Performance Evaluation of Linear Discriminant Analysis And Support Vector Machines to Classify Caesarian Sections

ABSTRACT (10 PT)

This study aims to examine the performance of the LDA-SVM method in predicting whether a prospective mother needs to perform a Caesarean section or not. The data used in this study is the dataset of caesarian section. This data consists of the results of 80 pregnant women who following cesarean section for the birth method with the most important characteristics of labour problems in the clinical field. The results obtained are the LDA-SVM method is able to classify properly with an accuracy of up to 100%. This research is also able to surpass the methods in previous studies. The results show that LDA-SVM for this case study generates accuracy of 100.00 %.

1. INTRODUCTION (10 PT)

Caesarean section is the process of giving birth to a baby which is done by cutting the abdomen to the uterus of the mother. The incision in the abdomen is a way out of the baby from the womb. Doctors usually make long incisions in a horizontal direction just above the pubic bone. The method of giving birth by cesarean section is usually done around the 39th week, or when doctors recommend a mother to undergo this operation. Usually the doctor will recommend a caesarean method if your pregnancy is at risk.

Compared to giving birth normally through the vagina, giving birth by caesarean section does require a longer healing time. That is why it is important to consult a doctor first before deciding to undergo this birthing procedure. Caesarean section should generally be done if you experience certain complications in pregnancy. These complications can usually complicate the process of giving birth normally or through the vagina.

Even if forced to carry out normal childbirth, the risk is endangering the health and safety of the mother and baby. This is where the doctor will suggest the choice to undergo labor by Caesarean method. Caesarean section can be planned from the beginning or the middle of pregnancy, or when complications arise before the time of giving birth.

Currently, worldwide, the rate of births for babies using c-section is increasing. By performing a cesarean section, the woman or the baby does not get any benefit in terms of health. So, this problem caught the attention of WHO. Due to the importance of limiting the increasing use of caesarean section in childbirth, this year, new guidelines on non-clinical interventions specifically designed to reduce unnecessary caesarean section were published by WHO.

In line with the directions given by WHO, some researchers have also developed an intelligent system to predict whether a mother needs to have a cesarean section or not. Some of them are carried out by Amin et al and Suwarno et al. Amin and Ali 2018 [1] analyzed the system created using machine learning methods such as Support Vector Machines (SVM), Random Forests (RF), Naive Bayes (NB), K-Nearest Neighbors (KNN) and Logistic Regression (LR). Whereas Suwarno and Santo 2019 [2] used the Neural Network method to predict eligibility in the caesarean method.

The SVM method has been widely used to solve cases in the medical field such as Diabetes [3, 4], Brain Detection [5], Cancer [6, 7], Heart Disease [8] and others and the accuracy obtained is also quite frustrating. However, in research conducted by Amin et al, the SVM method used only obtained an accuracy of around 76% [1]. In this study the author tries to modify the SVM method used by trying to combine it with a method for reducing the dimensions of the data, namely Linear Discrimination Analysis. In addition to measuring the accuracy obtained, researchers also to reduce data dimensions and SVM to do the classification and try to analyze the results obtained.

2. THEORIES

In this section, we will discuss the machine learning methods used in this study such as Linear Discriminant Analysis as a reduction dimension model and Support Vector Machines as a classifier.

2.1. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [9, 10, 11] is a statistical method that is quite popular in solving cases related to pattern recognition. Initially LDA was used to perform facial recognition. This method is known as Fisher's Linear Discriminant. However, at this time, LDA is more widely used to extract important information from data and reduce data dimensions.

This method is basically intended to find linear combinations of features that are the main characteristics, so that the data can be separated into two or more object classes or groups. Once the data has been processed using LDA, new data will have a more dispersed distribution and can finally increase recognition success. After that, the new data combination obtained can be reprocessed using both linear and non-linear classification methods.

In this study, LDA is used to find the main characteristics of the Caesar dataset and if possible, from the existing data, the researcher also reduces the dimensions so that it is expected that the results of the dimensional reduction can be classified properly by SVM. The following is the LDA algorithm: First of all we will define the covariance matrix in class (S_W) and the inter-class covariance matrix (S_B), each of which is exemplified as follows:

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \quad (1)$$

$$S_W = \sum_{i=1}^{c} N_i (x_k - \mu_i)(x_k - \mu_i)^T \quad (2)$$

Dimana: x_k = data ke-k
C = total class
N_i = amount of data in the class-i
μ = average from all data
μ_i = average data in class-i

So that the covariance matrix in class (S_W) can be minimized then we maximize the covariance matrix between classes (S_B). Next we can determine eigenvector (V) so that we can maximize the ratio of equation (3):

$$\frac{\det(VS_BV^T)}{\det(VS_WV^T)} \quad (3)$$

Resulting in a solution:

$$S_BV = \lambda S_WV \quad (4)$$

Then the eigenvalue (λ) and eigenvector (V) values of the covariance matrix equation (4) are found, namely:

$$cov = S_B S_W^{-1} \quad (5)$$

Because the LDA method discussed in this paper is the LDA method with 2 classes, the eigenvector used is c-1. The eigenvector value can be based on the largest eigenvalue of the covariance matrix. After the eigenvector is known, the LDA feature can be determined by calculating:

$$F_x = \sum_{i=1}^{k} (x_i - \mu)^T \times V \quad (6)$$

The data that has been reduced is then processed using SVM.

2.2. Support Vector Machines

The method used in this study is the Support Vector Machine (SVM). In addition to use in the health sector, SVM had also been widely applied in other areas i.e Pattern Recognition [12], Information Retrieval [13], Militer [14], Forest Detection [15], etc. SVM is a classification method in data mining. SVM can also make predictions in both classification [12] and regression [13] (learning to rank). Basically, SVM has a linear principle, but at this time SVM has evolved and then it can be utilized to solve non-linear cases. The way SVM works in non-linearly separable issues is to include the kernel concept. In this dimensionless field, we will seek a separator or often called a hyperplane. The Hyperplane which SVM want to find is the one which can maximize the margin between two classes. We can find the best hyperplane between two data by estimating the edge and afterward finding the maximum point from the separator. The effort to find the best hyperplane as a class separator is the core of the process in the SVM method. The kernel function used in this study is the RBF kernel :

$$K(x_i, x_j) = exp^{\left(\frac{\|x-x_i\|^2}{2\sigma^2}\right)} \quad (7)$$

3. EXPERIMENTAL METHOD

The framework of this study can be seen in Figure 1. In this section will explain related to the dataset and evaluation methods used by this research.

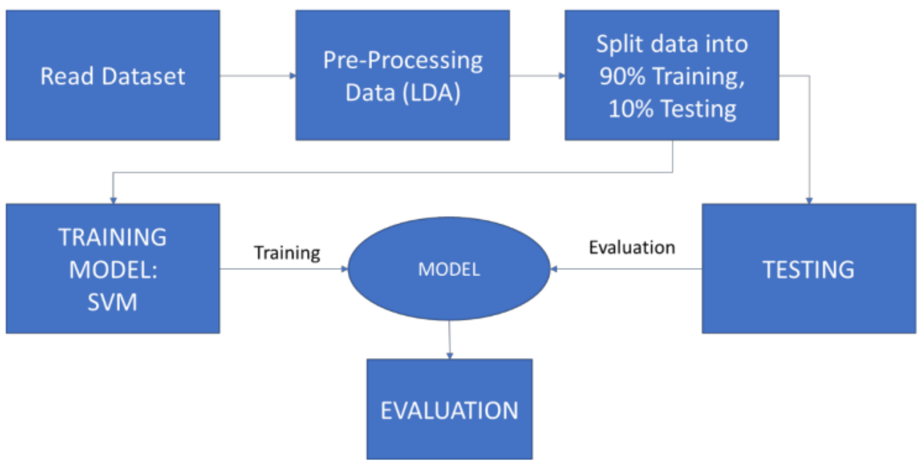

Figure 1. Research framework

3.1. Caesarian Dataset

The dataset utilized in this investigation is the dataset of caesarian section [18]. This dataset contains data about caesarian segment consequences of 80 pregnant mother with the most significant attributes of giving birth issues in the clinical field. The dataset comprises of 80 cases, made out of five ascribes which are the most significant attributes of childbirth issues specifically age, number of pregnant, conveyance time, circulatory strain and heart status. These credits are ordered in Table 1.

Table 1. Explanation of Dataset

Attribute	Value
'Age'	{ 22,26,28,27,32,36,33,23,20,29,25,37,24,18,30,40,31,19,21,35,17,38 }
'Delivery number'	{ 1,2,3,4 }
Delivery time'	{ 0,1,2 } -> {0 = timely , 1 = premature , 2 = latecomer}
'Blood of Pressure'	{ 2,1,0 } -> {0 = low , 1 = normal , 2 = high }
'Heart Problem'	{ 1,0 } -> {0 = apt, 1 = inept }
Caesarian Class	{ 0,1 } -> {0 = No, 1 = Yes }

3.2. Confusion Matrix

To evaluate the performance of classifier, in this article the writer uses confusion matrix (Table 2 and Table 3). Confusion matrix is a method commonly used to calculate accuracy in the concept of data mining or decision support systems. To measure the performance of the LDA-SVM method researchers used Accuracy as an evaluator. Accuracy can be described as the appropriate number of predictions divided by the total of all data. Prediction of the accuracy of the formation of classification models can be obtained by the following formula [19]:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Table 2: Confusion Matrix

		Actual Values	
		Caesar	**Non-Caesar**
Prediction	**Caesar**	TP	FP
	Non-Caesar	FN	TN

Table 3: Confusion Matrix Definition

Label	**Definition**
True Positive (TP)	The number of Caesar tuples which are correctly classified
True Negative (TN)	The number of non-Caesar tuples which are correctly classified
False Positive (FP)	The number of non-Caesar that are classified as Caesar
False Negative (FN)	The number of non-Caesar that are classified as Caesar

4. RESULTS AND ANALYSIS

The results of all the experiments can be seen in table 4. At the pre-process stage, the researcher divides the data into training data and testing data randomly so that each same experiment will produce a different data group. Training data and testing data consists of three groups namely 70% training data (30% test data), 80% training data (20% test data) and 90% training data (10% test data).

Table 4 shows the accuracy of all experiments using LDA-SVM. The parameter used in this research is RBF Kernel sigma (σ). This method was then tested on Caesar data with a composition of training data varying from 10% (90% testing data) to 90% (10% training data).

Based on the results of experiments that have been conducted, the results of classifying Caesar data obtained mixed results. Accuracy The highest accuracy of the LDA-SVM method is obtained when the data used as training is 80% and 90%. The classification results obtained are 100% of the data can be classified correctly. The sigma parameters that produce perfect scores are 4, -6.1 and -6.6 in the 80% training data and 4 and 16 in the 90% training data.

We also display the results of the average accuracy of training data and testing data from 10 experiments which can be seen in Figure 2. Based on Figure 2 it appears that the highest average accuracy in experiments with a composition of 70% training data is 74.70% which is obtained in the 10th experiment with the sigma parameter used is 4.0, the composition of the 80% training data is 100%, that is in experiments 3, 4, 6, and 7 with the sigma parameters used are -6.6, -6.1, 4.0 and 4.0, respectively. Meanwhile, the composition of 90% was obtained during the 8th and 10th experiments with the parameters of the 4th and 16th sigma.

Table 4. Accuracy For All Experiment

Exp	Sigma	70% Training	30% Testing	Sigma	80% Training	20% Testing	Sigma	90% Training	10% Testing
1	-5.9	67.86	70.83	-9.6	98.44	93.75	-9.6	97.22	100.00
2	-3.6	62.50	58.33	-9.4	96.88	81.25	-9.4	98.61	100.00
3	2.4	62.50	70.83	-6.6	100.00	100.00	-6.6	97.22	100.00
4	2.4	64.29	70.83	-6.1	100.00	100.00	-6.1	97.22	100.00
5	2.4	64.29	62.50	-4.4	98.44	100.00	-4.4	98.61	100.00
6	2.4	62.50	62.50	4.0	100.00	100.00	4.0	98.61	100.00
7	4.0	66.07	70.83	4.0	100.00	100.00	4.0	97.22	100.00

8	4.0	55.36	58.33	4.0	98.44	100.00	4.0	100.00	100.00
9	4.0	66.07	58.33	16.0	93.75	81.25	16.0	98.61	100.00
10	4.0	66.07	83.33	16.0	95.31	100.00	16.0	100.00	100.00
Akurasi Rata-rata		63.75	66.67		98.13	95.63		98.33	100.00

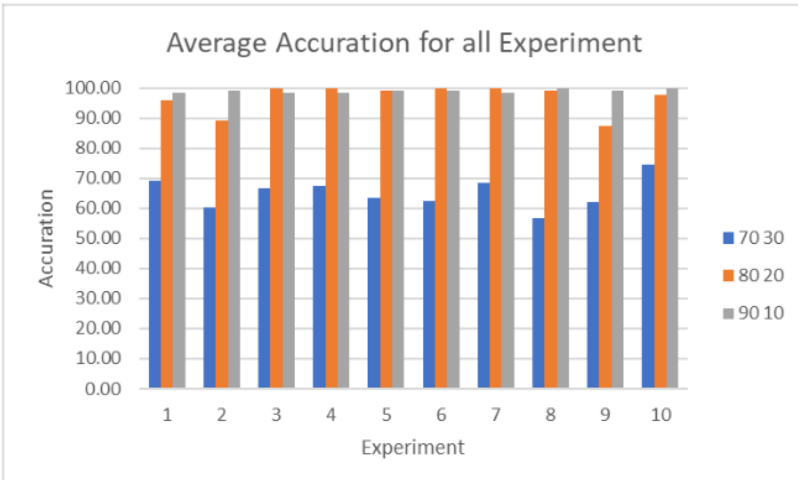

Figure 2. Average Accuration for All Experiment

Based on table 4 and figure 2 it can also be seen that the resulting increase in accuracy is not directly proportional to the greater sigma parameters used. In addition, based on the results of the average accuracy of each group of 70%, 80% and 90%, the greater the data used as training data, the greater the accuracy obtained.

Table 5 shows the results of comparison of accuracy using other methods in Caesar's data. When compared with other methods that have been done before (table 5), namely SVM, Random Forest, Naive Bayes, Logistic Regression, K-Nearest Neighbors and Neural Network, the performance of LDA-SVM is the best among the five other classifiers. We can say that because of the right combination of methods and use of sigma parameters. We hope this research can help the medical field in determining the birth method appropriately both in Indonesia and throughout the world.

Table 5. Performance Comparation with another Classifiers

Classifiers	Accuracy
SVM	76.3%
Random Forest	95.0%
Naïve Bayes	76.3%
Logistic Regression	77.5%
K-Nearest Neighbours	95.0%
Neural Network	77.5%
LDA-SVM*	100.0%

5. CONCLUSION

The main purpose of our research was to discover the performance of the Linear Discriminant Analysis classification method combined with Support Vector Machine (LDA-SVM) to classify Caesar data. LDA-SVM succeeded in doing the Caesar classification very well that is with an accuracy of 100%. The dataset used as training data in this study was 80% and 90%. In addition, based on the results of experiments and comparisons with previous methods (Table 1) that have been done it can be seen that LDA-SVM produces better accuracy than previous experiments. When compared with the SVM Method with no reduction in dimensions through LDA which only reached 76%, in the case of Caesar's data, it can be concluded that the accuracy of classification using SVM is affected by dimensional reduction using LDA.

ACKNOWLEDGEMENTS

The authors would like to thank Politeknik Negeri Jakarta, for the support of research funding, through the Research Grant in the Field of Science and Institutional Development in 2020.

Telkominika

ORIGINALITY REPORT

7%	7%	2%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	archive.ics.uci.edu Internet Source	3%
2	Submitted to Politeknik Negeri Bandung Student Paper	1%
3	Peng, S.H.. "Texture feature extraction based on a uniformity estimation method for local brightness and structure in chest CT images", Computers in Biology and Medicine, 201011/12 Publication	1%
4	Submitted to Study Group Australia Student Paper	1%
5	pure.ulster.ac.uk Internet Source	1%
6	Submitted to Universiti Teknikal Malaysia Melaka Student Paper	1%
7	link.springer.com Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		