Credit risk analysis using support vector machines algorithm

N H Putri¹, M Fatekurohman¹, and I M Tirta¹

¹Mathematics Department of Mathematics and Natural Sciences Faculty, University of Jember, Indonesia

E-mail: nurikaputri@students.unej.ac.id

Abstract. Credit risk or also known as bad credit risk is defined as the risk that occurs due to the inability of the customer to repay the loan and the interest within a certain period of time. Bad credit can be prevented by selecting proper customers during the loan application process so as not to give losses to the Bank. This research aims to analyze credit risk using machine learning methods. One of the machine learning algorithms that can be used for data classification is Support Vector Machine (SVM). The concept of SVM is to find optimal hyper plane with maximum margin to linearly separate the data set into two classes. The dataset were taken randomly from 2015 to 2018 at Bank XX, as many as 610 data. Dataset was divided into two parts along with percentage of the training data 80% and the testing data 20%. The variables used were gender, plafond, rate, term of time, job, income, face amount, warranty, and loan history as independent variable as well as credit status as dependent variable. The testing of SVM algorithm used linear, polynomial, Radial Basis Function (RBF), and sigmoid kernel obtained confusion matrices with accuracy respectively 0.9262, 0.9508, 0.8934, and 0.8361. Meanwhile, the AUC values were 0.9129, 0.9419, 0.9051, and 0.8285. The SVM model with polynomial kernel is the best model of the four models because it has the highest accuracy and AUC value. Thus, this model can be used to classify prospective customers into good credit or bad credit class with sufficiently high accuracy so as to help banks reduce the risk of bad credit.

1. Introduction
Bank is an institution which is to accumulate funding from community of people in the form of saving and to distribute it to them in the form of credit or others, in order to accelerate living standard of the people [1]. People who wants to do a credit must propose an apply to the bank. The regulation in accepting and rejecting credit proposal considers five requirements; character (customer’s personality), capacity (ability in paying the debt), capital (customer’s capital), collateral (credit risk guarantee), and condition (customer’s financial condition). The accepted applicant needs to sign an agreement to fully pay the debt in certain term. Furthermore, credit risk can rise if the customer is not able to pay his obligation within the term it is called bad credit.

Credit risk analysis can be done by applying some methods, one of them is machine learning. Machine learning is a combination of computer science, engineering, and statistics, so that it also includes as part of statistics [2]. Machine learning makes more use of prediction rules that will be generalized for new data [3]. Machine learning approach is relatively simpler than traditional approach (classical statistic) to deal with complex and big data. So it is suitable for banking sector which mainly processes loads information of customer data and it will take a lot of time if it is done using traditional one. One of the examples of a popular machine learning algorithm is Support Vector Machines (SVM). It is a classification technique using the most optimum hyperplane [4]. The application of
SVM can cope with some problems dealing with gen analysis [5], financial [6], medical field [7], and also banking. Some researches have been done using the SVM algorithm to classify potential customers who are able to pay debts or not in loan companies. However, the evaluation of the model used is only the accuracy and error values [8, 9]. Therefore, the authors are interested in using other evaluation model measures, such as the AUC value.

This research analyzes credit risk using machine learning method, namely SVM algorithm to classify prospective customers into good credit or bad credit class. The classification process of dataset was done by using programming language Python. The dataset were taken randomly from 2015 to 2018 at Bank XX, as many as 610 data. The variables used were gender, plafond, rate, term of time, job, income, face amount, warranty, and loan history as independent variable as well as credit status as dependent variable. This research was done in three stages; data preprocessing, model building, and model testing. Model was made using the training data then checked using the testing data to discover its performance. Model evaluation was counted using confusion matrix which are accuracy, sensitivity, specificity, precision, F1-score, false positive rate, false negative rate, and AUC (Area Under the Curve) values.

2. Methods
2.1 Data Preprocessing
Machine learning algorithms work with a specific data format. Therefore, the raw dataset we have must be prepared in advance to be the right data format. The data preprocessing stage consists of several stages, there are data cleaning, data type checking, and data normalization.

2.1.1 Data Cleaning. Dataset was cleaned before building the model in order to get the best decision quality, such as data whose attribute was lost, error, or incomplete. This matter can be overcome by filling up incomplete values or deleting it from dataset.

2.1.2 Data Type Checking. Data type on plafond, rate, term of time, salary, and face amount variables are float data type. Meanwhile, data type on variable gender, job, warranty, loan history and credit status are string data type. The string data type must be converted into float data type or category so that it can be processed using machine learning algorithm on Python.

2.1.3 Data Normalization. Data was normalized to avoid refraction on learning using min-max method.

\[
x_i^1 = \frac{x_i - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{newmax}_A - \text{newmin}_A) + \text{newmin}_A
\]

where

- \(x_i\) = value before normalized;
- \(x_i^1\) = value after normalized;
- \(\text{min}_A\) = the minimum value in attribute A;
- \(\text{max}_A\) = the maximum value in attribute A;
- \(\text{newmin}_A\) = the minimum limit in the range of new value;
- \(\text{newmax}_A\) = the maximum limit in the range of new value.

2.2 Support Vector Machines (SVM)
SVM is machine learning algorithm used to classify data into certain group in statistics. SVM concept is to find hyperplane that separate the dataset into two linear classes [10]. Hyperlane is a term used generally for all dimensions. For example, the separator of two classes in the 1st dimension is a point. The separator in the 2nd dimension is a straight line. The separator in the 3rd dimension is a flat plane. The separator in the 4th dimension is a three-dimensional plane and so on. So, the term
hyperplane refers to a point, a straight line, a two-dimensional plane, or other high-dimensional (three or more) planes.

The limitation in deciding two different classes in SVM uses optimal hyperplane with maximum margin. Margin is a distance between hyperplane and nearby pattern from each class. This nearby pattern is called support vector. Margin \( m \) can be formulated as \( \frac{2}{\|w\|} \). Margin is inversely proportional to \( \|w\| \), which means maximizing \( m \) is the same as minimizing \( \|w\| \).

Figure 1. Hyperplane SVM.

Dataset are generally non-linearly separated. In order to solve the problem, SVM uses kernel trick concept on higher dimension space. Kernel is generally used for turn dataset on input space into feature space with higher dimension. \( \Phi: R^p \rightarrow R^q \), where \( p < q \)

According Mercer theory, formulation of dot product from data on feature space can be replaced by kernel function \( K(x_i, x_j) \) which implicitly defines its \( \Phi \) transformation.

\[
K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)
\]  

(3)

Kernel trick concept gives a convenience in determining support vector that we only need to know the utilized kernel function, not the concrete shape or form of non-linear function \( \Phi \). There are generally four types of kernel functions that can be used in SVM [10]. The four types of kernel functions are shown in Table 1.

### Table 1. SVM Kernels.

| No. | Kernel Type                | Formula                                           |
|-----|----------------------------|---------------------------------------------------|
| 1   | Linear                     | \( K(x_i, x_j) = x_i^T x_j \)                    |
| 2   | Polynomial                 | \( K(x_i, x_j) = (x_i^T x_j + 1)^p \)            |
| 3   | Gaussian/RBF (Radial Basis Function) | \( K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \) |
| 4   | Sigmoid                    | \( K(x_i, x_j) = \tanh(\alpha x_i^T x_j + \theta) \) |

Optimum hyperplane searching which maximize margin can be formulated as Quadratic Programming (QP) that is solved using Langrange Multiplier.

\[
L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)
\]

(4)

with a limit \( \sum_{i=1}^{n} \alpha_i y_i = 0 \), \( \alpha_i \geq 0 \) (\( i = 1, 2, \ldots, n \)). \( w \) and \( b \) values can be obtained by counting:

\[
\sum_{i=1}^{n} \alpha_i y_i \Phi(x_i) = w
\]

(5)

\[
b = y_i - w^T \Phi(x_i)
\]

(6)
The maximization of this function results a number of positive $\alpha_i$. The data related to positive $\alpha_i$ is called support vector (SV), which is the most outside data of the two classes. After the solution to the QP problem is found (positive $\alpha_i$), the optimal hyperplane function is presented:

$$f(\Phi(x)) = w.\Phi(x) + b$$

$$= \sum_{i=1}^{n_S} \alpha_i y_i \Phi(x_i) + b$$

$$= \sum_{i=1}^{n_S} \alpha_i y_i K(x, x_i) + b$$

There are decision function of data class based on the optimal hyperplane function.

$$g(x_d) = \text{sign}(f(x_d)) = \begin{cases} 1, & f(x_d) > 0 \\ -1, & f(x_d) < 0 \end{cases}$$

2.3 Model Evaluation

At this stage, the evaluation of model uses a confusion matrix, namely the values of sensitivity, specificity, precision, FPR, FNR, F1-score, and AUC to see the model’s performance. Model evaluation can be done by using certain measure, where TP, TN, FP, and, FN are elaborated as follows [11].

1. TP (true positives), the amount of data with positive actual value and positive predictive value.
2. TN (true negatives), the amount of data with negative actual values and negative predictive values.
3. FP (false positives), the amount of data with negative actual values and positive predictive values.
4. FN (false negatives), the amount of data with positive actual value and negative predictive value.

Those four terms can be described as confusion matrix.

| Actual class      | Predicted class |
|-------------------|-----------------|
| Positive (bad credit) | Negative (good credit) | Positive (bad credit) |
| Negative (good credit) | TN | FP |
| Positive (bad credit) | FN | TP |

Evaluation using confusion matrix resulted in several values are as follows [12]:

1. Accuracy (recognition rate)

$$\text{ACC} = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN}$$

2. Recall or sensitivity or True Positive Rate (TPR)

$$\text{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN}$$

3. Specificity or True Negative Rate (TNR)

$$\text{TNR} = \frac{TN}{N} = \frac{TN}{FP+TN}$$

4. Precision or Positive Predicted Value (PPV)

$$\text{Precision} = \frac{TP}{TP+FP}$$

5. F or F1 or F-score

$$F - \text{score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

6. False Positive Rate (FPR)

$$\text{FPR} = 1 - \text{TNR} = \frac{FP}{FP+TN} = \frac{FP}{N}$$

7. False Negative Rate (FNR)

$$\text{FNR} = 1 - \text{TPR} = \frac{FN}{FP+TN} = \frac{FN}{P}$$
8. ROC (Receiver Operating Characteristic) Curve
   ROC curve is a technique to visualize and to test performance of a classifying model based on its performance [13]. ROC is two-dimension graphic which FPR represents horizontal axis (x) and TPR represents vertical axis (y). The better classification model is the one with a larger ROC curve.

9. Area Under the Curve (AUC)
   AUC is used to count the area under the ROC curve and it is counted by adding trapezoid area of the ROC curve. The AUC value is in the range of 0 and 1 and the closer to 1 the better the model in classify the data [14].
   The AUC value is divided into several groups [13].
   1. 0.90 – 1.00 = Excellent Classification
   2. 0.80 – 0.90 = Good Classification
   3. 0.70 – 0.80 = Fair Classification
   4. 0.60 – 0.70 = Poor Classification
   5. 0.50 – 0.60 = Failure

2.4 Variable Dummy
   Dummy variables are variables used to quantify qualitative variables, for example gender and job variables. Dummy variables change categorical variables by encoding them into binary variables. The dummy variable only has two values, namely 0 and 1. A value of 0 usually indicates a group that did not receive a treatment and a value of 1 indicates the group that received treatment. For example, the job variable has 3 categories, namely entrepreneurs, fisherman, and teachers. When changing the job variable into a dummy variable, it will become 3 columns, namely the entrepreneur job column, the fisherman job column, and the teacher job column. Label 1 for one category column and label 0 for other category column. If the data is entrepreneur then entrepreneur job column is labeled 1 and the fisherman and teacher job column is labeled 0.

3. Experimental Results
   The SVM algorithm has no theoretical guideline to find definitely better kernel type – between SVM with linear kernel, polynomial, RBF, or sigmoid, because each dataset is varied that can find its own optimum hyperplane. The research tested four kernel models to discover the quality of each decision-maker model. The important stages before building the model are data preprocessing where the incomplete data is cleaned. Then the data type checking stage is carried out. There are 5 variables with numeric data types and 5 variables with alphanumeric / string data types. The variables with the string data type then converted into a numeric data type by labeling it as follows, namely the variables of gender (2 categories: P = 1 and L = 0), collateral (2 categories: BPKB = 1 and SHM = 0), loan history (2 categories: good = 0 and bad= 1), and credit status (2 categories: good = 0 and bad = 1). While job variable because it has 5 categories (entrepreneur, salt pond owner, shrimp pond owner, shop owner, trader, transportation service, fisherman) is changed to the category type using dummy variables. So that the total variables dataset now are 16 variables. Then this dataset is split into the training data (80%) and the testing data (20%). Furthermore, the classification models are obtained using the training data for each SVM model with a linear, polynomial, RBF, and sigmoid kernel.

   One of popular model evaluations for measuring the quality of model performance is confusion matrix. There are confusion matrices of SVM models with kernel linear, polynomial, RBF, and sigmoid shown in Figure 2. Negative class (good credit) is labeled with 0 and negative class (bad credit) is labeled with 1.
Figure 2 shows the values of TP, TN, FP, and FN for the SVM model with linear, polynomial, RBF, and sigmoid kernels, and there are indicators of evaluation matrices of SVM models.

**Table 3.** The evaluation values of SVM models.

| SVM    | ACC   | TPR   | TNR   | PPV   | FPR   | FNR   | F1-score |
|--------|-------|-------|-------|-------|-------|-------|----------|
| Linear | 0.9406| 0.9174| 0.9482| 0.8538| 0.0518| 0.0826| 0.8845   |
| Polynomial | **0.9529** | **0.9339** | **0.9591** | **0.8828** | 0.0409 | 0.0661| **0.9076** |
| RBF    | 0.9324| 0.9339| 0.9319| 0.8188| 0.0681| 0.0661| 0.8726   |
| Sigmoid| 0.8135| 0.7025| 0.8501| 0.6071| 0.1499| 0.2975| 0.6513   |

In Table 3, the accuracy values of SVM linear, polynomial, and RBF are 0.9406, 0.9529, and 0.9324, respectively. Thus it has regarded as a good classifier. Meanwhile, the accuracy of SVM sigmoid is 0.8135. Thus it has considered as a fairly good classifier. The smallest error for the good class (FPR) and the smallest error for the bad class (FNR) are in the SVM model with a polynomial kernel. It means that the polynomial model has the smallest proportion of errors in classifying data both in good and bad classes. So that it is expected to reduce the level of bank losses. In the banking sector, the model is better predict good customers becoming bad customers than bad customers becoming good customers, because it is better to prevent before bad events actually occur. Therefore, it is necessary to pay attention the models with the lowest FNR value.

In order to see model performance in classifying new data, here are the confusion matrices from prediction the testing data.
Figure 3. Confusion matrices from prediction the testing data.

Figure 3.a shows the confusion matrix of the SVM model with linear kernel where the TN value of 89 means there are 89 good customers were predicted to be good, the TP value of 24 means there are 24 bad customers were predicted to be bad, the value of FP 6 means there are 6 good customers were wrongly predicted to be bad, and the value of FN 3 means there are 3 bad customers were wrongly predicted to be good by the SVM model with linear kernel. Likewise in Figure 5.b which shows the values of TN 91, TP 25, FN 2, and FP 4 from the SVM model with a polynomial kernel. Figure 5.c shows the values of TN 84, TP 25, FN 2, and FP 11 from the SVM model with the RBF kernel and Figure 5.d shows the values of TN 80, TP 22, FN 5, and FP 15 from the SVM model with a sigmoid kernel. The highest TN value is SVM polynomial kernel. The highest TP values were SVM polynomial kernel and RBF. The lowest FN values were polynomial SVM and RBF. The lowest FP value is SVM polynomial kernel.

Figure 3 shows the values of TP, TN, FP, and FN for the SVM model with linear, polynomial, RBF, and sigmoid kernels using the testing data, and there are model evaluation indicators are presented in Table 4.

Table 4. Model evaluation values from prediction the testing data.

| SVM     | ACC  | TPR  | TNR  | PPV  | FPR  | FNR  | F1-score |
|---------|------|------|------|------|------|------|----------|
| Linear  | 0.9262 | 0.8889 | 0.9368 | 0.8000 | 0.0632 | 0.1111 | 0.8421   |
| Polynomial | **0.9508** | **0.9259** | **0.9579** | **0.8621** | **0.0421** | **0.0741** | **0.8929** |
| RBF     | 0.8934 | 0.9259 | 0.8842 | 0.6944 | 0.1158 | 0.0741 | 0.7937   |
| Sigmoid | 0.8361 | 0.8148 | 0.8421 | 0.5946 | 0.1579 | 0.1852 | 0.6875   |
Table 4 shows the level of accuracy, TPR (sensitivity), TNR (specificity), PPV (precision), FPR, FNR, and f1-score for each SVM model. The SVM model with four kernels has an accuracy of 0.9262, 0.9508, 0.8934, and 0.8361 which means that all models have shown fairly good accuracy because the value is more than 0.8. The best model based on the accuracy value is the SVM model with polynomial kernel. The sensitivity values are 0.8889, 0.9259, 0.9259, and 0.8148 which shows the proportion of customers who are predicted to be bad class compared to customers who are actually in the bad class. Polynomial and RBF models have the highest sensitivity values. The specificity values are 0.9368, 0.9579, 0.8842, and 0.8421 which shows the proportion of customers who are predicted to be good class compared to customers who are actually good class with high scores. The best models based on the specificity value is the polynomial models. Then the SVM precision model values are 0.8000, 0.8621, 0.6944 and 0.5946 which shows the proportion of customers who are truly bad class compared to all customers who are predicted to be bad class. Here the precision values for the RBF and sigmoid models are much smaller than the other two models. Meanwhile, the best precision value is the SVM kernel polynomial model. The FPR value or the proportion of good class customers who are misclassified from the linear, polynomial, RBF, and sigmoid models are 0.0632, 0.0421, 0.1158, and 0.1579. Then the proportion of bad class customers who are misclassified (FNR) from the SVM model is 0.1111, 0.0741, 0.0741 and 0.1852. The best model based on the FPR value is the SVM model with a polynomial kernel, while the best model based on the FNR value is the SVM model with a polynomial kernel and RBF. Therefore, the SVM model with polynomial kernel can be said to have a fairly good classification performance for recognition in all classes. This can be seen from the high values of accuracy, TPR (sensitivity), TNR (specificity), PPV (precision), and f1-score also the quite small values of the FPR and FNR to avoid misclassification in good or bad class.

Another model evaluation that can be used to measure model performance is the AUC value. AUC looks at model performance based on FPR and TPR values. There are the ROC curves and AUC values of the 4 SVM kernels as shown in Figure 4.

![Figure 4. ROC Curves of SVM.](image)

The ROC curve threshold of SVM with linear, polynomial, RBF, and sigmoid kernels is on (0.0632, 0.8889), (0.0421, 0.9259), (0.1158, 0.9259), and (0.1579, 0.8148), respectively. The ROC curve is constructed by connecting each of these points with points (0,0) dan (1,1). Then, the AUC value can be obtained by finding the area under the ROC curve using the trapezium rule. The AUC values of SVM with linear, polynomial, and RBF kernel are 0.9129, 0.9419, and 0.9051. These models are included in the excellent classification group which means that the model is able to classify data very well. While the AUC value of SVM with sigmoid kernel is 0.8285. It is included in the good classification group which mean that the model is able to classify the data well.
The data visualization of SVM with linear, polynomial, RBF, and sigmoid kernel used income and loan history variables, where here the data of customers who have been bad history is coded 1 and those who have good history are coded as 0.

Figure 5. Plot of the SVM model with income and loan history variables.

The red pattern represents data classified into bad class (1) and the blue pattern represents data classified into good class (0). There are several errors where some good class data is incorrectly classified into bad class. The difference of those plots can be seen from the hyperplane shape. The SVM hyperplane with linear kernel in the form of a smooth straight line, meanwhile the SVM hyperplane with the polynomial, RBF, and sigmoid kernels are curved jagged lines. This is because the polynomial, RBF, and sigmoid kernel functions are nonlinear equations.

4. Conclusion
Based on the results of the testing data prediction using the SVM model with a linear kernel, rbf, polynomial, and sigmoid, confusion matrices are obtained with accuracy values of 0.9262, 0.9508, 0.8934, and 0.8361. Meanwhile, the AUC values were 0.9129, 0.9419, 0.9051, and 0.8285, respectively. It's means that all SVM models can be said to be able to classify data well, but the best model is the SVM model with polynomial kernel because it has the highest value of accuracy and AUC. This model can be used to assist Bank XX in choosing the decision to accept or reject a credit application by classifying the prospective customer data into good credit or bad credit class. So that Bank XX can reduce the risk of loss due to bad credit.

Acknowledgment
Thanks to Dr. Alfian Futuhul Hadi, S.Si., M.Si and Dr.Yuliani Setia Dewi, S.Si., M.Si for their advice and support in this research.

References
[1] Pemerintah Indonesia Undang-Undang Republik Indonesia Nomor 10 Tahun 1998 Perubahan atas Undang-Undang Nomor 7 Tahun 1992 tentang perbankan (Jakarta : Sekretariat negara)
[2] Dangeti P 2017 Statistics for Machine Learning (Mumai: Packt Publishing Ltd)
[3] Van Belle, Pelckmans, Surkens and Van Huffel 2008 Survival SVM: a Practical Scalable
Algorithm European Symposium on Artificial Neural Networks (ESANN) 23-25

[4] Vapnik V N 1998 Statistical Learning Theory (New York: John Wiley & Sons Inc)
[5] Vanitha, Devaraj D and Venkatesulu M 2015 Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection Procedia Computer Science 47 13-21
[6] Okasha and Mahmoud K 2014 Using Support Vector Machines in Financial Time Series Forecasting International Journal of Statistics and Applications 4(1) 28-39
[7] Janardhanan P 2015 Effectiveness of Support Vector Machines in Medical Data mining Journal of Communications Software and Systems 47 13-21
[8] Karaa and Krichene 2012 Credit-Risk Assessment using Support Vectors Machine and Multilayer Neural Network models: A Comparative Study Case of a Tunisian Bank Journal of Accounting and Management Information System 11 25-30
[9] Pandey et al 2017 Credit Risk Analysis using Machine Learning Classifier International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)
[10] Suyanto 2018 Machine Learning Tingkat Dasar dan Lanjut (Bandung: Informatika Bandung)
[11] Han J, Kamber M and Pei J 2012 Data mining: Concepts and Techniques (San Francisco: Morgan Kaufmann)
[12] Wei W W S 2006 Time Series Analysis Univariate and Multivariate Method (Canada: Addison Wesley)
[13] Gorunescu F 2011 Data Mining Concepts Model and Techniques (Springer-Verlag Berlin Heidelberg: Intelligent Systems Reference Library 12 325-326)
[14] Bradley A P 1997 The Use of The Area Under The Curve in The Evaluation of Machine Learning Algorithms Pattern Recognition: Department of Electrical and Computer Engineering 30(7) 1145-1159