validation the data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k-iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining k-1 folds are used for learning [13].

IV. RESULTS

In this section, we present numerical result on diabetes datasets. The result implies our classified about patients have diabetic or have not diabetic. In this case, we used binary logistic regression and support vector machines. Furthermore, diabetes dataset has many missing values, there are 0.49% missing values in our datasets. Therefore, we generate the missing value using predictive mean matching in R. In this problem, we using feature selection to reduce our dimension which it has high dimension.

In order to, we first compared the classification result of two method which is binary logistic regression and support vector machines. Furthermore, we evaluated result of classification using accuracy, sensitivity and specificity for each method. The last, we get the best model to solving diabetes problem.

A. Data Summary

Before performing classification, it is necessary to know the caracteristic differences of each features that influence the determiation whether the female patients of Pima Indian Heritage have a diabetes or not. Pima Indian Dataset consist of 768 patients with 8 features and 2 categories on dependent variable.

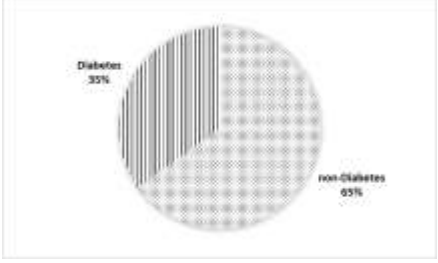

Figure 2 Proportion of Each Classes

There are 500 patients who are not diabetic and 268 patients with diabetes.

B. Preprocessing

Pima Indian Dataset consist of several attributes that have missing values, those attributes are Glucose, Diastole, Tricep Thickness, Insulin and BMI. All atributes in this dataset are not skewed, so missing values imputation method used in this case is the attribute mean for all samples in each class. After all the missing values are imputed, then the next step is detect the outlier in each class. There are 34 outliers for patients with diabetes and 129 outliers for patients who are not diabetic. There are several ways to handling the outlier, such as correct value, transformation, remove outlier, keep the [1] lier and so on [14]. In this study the outliers is kept by acknowledging the presence of an outlier and doing nothing to the outlier values prior to the analysis. Pima Indian Dataset is standardized because the range of data is too wide.

C. Logistic Regression on Dataset

Dependent variable on Pima Indian Dataset consists of two categories: patients who are not diabetic and patients with diabetes. The first step before doing binary logistic regression is to check whether there are multicollinearity cases in the dataset or not. Each attribute has a VIF value less than 10, there are no multicollinearity ceses in the Pima Indian Dataset.

Before classifying using binary logistic regression, the dataset is divided into training data and testing first. In this study there will be some combination of training and testing data split methods, those methods are repeated holdout stratification, randomly repeated holdout, cross validation stratification and randomly cross validation. Five repetitions are used in each of the repeated holdout and five folds are used in each cross-validation method. From each training and testing data will be evaluated by the accuracy and variance of accuracy so that will get the best training and testing dataset.

Table 3 Evaluation of Binary Logistic Regression for All Features

		1	2	3	4	5	Mean
Stratified Holdout	Accuracy	0,727	0,766	0,753	0,740	0,760	0,749
	Sensitivity	0,798	0,845	0,904	0,873	0,874	0,859
	Specificity	0,580	0,608	0,517	0,481	0,529	0,543
Random Holdout	Accuracy	**0,760**	0,740	0,701	0,779	0,747	**0,745**
	Sensitivity	**0,879**	0,885	0,816	0,843	0,876	**0,860**
	Specificity	**0,545**	0,500	0,500	0,654	0,526	**0,545**
Statifed Cross-Validation	Accuracy	0,747	0,745	0,773	0,758	0,773	0,759
	Sensitivity	0,830	0,810	0,900	0,910	0,860	0,862
	Specificity	0,593	0,623	0,537	0,472	0,611	0,567
Random Cross-Validation	Accuracy	0,805	0,758	0,740	0,752	0,766	0,764
	Sensitivity	0,862	0,827	0,896	0,894	0,893	0,874
	Specificity	0,667	0,636	0,483	0,525	0,510	0,564

From table 3 it can be seen that deviding method of training and testing data using all attributes that conduct the highest average accuracy is randomly repeated holdout with accuracy of 0.745, sensitivity of 0.86 and specificity of 0.545. Table 3 shows that from the results of deviding training and testing data using randomly repeated holdout, the training and testing dataset that produce the best accuracy are the first holdout

In determining whether the patients are diabetic or not there are some different attributes, so it is necessary to do the feature selection that are expected to determine which attributes that can classify the diabetic patients appropriately and effectively. Feature selection used in this study is backward elimination. The attributes selected from forward selection are Number of Pregnancy, DPF, Glucose and BMI

Table 4 Evaluation of Binary Logistic Regression for Selected Features

		1	2	3	4	5	Mean
Stratified Holdout	Accuracy	0,714	0,727	0,812	0,734	0,766	0,751
	Sensitivity	0,840	0,880	0,947	0,848	0,935	0,890
	Specificity	0,481	0,500	0,593	0,527	0,516	0,524
Random Holdout	Accuracy	**0,825**	0,805	0,799	0,818	0,779	**0,805**
	Sensitivity	**0,898**	0,909	0,914	0,909	0,863	**0,899**
	Specificity	**0,696**	0,618	0,551	0,655	0,644	**0,633**
Statifed Cross-Validation	Accuracy	0,747	0,765	0,786	0,752	0,773	0,764
	Sensitivity	0,850	0,840	0,910	0,860	0,910	0,874
	Specificity	0,556	0,623	0,556	0,547	0,519	0,560
Random Cross-Validation	Accuracy	0,747	0,752	0,773	0,765	0,805	0,768
	Sensitivity	0,909	0,845	0,928	0,819	0,885	0,877
	Specificity	0,455	0,589	0,509	0,678	0,585	0,563

Table 4 shows that deviding method of training and testing data using selected attributes that conduct the highest average accuracy is randomly repeated holdout with accuracy of 0.805, sensitivity of 0.899 and specificity of 0.633. Table 4 shows that from the results of deviding training and testing data using randomly repeated holdout, the training and testing dataset that produce the best accuracy is also the first holdout, same as the results of the full model. It can be seen in table 3 and table 4, by doing the feature selection can increase the value of classification accuracy. Then the best model for binary logistic regression is a model using the selected variable in the first random holdout.

Table 5 Parameter Estimates for the Best Model

	B	SE Estimate	Z	P-value	Odds Ratio exp(B)
Intercept	-0,833	0,105	-7,973	0,000	0,435
Num_preg	0,445	0,100	4,466	0,000	1,561
DPF	0,259	0,107	2,428	0,015	1,295
Glucose	1,050	0,116	9,040	0,000	2,858
BMI	0,550	0,108	5,067	0,000	1,733

Table 5 shows that each attribute is significant. The increase in the number of pregnancies, the risk of diabetes patients increased 1.56 times. Increased Diabetes pedigree function will increase the risk of diabetes patients by 1,295-times. For the attributes of Plasma glucose concentration a 2 hours in an oral glucose tolerance test showed that increased glucose tolerance would increase the risk of diabetes by 2,858-times. Any increase in body mass index will increase the risk of diabetes patients by 1.7333. Obtained binary logistic regression model is as follows.

$$\pi(x) = \frac{\exp(-0{,}833 + 0{,}445\text{Num\_preg} + 0{,}259\text{DPF} + 1{,}050\text{Glucose} + 0{,}550\text{BMI})}{1 + \exp(-0{,}833 + 0{,}445\text{Num\_preg} + 0{,}259\text{DPF} + 1{,}050\text{Glucose} + 0{,}550\text{BMI})}$$

From the model obtained, then performed the validation on the data testing.

Table 6 Confussion Matrix for Testing Data in Binary Logistic Regression

Actual	Prediction	
	0	1
0	88	10
1	17	39

Table 4 shows the confusion matrix for data testing and the accuracy of the classification using binary logistic regression with atribute selected is 82.5%.

D. Support Vector Machine

Support Vector Machine is one of non-linear classification method. This method included to supervised method. In this subsection, we imply the result of classification use SVMs. Previously we got training and testing data use hold-out and k-fold cross validation. We take 5th in k-fold cross validation.

Table 7 shown perform of classification model SVMs, it is result of analysis using SVMs with all of dimension in our problem. There are eighth features consist of number of times pregnant, plasma glucose, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function and the last is age of patient. Table 6 shown perform SVMs model with three kinds of kernel. We can see that the high accuracy is kernel linear with holdout stratified selection model. The mean of accuracy is 78%, meaning 78% model linear with holdout repeated stratified can be correctly classifier diabetic or not diabetic.

Table 8 is the perform our SVMs model which it shown how well our classifier can recognize tuples of different class. Table 8 shown results of analyze using model SVMs which there are three kinds of kernel consist of linear, radial base function and polynomial. In this result of our classification has six feature that we use in this part, consist of Glucose, BMI, Age, Insulin, Number of Pregnant and DPF. We got that from selection model using feature selection.

Table 7 Evaluation of Support Vector Machine for All Features

	Kernel	1	2	3	4	5	Mean
Stratified Holdout	linear	0.799	0.760	0.799	0.779	0.786	0.784
	Polynomial	0.740	0.682	0.734	0.740	0.766	0.732
	RBF	0.766	0.734	0.779	0.786	0.760	0.765
Random Holdout	linear	0.727	0.805	0.779	0.708	0.779	0.760
	Polynomial	0.714	0.786	0.760	0.721	0.701	0.736
	RBF	0.727	0.786	0.773	0.682	0.760	0.745
Statifed Cross-Validation	linear	0.805	0.727	0.753	0.708	0.805	0.760
	Polynomial	0.688	0.779	0.753	0.662	0.760	0.729
	RBF	0.740	0.721	0.766	0.721	0.786	0.747
Random Cross-Validation	linear	0.766	0.714	0.766	0.825	0.753	0.765
	Polynomial	0.740	0.675	0.714	0.786	0.753	0.734
	RBF	0.753	0.695	0.760	0.805	0.753	0.753

The below table shown about perform our SVMs model. We utilize iteration the 5th. The average of accuracy model for radial base function using stratified Cross-Validation has greater accuracy than another model. We can see the accuracy is 77%. The meaning of it is 77% accurately our model predicted tuples of different classes in another word its means RBF model with stratified Cross-Validation can be correctly classifier diabetic or not diabetic. Thus, we note that although the classifier has a high accuracy, but it has recognized positive classes poor.

Therefore, the result of our analysis explained there is no influence about evaluation model on our case. Because of there are different results between the mean of accuracy, sensitivity, and specificity. We can see on below table the best accuracy is radial base function with selection model k-fold repeated stratified. The value of its 0.77 or 77% which means 77% RBF model with stratified Cross-Validation can be correctly classifier diabetic or not diabetic.

Table 8 Evaluation of Support Vector Machine for Selected Features

	Kernel	1	2	3	4	5	Mean
Stratified Cross-Validation	linear	0.753	0.766	0.766	0.779	0.779	0.769
	Polynomial	0.740	0.734	0.714	0.753	0.747	0.738
	RBF	0.773	0.773	0.779	0.760	0.766	0.770
Random Cross-Validation	linear	0.760	0.714	0.779	0.825	0.760	0.768
	Polynomial	0.727	0.669	0.721	0.799	0.760	0.735
	RBF	0.779	0.675	0.773	0.825	0.766	0.764
Random Holdout	linear	0.792	0.753	0.747	0.773	0.721	0.757
	Polynomial	0.766	0.701	0.760	0.721	0.734	0.736
	RBF	0.792	0.760	0.753	0.766	0.740	0.762
Stratified Holdout	linear	0.734	0.753	0.760	0.753	0.792	0.758
	Polynomial	0.740	0.740	0.760	0.714	0.682	0.727
	RBF	0.753	0.760	0.799	0.753	0.773	0.768

V. CONCLUSIONS

1 The best training dan testing dataset that obtain the best accuracy for binary logistic regression is Random repeated Holdout on the first holdout. The best accuracy (82,5%) is obtained form model with selected attributes.

2 Feature selection in this case cannot influencing accuracy using model SVMs. We can see from accuracy which it is get from SVMs model with all of feature and SVMs model with feature selection model. The accuracy of both model is 77% and 78%.

3 Because of there are difference in the value of accuracy, sensitivity, and specificity on each evaluation model. We cannot sure that evaluation model influence to increase perform of our model with SVMs in particular for this case.

4 Therefore, each method of partition training and testing data use concept of random selection, cannot be sure what method is better, it needs to be done several times trial to get the partition training and testing dataset to obtain the best model.

REFERENCES

[1] I. H. Witten and E. Frank, Data Mining Practical Learning Tools and Techniques, 2nd Edition ed., United States of America: Morgan Kaufmann, 2005.

[2] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, 3rd Edition ed., USA: Morgan Kaufmann, 2012.

[3] S. Kim, Z. Yu, R. M. kil and M. Lee, "Deep learning of support vector machines with class probability output networks," *Neural Networks*, pp. 19-28, 2015.

[4] R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis, New Jersey: Pearson Education, 2007.

[5] F. P. S. Rachman, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer dengan Menggunakan Regresi Logistic Ordinal dan Support Vector Machine," *Jurnal Sains dan Seni ITS*, vol. 1, 2012.

[6] D. Hosmer, S. Lemeshow and R. Sturdivant, Applied Logistic Regression, 3rd Edition ed., New Jersey, USA: John Wiley & Sons, 2013.

[7] S. Purnami, S. Rahayu and A. Embong, "Feature Selection and Classification of Breast Cancer Diagnosis Based on Support Vector Machines," *IEEE*, pp. 1-6, 2008.

[8] B. Sitthidah and J. S. Maurice, "Comparing Training Method for a New Interactive Whiteboard," *International Symposium on Human Factors and Ergonomics in Health Care: Improving Outcomes*, pp. 15-18, 2016.

[9] A. O. Kusakci, B. Ayvaz, a. Karakaya and E., "Towards An Autonomous Human Chromosome Classification System using Competitive Support Vector Machines Teams (CSVMT)," *Expert Systems With Applications*, pp. 224-234.

[10] X. Peng and J. Shen, "Twin-Hyperspheres Support Vector Machine with Automatic Variable Weights for Data Classification," *Information Sciences*, pp. 216-235, 2017.

[11] S. M. J. a. M. J. Maldonado, "Redefining Support Vector Machines with The Ordered Weighted Average," *Knowledge-Based System*, pp. 41-46, 2018.

[12] B. Jason, An Introduction to Feature Selection, 2014.

[13] P. Refaeilzadeh, L. Tang and H. Liu, Cross-Validation, Arizona State University, 2008.

[14] H. Aguins, R. K. Gottfredson and H. Joo, "Best-Practice Recommendation for Defining, Identifying, and Hangling Outliers," *Organizational Research Methods*, pp. 270-301, 2013.