e-ISSN: 2614-8404
p-ISSN: 2776-3234

Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results

Saikin *), Sofiansyah Fadli[2], Maulana Ashari[3]
[1,3]Program Studi Sistem Informasi, STMIK Lombok
[2]Program Studi Teknik Informatika, STMIK Lombok
Email: [1]eken.apache@gmail.com, [2]sofiansyah182@gmail.com, [3]aarydarkmaul@gmail.com

Abstract – **The performance of the organizations or companiesare based on the qualities possessed by their employee. Both of good or bad employee performance will have an impact on productivity and the impact of profits obtained by the company. Support Vector Machine (SVM) is a machine learning method based on statistical learning theory and can solve high non-linearity, regression, etc. In machine learning, the optimization model is a part for improving the accuracy of the model for data learning. Several techniques are used, one of which is feature selection, namely reducing data dimensions so that it can reduce computation in data modeling. This study aims to apply the method of machine learning to the employee data of the Bank Rakyat Indonesia (BRI) company, so that it can improve the performance of the classification algorithm by removing some features that have no correlation to the objective label and have a significant effect on the classification results. The method used is SVM method by increasing the accuracy of learning data by using a feature selection technique using a wrapper algorithm. From the results of the classification test, the average accuracy obtained is 72 percent with a precision value of 71 and the recall value is rounded off to 72 percent, with a combination of SVM and cross-validation. Data obtained from Kaggle data, which consists of training data and testing data. each consisting of 30 columns and 22005 rows in the training data and testing data consisting of 29 col-umns and 6000 rows. The results of this study get a classification score of 82 percent. The precision value obtained is rounded off to 82 percent, a recall of 86 percent and an f1-score of 81 percent**.

Keywords – K-Fold Algorithm;SVM Method; Classification; Machine Learning

I. INTRODUCTION

The performance of the organizations or companies are based on the qualities possessed by their employee. both of good or bad employee performance will have an impact on productivity and the impact of the benefits that are obtained by the company. Machine learning enables companies to measure employee performance based on Key Performance Indicator (KPI), by applying learning to historical data, making it easier for companies to make decisions. Machine learning algorithms are often used to be solutions to solve problems related to data learning. Algorithms used are classification, clustering and regression algorithms.

Support Vector Machine (SVM) is a machine learning method based on statistical learning theory and can solve high nonlinear, regression, etc. in the sample space and can also be used as a predictive system identification tool [1]. This algorithm is also flexible, it can be applied to the field of data modeling where data classification and data analysis are regression in nature. SVM is an algorithm for making predictions, both predictions in regression and classification cases. which is how it works to get the optimal separator function (hyperplane) to separate observations that have different target variable values, from the concept promoted by this SVM algorithm which makes SVM work well on high-dimensional data sets, even SVM also uses kernel techniques to map the original data from the original dimension to another dimension which is relatively higher [2]. Prediction using SVM is very sensitive to the value of the parameters, being the soft-marginal value constant C of various kernel parameters [3].

In machine learing, model optimization is part of improving the results of model accuracy for data learning. Several techniques are used, one of which is feature selection, namely reducing data dimensions so that it can reduce computation in data modeling. The main purpose of feature selection is to reduce the number of features used in the classification while maintaining an acceptable classification accuracy [4]. Feature selection can have a big impact on the effectiveness of the resulting classification algorithm [5], in some cases, as a result of feature selection, the accuracy of future classification can be improved [6].

One of the algorithms used for feature selection is the Wrapper method. The wrapper method is a method of selecting features as a blackbox to find the best sub-set of attributes. in a previous study [7] in the form of "Combination of the Correlated Naive Bayes Method and the Wrapper Feature Selection Method for Health Data Classification". The aim of this study was to combine the Correlated Naive Bayes method and Wrapper-based feature selection for health data classification. The stages of this research consisted of several stages, namely (1) collecting the Pima Indian Diabetes and Thyroid dataset from the UCI Machine Learning Repository, (2) pre-processing data such as transformation, scaling, and Wrapper-based feature selection, (3) classification using Correlated Naive Bayes and Wrapper Feature Selection Method, and (4) performance testing based on its accuracy using 10-fold cross validation method. Based on the results of the tests

e-ISSN: 2614-8404
p-ISSN: 2776-3234

that have been carried out, the combination of the Correlated Naive Bayes method with Wrapper-based feature selection gets the best accuracy of both the data used. For the Pima Indian Diabetes dataset, the accuracy is 71.4% and the Thyroid dataset accuracy is 79.38%. Thus, the combination of the Correlated Naive Bayes method and Wrapper-based feature selection resulted in better accuracy without feature selection with an increase of 4.1% for the Pima Indian Diabetes dataset and 0.48% for the Thyroid dataset.

Based on previous research, this study aims to apply the machine learning method to the employee data of the Bank Rakyat Indonesia (BRI) company [19]. The method used is the Support Vector Machine method by increasing the accuracy of data learning with feature selection techniques using the wrapper algorithm. Data obtained from Kaggle data, which consists of training data and testing data. each of which consists of 30 columns and 22005 rows of training data and data testing consisting of 29 columns and 6000 rows. modeling results will predict employees into the best performance class and not.

II. RESEARCH METHODOLOGY

A. Literature Review

Research conducted by [8] with the research title "Seleksi Fitur Warna Citra digital Biji Kopi MenggunaanMetode Principal Component Analysis". The results of this study show that the average training process on feature data after the feature selection process has increased compared to without feature selection. This can be seen from the 5 times the training process with feature selection, the best accuracy value is 90.8%, while without feature selection the best accuracy is 89.6%. in a study conducted by [9] entitled " Principal Component Analysis Support Vector Machine (PCA-SVM) untukklasifikasi Kesejahtraaan Rumah Tanggak di Kabupaten Brebes" The results of household poverty classification in Brebes Regency use PCA-SVM with the RBF kernel approach to training data. obtained 667 respondents who were classified appropriately. There are 93 households classified as not poor and 574 households classified as poor.

In the testing data, there were 285 respondents who were classified appropriately. There are 35 households classified as not poor and 250 households classified as poor. in a study conducted [9] entitled " Seleksi Fitur Dan Optimasi Parameter K-NN BerbasisAlgoritmaGenitika Pada Dataset Medis". This study concludes that a genetic algorithm is applied to select features and optimize k parameters for k's closest neighbors to improve the accuracy of the five medical data sets used as benchmarks. The proposed method proved to be effective in increasing accuracy, and the difference in test results between the five datasets resulted in a significant difference.

Support Vector Machine (SVM) was developed by Vapnik in 1992 together with Bernhard Boser and Isabelle Guyon [10]. SVM is a machine learning method that performs a technique to find a classifier function that can split data into two different classes. 11]. The strategy used is to minimize errors in training data and the Vapnik-Chervokinensis (VC) dimension called Structural Risk Minimization (SRM). The aim of SVM is to get the best hyper-plane separating the two classes [12]. Getting the best hyperplane is the same as maximizing the distance between the hyperplane with the closest pattern from each class (margin). The advantage of the SVM method is its generalizability, which is the ability to classify other data that is not included in the data used in machine learning [13]. Feature selection is a process that involves a subset of feature sets that produce output such as the entire feature set. Feature selection is usually used to select optimal features, reduce dimensions, increase accuracy of classification algorithms, and remove irrelevant features [14]. The feature selection technique is divided into 3 groups namely Filter, Wrapper, and Embedded [15]. Filter-based feature selection research was carried out by [16].

B. Method

The stages of this research were carried out from the stage of collecting the dataset, processing the classification data to testing the classification results. The details of the research stages are drawn as shown in Figure 1.

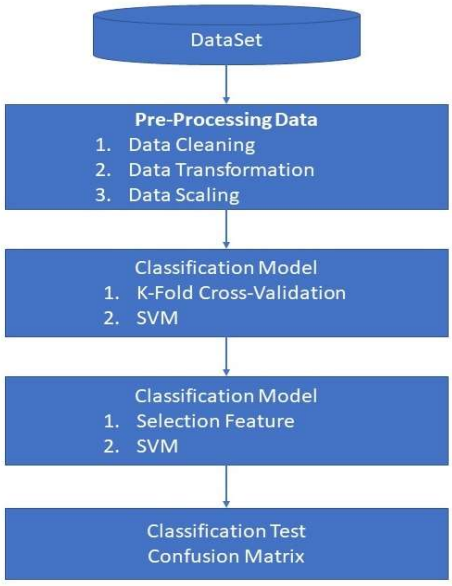

Figure 1. Research Flow

C. Data Retrieval

Data and research variables are secondary data taken from the Kaggle website, in the form of BRI bank employee data consisting of training data and testing data. training data consists of 28 columns and 22005 rows, and testing data consists of 28 columns and 6000 rows. The following displays the features for training data.

Table 1. Dataset Features

	Variabel
1	job_level
2	job_duration_in_current_job_level
3	person_level
4	job_duration_in_current_person_level
5	job_duration_in_current_branch
6	Employee_type
7	Employee_status
8	Gender
9	Age
10	marital_status_maried(Y/N)

e-ISSN: 2614-8404
p-ISSN: 2776-3234

11	number_of_dependences
12	number_of_dependences (male)
13	number_of_dependences (female)
14	Education_level
15	GPA
16	year_graduated
17	job_duration_as_permanent_worker
18	job_duration_from_training
19	branch_rotation
20	job_rotation
21	assign_of_otherposition
22	annual leave
23	sick_leaves
24	Best Performance
25	Avg_achievement_%
26	Last_achievement_%
27	Achievement_above_100%_during3quartal
28	achievement_target_1
29	achievement_target_2
30	achievement_target_3

D. Pre-Processing Data

Data cleaning is the process of data cleaning from several inconsistent data values. Its main purpose is to eliminate misinformation related to data. Data cleaning is carried out on data with high outlier values and data on empty and nan value data. empty data is used for the following, which is a display of missing data based on the highest order:

	Mising Values	% of Total Values
achievement_target_3	6727	30.570325
achievement_target_2	6727	30.570325
achievement_target_1	6727	30.570325
Achievement_above_100%_during3quartal	6302	28.638946
Last_achievement_%	6302	28.638946
Avg_achievement_%	6289	28.579868
Education_level	3608	16.396274
GPA	3503	15.919109
year_graduated	3503	15.919109
job_duration_as_permanent_worker	2055	9.338787
Employee_type	12	0.054533

Figure 2. Data Missing Values

E. Data Transformation

The classification method works with numeric data types, data transformation is used to change the form of data types other than Number into number data types. In the data set used in this study, there are several features that are object data types or categorical data types.

F. Scaling

The use of scaling to reduce the dominance of features whose range values are higher than features with smaller ranges. The use of scaling will affect the classification results because there will be no features that have a dominant value in the classification results, all will be calculated based on a maximum range value of 1 and a minimum. In the dataset, there are several features whose range values are higher than other features.

G. K-Fold Cross-Validation

Determination of training data and test data using the K-fold cross-validation technique where the amount of K is used, namely 4. where using 4 K, will divide the data into 3 subsamples into training and 1 sub sample into testing data. The following illustrates the distribution of training data and testing data using 4K.

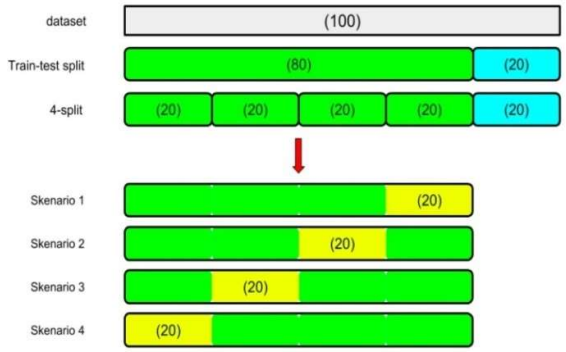

Figure 3. K-Fold Cross-Validation Scenario

H. Feature Selection

Feature selection is used here by looking for correlations that have a high impact on the classification results. features that do not have a significant effect on the classification result will be discarded. The method used is the analysis of variance (ANOVA), which is a statistical method that functions to test the significant effect on the average between groups of variables. The results of data visualization, there are several features that do not have a significant effect on the classification results such as Achievement_above_100% _during3quartal, Best Performance, sick_leaves, GPA, Education_level and gender.

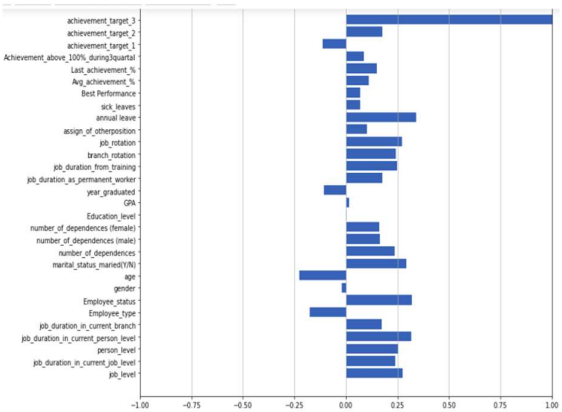

Figure 4. Features and Targets Correlation

I. Support Vector Machine

The fourth step is classification using SVM. The SVM algorithm has the ability to analyze data and perform pattern recognition [12]. SVM does its job by searching for the best hyperline, where what is meant by hyperline is the

e-ISSN: 2614-8404
p-ISSN: 2776-3234

dividing line between two classes [1]. The bigger the margin or dividing line, the smaller the level of misclassification that occurs [24]. SVM is also known as using the parameter penaltykernel method, the kernel contained in SVM such as kernellinear, radial Basis Function and Polynomial. In this study, kernelliniear will be used and the number of parameter penalties of C = 1.0.

III. RESULTS AND DISCUSSION

A. Classification Results Testing

Testing the results of classification using confusionmatrix is to look for true positive, true negative, false positive and false negative values. By applying two stages, namely the first stage by testing the classification results of the combination of the SVM algorithm with cross-validation. The second stage is to test the SVM classification results from feature selection.

B. SVM and K-Fold Cross-Validation Testing

In the classification using the SVM linear kernel and the parameter value C = 1.0 and using 4-fold cross-validation, the resulting average accuracy value is 72 percent.

```
=======================Score SVM dengan 4-fold cross-validation===
[0.72952012 0.72370334 0.73291323 0.72678788]
==========================================================



==================================Rata-rata Score===============
0.7282311432306585
==========================================================
```

Figure 5. SVM Classification with 4-Fold Cross-Validation

The results of the classification test using confusion matrix for classification with 4-fold cross-validation show that the predicted true positive value (true as best performance) is 1939, while the true negative value (correct prediction is no best performance) is 2105. While the false negative value (which is wrongly predicted as (best performance) is 831 and false positive (wrongly predicted as no best performance) is 673.

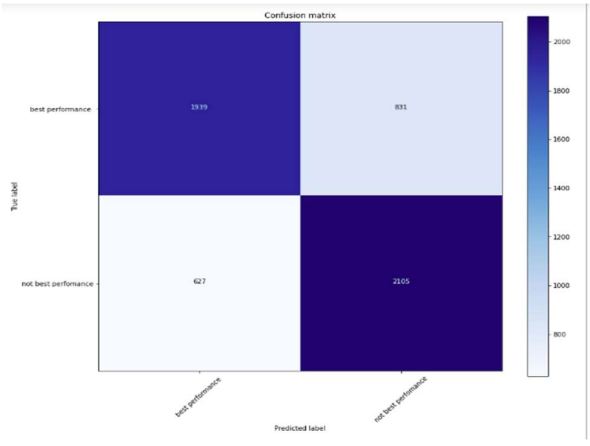

Figure 6. Confusion Matrix for SVM 4-Fold Cross-validation classification

C. SVM Testing With Feature Selection

In the SVM test with a 4-fold cross-validation score, an average of 72 percent was obtained, while using a combination of SVM with feature selection resulted in 82 percent, an increase of about 10 percent. The precision value obtained is rounded off to 82 percent, 86 percent recall and 81 percent f1-score.

```
==========================================================
Hasil klasifikasi SVM dengan Seleksi fitur = 0.8167605889838211
==========================================================
Nilai accuracy        = 0.8158851326790258
Nilai Precision Score = 0.8194404055703874
Nilai Recall          = 0.8667642752562226
nilai F1 Score        = 0.8154561244293783
==========================================================
```

Figure 7. SVM Score ValueandFeature Selection

The test results use confusion matrix, the predicted true positive value is 2121 best performance, and the true negative value is 2368. Meanwhile, the true positive value is 268 while the false negative value is 649.

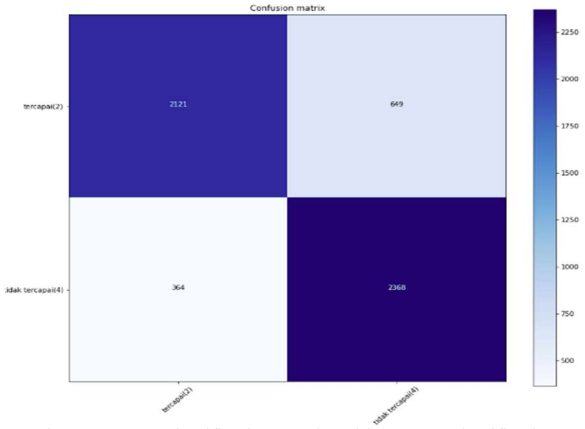

Figure 8. SVM Classification Matrix With Feature Classification

D. Discussion

The results of the confusion matrix measurement show a false positive value and false shows a high enough value. Testing the results of classification using confusion matrix is to find true positive, true negative, false positive and false negative values. By expecting two stages, namely the first stage by testing the classification results of the combination of the SVM algorithm with cross-validation. The second stage is to test the SVM classification results from feature selection.

IV. CONCLUSION

From the results of the classification test, the average accuracy obtained is 72 percent with a precision value of 71 and the recall value is rounded off to 72 percent, with a combination of SVM and cross-validation. To improve the accuracy, a feature selection experiment was carried out and searched for several features by looking for high correlation values and removing some features with low correlation values to the target data. The results obtained from the SVM modeling trials with feature selection obtained accuracy rounded to 82 percent. The accuracy value is 81 percent, the precision value is 82, the recall value is 86 percent and the fi-score is 81 percent. Confusion matrix testing results in true positive 2121 best performance, true negative 2368, true positive 268 while false negative is 649.By applying feature selection to the SVM algorithm, the accuracy value increased from 72 percent to 82 percent, an average increase of 10 percent.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

REFERENCES

[1] Agustian Noor. (2018). *Perbandingan algoritma Support Vector machine biasa dengan support vector mechine berbasis particale swarm optimization untuk prediksi gempa bumi.* Jurnal Humaniora dan Teknologi.DOI:10.34128/jht.v4i1.37.

[2] Prasetyo. (2014). *Data Mining Mengolah data menjadi informasi*. Andi. Yogyakarta

[3] Ultach Enri. (2018). *Optimasi Parameter Support Vector Machine Untuk Prediksi Nilai Tukar Rupiah Terhadap Dolar Amerika.* Jurnal Gerbang.Vol 8 No 1.

[4] Raymer, M. L. Punch, W. F., Goodman, E. D., Kuhn, L. A., & Jain, A. K. (2000). *Dimensionality reduction using genetic algorithms.* IEEE Transactionson Evolutionary Computation, 4(2), 164-171.

[5] Jain, A., & Zongker, D. (1997). *Feature Selection: Evaluation, Application and Small Sample Performance.* IEEE Transactions on Pattern Analysis andMachine Intelligence. 19(2). 153-158.

[6] Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook (Second Edition ed.).* New York: Springer.

[7] Hairani., Muhammad Innuddin. (2019).*Kombinasi Metode Correlated Naive Bayes dan Metode Seleksi Fitur Wrapper untuk Klasifikasi Data Kesehatan.* Jurnal Teknik Elektro Vol. 11 No. 2

[8] Rizki Tri Prasetio. (2020). *Seleksi Fitur Dan Optmiasi Parameter K-NN Berbasis Algoritma Genitika Pada Dataset Medis.*Jurnal Renponsif. Vol. 2. No. 2.

[9] Diani. (2017). *Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data Microarray untuk Deteksi Kanker,* Indonesian Jurnal Of Computing. Bandung.Vol. 2 No. 1. DOI: 10.21108/INDOJC.2017.2.1.169.

[10] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques 3rd Edition.* USA: Morgan Kaufmann.

[11] Vapnik, V. N. (2002). *The Nature of Statistical Learning Theory 2nd Edition.* New York: Springer-Verlag

[12] Gunn, S. (1998). *Support Vector Machines for Classification and Regression.* Southampton: University of Southampton.

[13] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, " Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 13, no. 5, pp. 971–989, 2016, DOI: 10.1109/TCBB.2015.2478454.

[14] E. Hancer, B. Xue, and M. Zhang, " Differential evolution for filter feature selection based on information theory and feature ranking," Knowledge-Based Syst., vol. 140, pp. 103–119, 2018, DOI: 10.1016/j.knosys.2017.10.028.

[15] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," Genomics, vol. 112, no. 2, pp. 1173–1181, 2020, DOI: 10.1016/j.ygeno.2019.07.002.

[16] Abdillah, Abdul, Azis., Prianto, Budi. (2019). *Pembelajaran Mesin Menggunakan Principal Component Analysis dan Support Vector Machines untuk Mendeteksi Diabetes.* J. Matem. Sains. DOI Number: 10.5614/jms.2019.24.1.2.

[17] Buntoro, G.A. (2017). *Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter.* INTEGER: Journal of Information Technology, Vol. 2, Ed. 1. DOI: 10.31284/j.integer.2017.v2i1.95

[18] Hikmawan, S., Pardamean, A., Khasanah, S.N., (2020).*Sentimen Analisis Publik Terhadap Joko Widodo Terhadap Wabah Covid-19 Menggunakan Metode Machine Learning.* Jurnal Kajian Ilmiah, Vol. 20. Ed. 2. DOI:10.33633/tc.v19i4.4044

[19] Sari, Erna DH., Irhamah. (2019). *Analisis Sentimen Nasabah Pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner, Naïve Bayes Classifier (NBC), dan Support Vector Machine*

e-ISSN: 2614-8404
p-ISSN: 2776-3234

(SVM). Jurnal Sains dan Seni. Vol. 8. No. 2. ISSN.
2337-3520. Institut Teknologi Sepuluh Nopember.