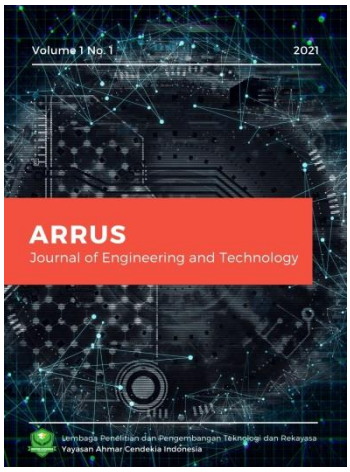
https://doi.org/10.35877/jetech001

OPEN ACCESS

ISSN : 2776-7914 (Print) / 2807-3045 (Online)



*Corresponding author: Zulkifli Rais, Department of Statistics, Universitas Negeri Makassar.

E-mail: zulkifli_rais89@unm.ac.id

RESEARCH ARTICLE

The Support Vector Machine (SVM) And Random Forest Methods For Classification Graduation Rate

Lili Magfirah Rahma Sudirman, Ruliana, Zulkifli Rais*

Statistics Department, Universitas Negeri Makassar, Indonesia

Abstract: Efforts towards an independent nation with high competitiveness can't be separated from educational programs. Therefore, education must be able to produce quality graduates who have knowledge, master technology, and have technical skills, and adequate life skills. The timeliness of students in completing their studies is one of the supports for assessing the quality of higher education. Classification analysis can be used to predict whether a student is said to pass on time or not. Support Vector Machine (SVM) and Random Forest methods are part of the classification method. SVM and Random Forest classification analysis is done by using historical data alumni from FMIPA UNM of the graduation year 2019-2020 which come from the Administration, Academic and Student Affair Bureau of UNM. SVM accuracy level of RBF kernel with optimum value C = 1 and gamma = 1 is 68% and Random Forest accuracy with optimum value m = 2 and k = 500 is 72%. Therefore, the best method for determining the accuracy of the study duration of FMIPA UNM students is Random Forest..

Keywords: Education, Collage, Classification, SVM, Random Forest

1. INTRODUCTION

The main section.

Along with the development of science and technology which is increasing rapidly, it is necessary to have quality human resources and various competencies (Saepuzaman & Retnawati, 2021). Efforts towards an independent and highly competitive nation cannot be separated from education programs. Therefore, education must be able to produce quality graduates who have knowledge, master technology, and have technical skills, and adequate life skills (Adelina, 2018). Every university has an obligation to control the learning achievement of each student and produce quality graduates (Fajrila, 2018).

Makassar State University (UNM) is one of the best public universities in Indonesia. As a fairly old university in Indonesia, UNM has an A accreditation and has produced many alumni from various regions and backgrounds. UNM has 11 faculties, one of which is the Faculty of Mathematics and Natural Sciences (FMIPA). To continue to improve its quality, of course, UNM must also consider the aspect of timeliness of student studies. The length of student study can of course occur due to many things or factors including social, economic, academic, and others. To find out the factors that significantly influence the length of student study, it is necessary to carry out statistical analysis.

Data mining is a semi-automatic process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify potentially useful and useful knowledge information stored in large databases (Turban, 2005). One of the data mining techniques is classification(Fajrila, 2018). According to Han (2001), classification is

https://doi.org/10.35877/jetech001

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

the process of finding a collection of data patterns with one another to be used to predict data that does not yet have a certain data class. Support Vector Machine (SVM) and Random Forest are part of the classification method.

Basically, SVM is a classification for linear data. However, if there is non-linear data then it is processed using the best choice of SVM kernel (Setyorini, 2020). According to Gomes, Prudêncio, Soares, Rossi, and Carvalho (2012) the performance of SVM is highly dependent on the selection of the kernel. Selection of the right kernel function in SVM is very important because this kernel function will determine the feature space in which the classifier function will be searched (Naufal et al. 2015). In this study, the RBF kernel is used because it can handle linear separation of high-dimensional nonlinear input data (Ahmed et al., 2017).

Random forest is a combined tree method derived from the development of the Classification and Regression Tree (CART) method, namely by applying the bootstrap aggregating (bagging) and random feature selection methods (Breiman, 2001). Random Forest will create a large number of classification trees and then combine them to find a high level of accuracy (Sartono and Syafitri, 2010).

Based on the circumstances, this study classifies the graduation rate Faculty of Mathematics and Natural Science Makassar State University using Support Vector Machine and Random Forest Methods.

2. Literature Review

According to (Elly Susilowati, 2015) classification is the process of developing a model that classifies an object according to its attributes.

2.1. Support Vector Machine (SVM)

The SVM method is a machine learning method that works on the principle of Structural Risk Minimization (SRM) which aims to find the best hyperplane to separate the two data classes. SVM works by maximizing the margin which is the distance between the two data classes (Pratiwi 2017). The best-separating function or hyperplane is a hyperplane located in the middle of two objects from both classes (Fauziah et al, 2022).

In general, problems in the real-world domain (real-world problems) are rarely linearly separable but are non-linear. To solve the non-linear problem, SVM was modified to include kernel functions. In the case of non-linear optimization, the equation becomes as follows:

$$\hat{a} = \arg\min \frac{1}{2}\sum_{i=1}^{n} a_i, a_j, y_i, y_j K\left(x_i x_j\right) - \sum_{i=1}^{n} a_i$$

$K(x_i x_j)$ is a kernel function used to handle non-linear data.The function of the kernel allows to implementation a model in a higher dimensional space (feature space).

According to Gomes, Prudêncio, Soares, Rossi, and Carvalho (2012) the performance of SVM is highly dependent on the selection of the kernel. The kernel trick provides several conveniences including determining the support vector, so it is enough to know the Kernel function used, and no need to know the shape of the non-linear function (Setyorini, 2020). The kernel function formula is found in Table 2.1 (Prasetyo, 2012)

Table 1 Kernel type

https://doi.org/10.35877/jetech001

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

Kernel Name	Function Definition
Linear	$K(x, y) = x, y$
Polynomial	$K(x, y) = (x, y + c)^d$
GaussianRBF	$K(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right)$
Sigmoid	$K(x, y) = \tan(\sigma(x, y) + c)$

Selection of the right Kernel function is very important because it will determine the feature space, where the classifier function will be searched.). According to Scholkopf and Smola (1997), the RBF Kernel function has the advantage that it automatically determines the value of an infinite range. The RBF kernel also effectively avoids overfitting by selecting the correct values for the C and parametersand a good RBF is used when there is no prior knowledge.

2.2. Random Forest

Random Forest is one of the ensemble methods to increase the accuracy of a data classification from an unstable single disaggregator through a combination of multiple sorters from the same method as the voting process to obtain a final classification prediction (Wezel & Potharst, 2007). The advantages of Random Forests include being able to produce lower errors, giving good results in classification, being able to handle very large amounts of training data efficiently, and being an effective method for estimating missing data (Breiman, 2001).

The sample size of the explanatory variable (m) when using the random forest method greatly affects the correlation and strength of each tree. To determine m, namely the **number** of predictor variables taken at random with a p-value of many independent variables (Breiman & Cutler, 2003). According to Breiman (2001), the right use of m will produce a random forest with a fairly small correlation between trees but the strength of each tree is large enough as indicated by the acquisition of a small OOB error.

2.3. Accuration

Counting accuracy aims to determine how much accuracy the data is categorized based on the confusion matrix (Suniantara dkk., 2020). Confusion matrix serves to provide an assessment of the classification ability based on true and false (Rahayuningsih, 2019). Table 1 will describe in general the confusion matrix.

Table 2 Confusion Matrix

	Positive	**Negative**
Actual Positive	True Positif (TP)	False negative (FN)
Actual Negative	False Positif (FP)	True Negative (TN)

3. Research Method and Materials

1. The steps taken in this study are as follows :
2. The first stage is data collection, this data collection is then formed a dataset that will be used in this study.
3. This descriptive analysis was carried out on each variable using tables or graphs.
4. Furthermore, data standardization is carried out.
5. In the classification of data types, it is generally divided into two, namely training data and testing data.
6. Training data is the training data used, the training data in this study used is 60% of the total data
7. Data testing is data that is used to test the level of accuracy of the classification method. Testing data used in the study was 40%.
8. The training data is used to perform SVM classification analysis for the RBF kernel. In performing the analysis using the RBF kernel function, the Cost (C) and Gamma (γ)

https://doi.org/10.35877/jetech001

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

parameters were optimized. The SVM kernel RBF process is to classify the SVM on the training data using the optimum C and gamma parameter values in order to produce the best analytical model.

9. Random Forest Classification also uses training data and determines the number of trees formed (k or ntree) and determines the number of predictor variables (m) whose size is determined by the researcher based on previous research. The next step is to optimize the parameter values with a turning process to determine the optimum parameter values. The next step in the Random Forest process is to classify the training data using the optimum k and m values to produce the best analytical model
10. The next stage after obtaining the best model for the classification of SVM and Random Forest on the training data, the classification process will be carried out with data testing to test the accuracy level of each method.
11. Furthermore, after the process for data testing is carried out, it will produce a confusion matrix table for data testing, which can show the level of accuracy of each classification method.
12. Determine the best method based on the highest accuracy value of the two methods

4. Results and Discussion

4.1. Results

The following is a presentation of the length of study for FMIPA UNM students who graduated 2019-2021 in Figure 4.1.

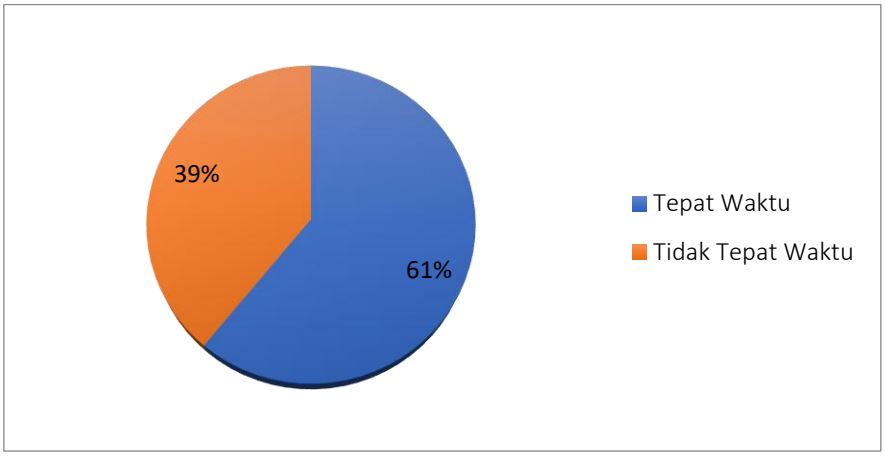


Figure 1 Percentage of Student Graduation Timeliness

Based on Figure 1 above, it can be seen that the percentage of FMIPA UNM students who graduated on time was 61%, while the percentage of FMIPA students who did not graduate on time was 39%. From this percentage, it can be seen that there are more FMIPA UNM students who graduate on time.

4.1.1. Support Vector Machine

The alumni data used is data consisting of various categorical and numerical variables, and the data is not linear, so the SVM method used in this study is to use the Kernel Radian Basis Function (RBF). In performing the analysis with the RBF kernel function, optimization of the Cost (C) and Gamma ( ) parameters was carried out. Therefore, the researchers here used a test of 5 C values, namely 0.1; 1; 5; 10 and 50 as well as 4 gamma value trials, namely 1, 2, 3, and 4. In determining the best parameters in the RBF kernel, trial and error was also carried out so that the following Table 4.2 will be obtained:γ

https://doi.org/10.35877/jetech001

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

Table 3. Cost and Gamma . error values

Cost	Gamma	Error	Dispersion
0.1	1	0.4076530	0.049156095
1.0	1	0.3963587*	0.04477515
5.0	1	0.4104391	0.03453860
10.0	1	0.4104303	0.03794540
50.0	1	0.4094869	0.04186257
0.1	2	0.4076530	0.04156095
1.0	2	0.4245371	0.04146833
5.0	2	0.4254981	0.04807815
10.0	2	0.4254981	0.04807815
50.0	2	0.4254981	0.04807815
0.1	3	0.4076530	0.04156095
1.0	3	0.4301887	0.04390396
5.0	3	0.4292453	0.04524613
10.0	3	0.4292453	0.04524613
50.0	3	0.4292453	0.04524613
0.1	4	0.4076530	0.04156095
1.0	4	0.4339270	0.04649745
5.0	4	0.4301887	0.04390396
10.0	4	0.4301887	0.04390396
50.0	4	0.4301887	0.04390396

*error smallest

Table 3 shows that the SVM method using the RBF kernel produces the smallest error values gamma for the parameters cost = 1 and gamma = 1, so that the cost and gamma values error used for the parameter values error in the RBF kernel are those that produce the smallest errors, namely cost = 1 and gamma. =1.

The support Vector is the data point closest to the hyperplane of each class.

Below are the prediction results by testing the dataset

Table 4. Testing data prediction results

Prediction	Testing Data	
	On time	Not on time
On time	393	177
Not on time	59	83

Based on Table 5 the number of support vectors is 978 data contained in Parameter cost = 1 and gamma = 1. This means that there are 992 data located near each hyperplane group, the number of students graduating on time and students graduating not on time.

4.1.2. Random Forest

Classification analysis using the Random Forest method by determining how many trees are formed (Ntree) and determining how many random samples are taken for each experiment (Mtry). The Mtry value or explanatory variable and the entree value are used to obtain an optimal model and a small OOB error value. The following is a graph of the OOB error value according to the previously searched Mtry value.

https://doi.org/10.35877/jetech001

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

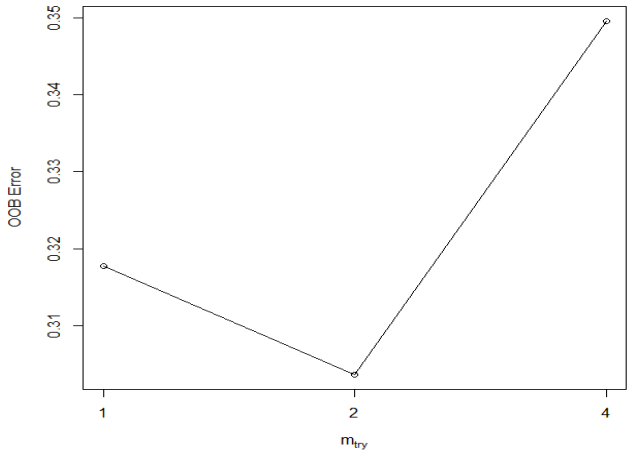


Figure 2. Graph of OOB . error values

Based on Figure 2, it can be seen that the lowest OOB error value was obtained when Mtry was 2, so the researcher decided to use Mtry = 2. The following is the error value generated by each number of trees with an Mtry value of 2.

Table 6. Testing the OOB Error value with different Ntree

Ntree	OOB Error
25	32.43%
50	32.33%
100	31.77%
500*	30.65%*
1000	32.5%

After testing with different numbers of trees (Mtry), the lowest OOB error value was obtained when the number of trees (Ntree) was 500. The lowest OOB error value when the number of trees (Ntree) was 500 was 30.65%. Therefore, it can be used to see the model and also the accuracy of the random forest data method for alumni of the Faculty of Mathematics and Natural Sciences UNM in 2019-2021.

Table 7. Prediction Results with Testing Data

	On time	Not on time
Appropriate	375	112
Not on time	77	148

Based on Table 7, it is known that the predictions of students who graduated on time were 487 students of which there were 375 students with correct predictions and 112 wrong predictions. Meanwhile, the number of students who graduated not on time was 225 students where there were 148 students with correct predictions and 77 students with incorrect predictions.

4.2. Discussion

Based on the SVM and Random Forest methods, the results of the 2019-2021 FMIPA UNM student graduation classification results will be determined. The two classification results from the SVM and Random Forest methods will determine the best method with the most accurate accuracy value. The following is a comparison of the accuracy of the two methods used

https://doi.org/10.35877/jetech001

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

Table 8. Comparison of SVM and Random Forest methods

Method	Accuracy(%)
SVM	67%
Random Forest	73%

Based on Table 8, it can be said that the best method in classifying the length of study for students of the Faculty of Mathematics and Natural Sciences, Makassar State University in 2019-2021 is using the Random Forest method with an accuracy rate of 73%.

5. Conclusion

Based on the results of research and discussions that have been carried out to classify the length of study for students of the Faculty of Mathematics and Natural Sciences, Makassar State University with the SVM and Random Forest methods, it is concluded that the SVM method using parameters cost = 1 and gamma = 1 produces an accuracy of 0.67, while Random Forest method using m=2 and k=500 produces an accuracy value of 0.73. This study resulted in the best method for analyzing the accuracy of the study duration for the students of the Faculty of Mathematics and Natural Sciences, Makassar State University in 2019-2021, which was the Random Forest method.

References

Adelina S. (2018). Factors Causing the Length of Students Completing Thesis at the Faculty of Economics, Padang State University. EcoGen. 1(1). 184-196.

Ahmed, E., Sazzad, MAU, Islam, MT, Azad, M., Islam, S., & Ali, MH (2017, March). Challenges, comparative analysis and a proposed methodology to predict sentiment from movie reviews using machine learning. International Journal of Pure and Applied Mathematics. 118(20). 3277-3284.

Breiman, L. (2001). Random Forest Machine Learning. Netherlands: Kluwer Academic Publisher.

Elly, S., Mira, KS & Alfian, AG (2015). Implementation of the Support Vector Machine Method to Classify Traffic Congestion on Twitter. Makassar. e-Proceeding of Engineering. 2(1). 1478.

Fajrila, Erena. (2018). Comparison of Classification Timeliness of Student Graduation Using Binary Logistic Regression and Naïve Bayes Classifier. Thesis. Statistics Study Program, Faculty of Mathematics and Natural Sciences, Islamic University of Indonesia.

Fauziah., Tiro. MA, Ruliana. (2022).Comparison of k-Nearest Neighbor (k

-NN) and Support Vector Machine (SVM) Methods for Classification of Poverty Data in Papua.ARRUS Journal of Mathematics and Applied Science. 2(2). 83-91.

Gomes, TAF, Prudêncio, RBC, Soares, C., Rossi, ALD, & Carvalho, A. (2012). Combining meta-learning and search techniques to select parameters for support vector machines. Neurocomputing, 75(1), 3-13.

Han, KJ (2001). Data Mining: Concepts and Techniques. San Francisco : John Wiley & Sons Inc

Hanifa, T.T., Adiwijayah., Al-Faraby, S. (2017). Churn Prediction Analysis on Customer Data PT . Telecommunication with Logistic Regression and Underbagging.e-Proceeding of Engineering.4(2). 3210–3225.

Naufal, AR, Wahono, RS, & Syukur, A. (2015). The application of bootstrapping for class imbalance and weighted information gain for feature selection on the support vector machine algorithm for predicting customer loyalty. Journal of Intelligent Systems. 1(2), 98-108.

Pratiwi, YR (2017). Comparison of Sentiment Analysis on Pertalite through the Twitter Social Network using Support Vector Machine and Maximum Entropy Methods. Skippy. Faculty of Mathematics and Natural Sciences, Islamic University of Indonesia. Yogyakarta.

https://doi.org/10.35877/jetech001

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

Rahayuningsih, P. A. (2019). Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Tingkat Kematian Dini Kanker dengan Dataset Early Death Cancer. Jurnal Teknik Informatika Kaputama (JTIK), 3(2).

Republic of Indonesia. (2003). Law of the Republic of Indonesia Number 20 article 19 paragraph 1. Jakarta: State Secretariat.

Saepuzaman D, Retnawati H. (2021). Discriminant Analysis and Classification of the Accuracy of Study Period for Physics Education Students. Journal of Physics Education. 9(2). 92-102.

Sartono, B., & Syafitri, UD (2010). Combined Tree Method: Optional Solution To Overcome Regression Tree Weaknesses. Statistics and Computing Forum, 1-7.

Setyorin, EAM (2020). Comparative Analysis of Machine Learning Methods: Random Forest and SVM for Lung Cancer Detection. Essay. Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Jember.

Suniantara, I. K. P., Suwardika, G., & Soraya, S. (2020). Peningkatan Akurasi Klasifikasi Ketidaktepatan Waktu Kelulusan Mahasiswa Menggunakan Metode Boosting Neural Network. Jurnal Varian, 3(2), 95–102. https://doi.org/10.30812/varian.v3i2.651

Turban, Ephraim. (2005). Decision Support Systems and Intelligent Systems Indonesian Edition Volume 1. Andi: Yogyakarta.

Wezel, M. V & Potharst, R. (2007). European Journal of Operational Research 2007. Vol.181. Issue 1, p. 436-452.