Support vector machines for classification of low birth weight in Indonesia

Ning Eliyati, Alfensi Faruk *, Endang Sri Kresnawati, Ika Arifieni
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Sriwijaya University

*alfensifaruk@unsri.ac.id

Abstract. This paper proposes support vector machines (SVMs), which is currently one of the most popular algorithms in machine learning (ML), in order to classify the low birth weight (LBW) data. The main objectives of this study are to predict the classification of LBW data in Indonesia based on the SVMs and to compare the performance of the proposed SVMs with the binary logistic regression as the most common model for classification of LBW data. The obtained samples were based on the results of Indonesian Demographic and Health Survey in 2012. The results showed that SVMs with four kernel functions (linear, radial, polynomial and hyperbolic tangent) were fit well to the LBW data in Indonesia. Furthermore, the constructed SVMs based on linear kernel function had the best performance among the SVMs with the other proposed kernel functions. This research also concluded that the SVMs based on linear kernel competed well with the binary logistic regression for classification LBW data in Indonesia.

1. Introduction
Low birth weight (LBW) can be defined as a birth weight less than 2,500 gr. Many studies have shown that there is a high correlation between LBW and the number of infant mortality cases [1]. In particular, the LBW can increase infant mortality rate. To reduce the number of LBW cases, therefore, the first effort that should be conducted by the policymakers is to identify what factors that could potentially influence the incidence of LBW.

LBW commonly occurs in third-world countries including Indonesia. Moreover, as one of the most populous countries in the world, Indonesia has more LBW incidence than its neighbour countries in South East Asia region [2]. According to the most recent basic health survey in Indonesia [3], the percentage of the incidence of LBW is about 10.2% where the highest rate is Middle Sulawesi Province (16.8%) and the lowest rate is in North Sumatra Province (7.2%).

Many works have been conducted in order to examine the potential factors affecting the LBW. The incidence of LBW may be caused by seven main potential factors, namely genetics (gene and interaction with environment), nutrient adequacy (pregnant mother’s nutrients, protein and energy adequacy), mother’s weight and characteristics (pregnant mother’s body weight and parity), medical records of the mother (anemia, malaria, syphilis and rubella), pregnancy complication (infection when giving birth), mother’s lifestyle (smoking and alcohol consumption) and environment (population and socio-economic factors) [2]. Meanwhile, the recent work has shown that five significant socio-economic factors affecting the LBW in Indonesia are the place of residence, time zone area,
household’s wealth index, father’s educational level and the number of child born when giving birth [4].

The majority of works employed conventional statistics such as binary logistics regression in order to predict the LBW rates based on the potential factors. However, Faruk et al [4] have shown that random forest as one of the machine learning (ML) approaches outperformed the binary logistic regression in order to classify the LBW cases in Indonesia. They used the secondary data from the results of Indonesian Demographic and Health Survey (IDHS) in 2012. However, there are other approaches in ML that could be proposed for classification task of LBW data, such as support vector machines (SVMs), Naive-Bayes and decision tree. These ML techniques have been widely applied in various real-life problems and they often outperformed the conventional statistical methods [5,6].

One of the advantages of ML techniques is their excellent ability to handle big data which is commonly found in few recent years [7]. This task is very difficult to be conducted by using any conventional statistical method since the computational process of the parameter estimations are too slow or to complicated. Moreover, unlike the conventional parametric statistical methods, ML approaches do not give any distributional assumption on the dataset. Therefore, the overall aim of this study was to propose an ML approach for predicting the classification task of the LBW in Indonesia. The proposed ML approach was only limited to SVMs since it is a popular ML technique for binary classification task. Furthermore, the data used in this research was based on the results of the work by Faruk et al [4] and the whole computational process was conducted by using R-Studio version 1.1.383 and R Data Miner (Rattle) GUI version.

2. Materials and methods
This research employed the SVMs approach to the LBW dataset provided by Faruk et al [4]. The dataset was extracted from the raw dataset from the results of 2012 IDHS. In the beginning, the dataset contained 45607 women aged 15–49 years but after preprocessing the dataset was reduced to 12055 women aged 15–49 who gave birth between May 2007 and July 2012 as the observation period. The considered independent variables in Faruk et al [4] consisted of seven categorical variables and one continuous variable. However, this work transformed the continuous variable, namely age of the mother, to a categorical variable which contained three age categories (under 20, between 20 and 30, and more than 30). Moreover, all independent variables used in this research are described in Table 1.

The steps of the research methods conducted in this work are data preprocessing, data transformation, dividing the dataset into two groups (training data and testing data), model building using training data, target prediction using testing data, data evaluation, comparing the predicting performance of SVMs, and data deployment based on the proposed models. In addition, this research also compared the prediction performance of SVMs for the LBW dataset with binary logistic regression as the most applied model for LBW classification.

3. SVMs for binary classification
SVMs has been known as a popular ML approach for classification, regression, and other tasks. The algorithm of SVM was introduced by Cortes and Vapnik [8] for binary classification. Their approach may be roughly described as follows:

1. The fundamental concept of SVMs algorithm is to find the optimal hyperplane which separates two considered classes by maximizing the distance between the classes’ closest points. Any point lying on the borderline is defined as support vector. Furthermore, the middle of the margin is called the optimal hyperplane.
2. The points that lie on the wrong side are weighted down.
3. To find the optimal hyperplane, a linear separator can be used. However, if it cannot be determined, kernel techniques should be employed to find the optimal nonlinear hyperplane.
4. The nonlinear optimal hyperplane can be assumed to follow the form of a certain function. The most widely used functional forms of the hyperplane are radial, polynomial and hyperbolic tangent function.
Table 1. The summary of independent variables

| Variables          | Symbols | Categories          |
|--------------------|---------|---------------------|
| Place of residence | res     | 0: Urban            |
|                    |         | 1: Rural            |
| Time zone          | tz      | 0: West             |
|                    |         | 1: Middle           |
|                    |         | 2: East             |
| Wealth index       | wealth  | 0: Poor             |
|                    |         | 1: Middle & Above   |
| Mother’s education | m_edu   | 0: No               |
|                    |         | 1: Yes              |
| Father’s education | h_edu   | 0: No               |
|                    |         | 1: Yes              |
| Age of the mother  | agecat  | 0: < 20             |
|                    |         | 1: between 20 and 30|
|                    |         | 2: > 30             |
| Mother’s job       | job     | 0: No               |
|                    |         | 1: Yes              |
| Number of children | child   | 0: < 4              |
|                    |         | 1: ≥ 4              |

4. Results and analysis

This section consists of three subsections, namely data preprocessing, model building, data prediction and data evaluation. Generally, the data preprocessing in this research follows the results from the previous work [4] but this work made a little modification, that is by transforming the variable age of the mother from a continuous variable to a categorical variable with three categories. As a result, this research skipped the skewness checking since the skewness values are not used in the categorical variable. In model building, this research employed SVMs to be fitted to the LBW data as well as the binary logistic regression as the most popular model for the classification of a binary dependent variable like the considered LBW data. The final subsection, namely data prediction and evaluation, is addressed mainly for assessing the proposed model performances. The performances of four proposed kernel for SVMs were compared with each other in addition to binary linear regression.

4.1. Data preprocessing

In order to visualize the data, this work uses histogram for each independent variable. The x-axis represents the related categories while the y-axis is the number of low birth weight babies for each characteristic (Figure 1). It can be seen from Figure 1 that there are eight plots which related to all independent variables. Furthermore, the plots do not indicate that there is no plot which follows a bell-shape pattern. In order words, it means that the LBW data based on each independent variable do not follow a normal distribution. However, this fact does not make any impact on the proposed models since they do not require any distributional assumption on the dataset.
Figure 1 also shows that there are several independent variables which contain outliers, namely \textit{m\_edu}, \textit{h\_edu} and \textit{agecat}. These outliers data are not removed from the dataset since they are very crucial to represent their respected categories. Checking the skewness in the dataset is also an important task since a zero skewness indicates a symmetrical distribution of the data. However, the skewness of this data cannot be computed since it is not meaningful for categorical independent variables.

4.2. Model building

Although in this work the proposed model for the classification of LBW data is SVMs, binary logistic regression is also employed to the considered LBW for model comparison. The model building by using binary logistic regression is not described in this paper since the main role of this model is just only to assess the performance of the SVMs approach. Instead, the binary logistic regression is described in the data prediction and evaluation subsection.

The proposed model has to be built in the training data set meanwhile the prediction task will be based on the testing dataset. Therefore, to obtain these two group of the dataset, the original data which contain 12055 samples are divided into two groups. The first group became the training data (80% of the total samples) and the second group became the testing data (20% of the total samples). Furthermore, the SVMs method is fitted to the training data by using four common kernel functions of the hyperplane, namely Gaussian radial basis, quadratic polynomial, linear and hyperbolic tangent.
Table 2 demonstrates the complete results of the model fitting with four hyperplane functions. According to Table 2, the number of vectors obtained from the Gaussian radial basis is the highest compared with the results from the other kernel functions. The high number of vectors indicates the lower of the training error of the proposed model. This has been shown by the training error of the SVMs based on the Gaussian radial basis which has the lowest value (6.7%) among the other models. Therefore, it can be concluded that the Gaussian radial basis is the best kernel function for the SVMs approach in the training data.

4.3. Data prediction and evaluation

The prediction performance of the proposed model should be applied to the testing data. In this phase, the SVMs with four kernel functions were employed to the testing data as well as the binary logistic regression. The three methods that were used to assess the prediction performance in this work are confusion matrix, area under curve (AUC) and receiver operating characteristic (ROC) curve. The computation results of the first two criteria are depicted in Table 3. Meanwhile, the results of the ROC curves for each proposed model can be seen in Figure 2.

Based on Table 3, the SVMs with three kernel functions (Gaussian radial basis, 2-degree polynomial and linear) and binary logistic regression have the lowest overall confusion matrix error (7.1%). This means that the prediction performance of these four models in order to classify the LBW data is the best among the other models (hyperbolic tangent SVMs). However, the different results have been obtained from the other criterions (AUC and ROC curves). The higher the AUC value, the better the performance of the model is. Meanwhile, if the ROC curve of the model is above the diagonal curve, it represents the good classification results (better than random). On the other hand, if the ROC curve of the model is below the diagonal line, it represents the bad results (worse than random). The ROC curves of the linear SVMs (Figure 2 (d)) and binary logistic regression (Figure 2 (e)) are above the diagonal line which means their performance in LBW classification is good. The other three ROC curves (radial SVMs, quadratic polynomial SVMs and hyperbolic tangent SVMs), on the contrary, are below the diagonal line which shows bad prediction performance of the models.

Table 2. The results of fitting SVMs to training data

| Kernel Function             | Number of Vectors | Objective Function Value | Training Error | Time Taken (Seconds) |
|-----------------------------|-------------------|--------------------------|----------------|----------------------|
| Gaussian Radial Basis       | 1085              | -9204.135                | 0.067397       | 13.4                 |
| Polynomial (2 Degree)       | 1124              | -9280                    | 0.068731       | 26.29                |
| Linear                      | 960               | -9280                    | 0.068731       | 2.46                 |
| Hyperbolic Tangent          | 870               | -93150.57                | 0.124278       | 6.38                 |

Table 3. The assessment of model performance

| Kernel Function              | Overall Confusion Matrix Error (%) | Area Under Curve (AUC) |
|------------------------------|-----------------------------------|------------------------|
| Gaussian Radial Basis        | 7.1                               | 0.4823                 |
| Polynomial (2 Degree)        | 7.1                               | 0.4995                 |
| Linear                       | 7.1                               | 0.5495                 |
| Hyperbolic Tangent           | 12.6                              | 0.5083                 |
| Binary Logistic Regression   | 7.1                               | 0.5643                 |
Figure 2. ROC curves for all proposed models. The curves (d) and (e) represent the ROC curves for linear SVMs and binary logistic regression, respectively. Both models have the better prediction performance compared with other models (radial SVMs, quadratic polynomial SVMs and hyperbolic tangent SVMs).

5. Conclusion
The SVM method with four kernel functions (linear, radial, polynomial and hyperbolic tangent) can be applied to predict the binary classification of the LBW infants in Indonesia. Their predictive performance can be categorized as good since their average predictive error is approximately under 10%. However, the performance of SVMs with linear kernel function is better than SVMs with the other three kernel functions. Furthermore, linear SVMs can compete very well with binary logistic regression as the most widely used model for classifying LBW data.

References
[1] Boardman J D, Powers D A, Padilla Y C and Hummer R A 2002 Low birth weight, social factors, and developmental outcomes among children in the United States Demography, 39 :353-68
[2] UNICEF and WHO 2004 Low birthweight:country, regional and global estimates(New York: UNICEF)
[3] Kemenkes R I 2014 Profil kesehatan Indonesia (Jakarta: Kementerian Kesehatan Republik Indonesia)
[4] Faruk A, Cahyono E C, Eliyati N and Arifieni I 2018 Prediction and classification of low birth
weight data using machine learning techniques, *Indonesian Journal of Science & Technology* 3 18-28

[5] Song X, Mitnitski A Cox J, Rockwood K 2004 Comparison of machine learning techniques with classical statistical models in predicting health outcomes *Stud Health Technol Inform* 107 736-40

[6] Khondoker M, Dobson R, Skirrow C, Simmons A and Stahl D 2016 A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies *Statistical Methods in Medical Research* 25 1804–1823

[7] Ratner B 2011 *Statistical and machine-learning data mining: techniques for better predictive modeling and analysis of big data* (second edition)(Boca Raton: CRC Press)

[8] Cortes C and Vapnik V 1995 Support-vector networks *Machine Learning* 20 273-297