JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 21, No. 1, September 2024, pp. 205-214
DOI: 10.20956/j.v21i1.36257

Classification of Unisba Students' Graduation Time using Support Vector Machine Optimized with Grid Search Algorithm

Ilham Faishal Mahdy[1], Muthia Nadhira Faladiba[2], Nur Azizah Komara Rifai[3], Indah Siti Rahmawati[4], Andhika Sidiq Firmansyah[5]

[1,2,3,4,5]*Department of Statistics, Universitas Islam Bandung*
Email: *ilham.faishal@unisba.ac.id[1], muthia.nadhira@unisba.ac.id[2], nur.azizah@unisba.ac.id[3], indahrahma387@gmail.com[4], kasid630@gmail.com[5]*

Abstract

Support Vector Machine is a classification method that finds the optimal hyperplane to separate two data classes. SVM has much better generalization performance than other methods. However, SVM needs to improve in determining hyperparameter values. Therefore, parameter optimization is necessary to determine the optimal hyperparameter value. Grid search is one of the parameter optimization methods that can improve the quality of SVM models. This study aims to assess the level of accuracy in predicting student graduation times by using five features that affect it. This study shows that the resulting SVM model optimized with the Grid Search Algorithm is quite consistent and prevents overfitting. By utilizing the results of SVM modelling, UNISBA is expected to improve the quality of graduates. The risk of delays in graduation can be considered early by paying attention to the background and achievements of students.

Keywords: Classification, Graduation Time, Grid Search, Optimization; Support Vector Machine

1. INTRODUCTION

Students have a position as the nation's next generation in the future. Students must be able to become pioneers in society, provide changes that positively impact people's lives and instil positive values in society. In other words, students can be referred to as agents of change [12]. In addition, choosing a career is also considered important for students. In college, this is reflected in choosing a primary or skill set designed for a specific type of job in a particular field. While studying at university, students are taught skills that are taught during the teaching and learning process in the classroom and during practicum. This is expected to provide an overview and give the students soft skills in their respective fields.

Universitas Islam Bandung (UNISBA), established in 1958, is one of the favourite Islamic universities in West Java Province. Currently, the university has 10 Faculties and 19 Undergraduate

Study Programs. The following is the number of students in UNISBA based on the Undergraduate Study Program registered in PDDikti in the Odd Semester of 2023.

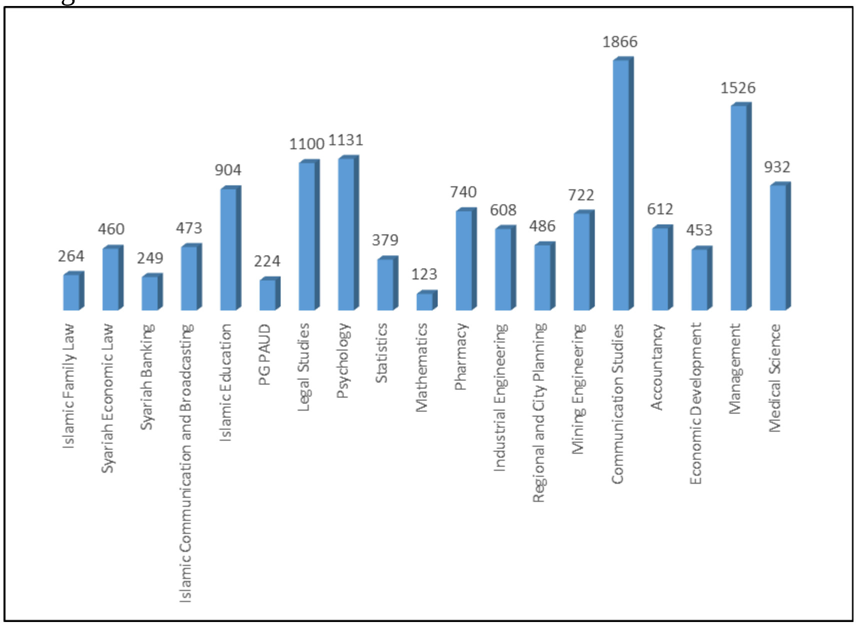

Figure 1.1. Number of Students Based on Undergraduate Study Programs at UNISBA

As seen in Figure 1.1, the Communication Studies Study Program has the highest number of students at UNISBA, 1,866. Meanwhile, the Mathematics Study Program has the lowest number of students, 123.

The punctuality of a student's graduation has varying criteria for each program offered in the college. Based on the academic regulations of UNISBA [20], undergraduate program students are declared to have graduated if they have taken a study load of at least 144 credits with a cumulative grade point average (GPA) greater than or equal to 2.00, a minimum study time of 7 semesters, and a maximum of 14 semesters, as well as research seminars published in Unisba, proceedings, or journals repositories. However, in the academic environment of higher education, there are still problems regarding students who need help to complete their studies within the specified time frame. To solve this problem, a method or technique must be used to group students so that the graduation time can be predicted. One of the methods that can be used is Support Vector Machine.

Support Vector Machine (SVM) was first introduced by Vapnik, Bernhard Boser, and Isabelle Guyon. The main goal of SVM is to find a hyperplane that maximizes the distance to the nearest data. This distance is called the margin. The larger the margin, the less likely it is to produce errors [14]. SVM is a classification method that offers advantages in generalization, Curse of Dimensionality, and Feasibility. With these advantages, SVM can classify data that does not pass the algorithm training stage and produce a classification model even though the training data is relatively small [5].

Several studies have been conducted to Create a Student Graduation Time Classification Model. Zeniarja et. al. [21] applied the CRISP-DM method using 12 attributes and the Random Forest algorithm to classify student graduation status with an accuracy of 77.35%. In this study, only students from Dian University were involved. Suhada et. al. [17] applied the CRISP-DM method through algorithms C4.5. The study results show that GPA, Gender, and Regional Origin can be considered in predicting student graduation with a model accuracy of 85.83%.

Some factors that can affect student graduation, especially at the Islamic University of Bandung, are as follows: 1) Gender, as contained in the results of the study by Agwil et. al. [2] and Triyaningsih et. al. [19]; 2) Regional origin, as contained in the research results by Ariani et. al. [3] and Suhada et. al. [17]; 3) Pesantren scores, which are required for UNISBA students to follow as a condition for participating in Undergraduate Conferences. 4) TOEFL scores, which are required for UNISBA students to follow as a condition for participating in Undergraduate Conferences; and 5) GPA, as stated in the results of the study by Suhada et. al. [17]. Armed with this information, it is hoped that it can be used as a reference in the learning strategy made by the University and Study Programs to increase student graduation rates.

2. MATERIALS AND METHOD

2.1. Materials

The data used in this study were collected from a database owned by the Information and Technology System Development Section (PSITEK) UNISBA. The data collected consists of 2346 graduates who graduated in the 2022-2023 Academic Year, starting in September 2022 and ending in August 2023. Table 2.1 shows the list of variables that have gone through the encoding process.

Table 2.1. Variables

Variables	Categories
Graduation Time	On Time Over Time
Gender	1: Female 2: Male
Regional Origin	1: Bandung Area 2: Outside Bandung Area
TOEFL Scores	1: A1 2: B1 3: B2
Pesantren Scores	1: B 2: B+ 3: A- 4: A
Latin Honors Grades	1: No Predicate 2: Satisfactory 3: Very Satisfactory 4: With Praise

2.2. Data Preprocessing

Data preprocessing is the process of converting raw data into a format that is easier to understand. Data preprocessing must be done because not all data attributes can be used. This process is carried out so that the data used is what is needed [13]. Some of the things that can be done in data preprocessing are data cleaning and encoding.

During data cleaning process, raw data obtained must be re-selected, discarding or eliminating incomplete, irrelevant, and inaccurate data. Two things need to be considered are ensuring that the data collected does not contain missing values and all data is needed during the data analysis process.

In this study, the encoding technique to be used is label encoding. The encoding label converts each value in a column into a different numeric value. Each category value is consecutively labelled as an integer [7]. Categorical data must first be converted into numerical data because machine learning including Support Vector Machine generally cannot handle categorical data.

2.3. Support Vector Machine

Support Vector Machine (SVM) is a machine learning technique that uses a hyperplane to separate data into two different classes. A hyperplane is a decision boundary line used to separate two data classes. In simple terms, SVM involves finding the optimal hyperplane. Maximizing the distance between the hyperplane and the nearest pattern of each class will result in an optimal hyperplane [1]. This distance is known as the margin. The larger the margin, the less likely it is to produce generalization errors. SVM is a linear classification method that has also been developed for non-linear problems.

Linear SVM classification is the application of SVM algorithms to linearly separable data. If a dataset can be separated into two classes using a straight line, the data is said to be linearly separable [18]. As in Figure 2.1, which is divided into two classes, the data with negative signs is the first class, and the data with positive signs is the second class. The data are separated linearly (linear separable) by a dividing line (γ).

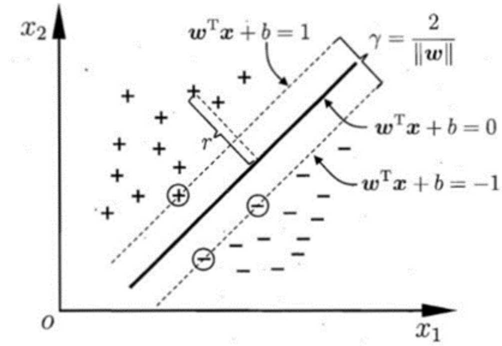

Figure 2.1. Illustration of SVM Classification [10]

The optimal hyperplane in SVM maximizes generalization capabilities. However, when the training data cannot be separated linearly, the resulting classification may not have reached the maximum generalization ability, even though an optimal hyperplane has been found. Therefore, to overcome this problem, SVM is modified by adding kernel functions. If the data cannot be separated linearly, then the hyperplane is formulated as follows:

$$\boldsymbol{f}(x) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b \tag{2.1}$$

with

$\boldsymbol{w}$ = weight vector

$\phi(\boldsymbol{x})$ = mapping function for x

b = bias

The kernel is used to map data onto higher dimensions in nonlinear classifications, where data cannot be linearly classified [9]. The kernel functions can be formulated as follows:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) \tag{2.2}$$

An optimal hyperplane can be achieved by minimizing $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{M} \xi_i$ where ξ_i is variable slack and C is a penalty parameter [15]. Using the kernel function, the lagrange equation as a dual problem is obtained with the following formula:

$$max_\alpha L_D = \sum_{i=1}^{M} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{2.3}$$

with

α_i = lagrange multiplier

y_i = target variable

Equation (2.3) must meet the conditions $0 \le \alpha_i \le C$ and $\sum_{i=1}^{M} \alpha_i y_i = 0$ to obtain the optimal lagrange multiplier value. In order to make class predictions on data, the equation used is as follows [6]:

$$G(x) = \sum_{i=1}^{S} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \tag{2.4}$$

where

$$\boldsymbol{x} \in \begin{cases} \text{Class 1, if } G(\boldsymbol{x}) > 0 \\ \text{Class 2, if } G(\boldsymbol{x}) < 0 \end{cases} \tag{2.5}$$

Several kernel functions that can be used to solve non-linear problems in SVM are as follows:

1. Linear

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{2.6}$$

2. Polynomial

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + 1)^d \tag{2.7}$$

3. Gaussian/Radial Basis Function (RBF)

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \parallel \boldsymbol{x}_i - \boldsymbol{x}_j \parallel^2) \tag{2.8}$$

In this study, we used RBF Kernel in the analysis stage.

2.4. Grid Search Optimization

Grid search is an optimization technique that helps determine the optimum values by comprehensively searching a specific model and selected parameter values. Grid search is simple in terms of implementation, but the experience is required as to which parameter values [8]. The grid search is originally an exhaustive search based on a defined subset of the hyperparameter space. The hyperparameters are specified using a minimal value (lower bound), a maximal value (upper bound), and a number of steps. Three different scales can be used: linear, quadratic, and logarithmic. The performance of every combination is evaluated using some performance metrics. The grid search algorithm is explained as follows [16]:

1. Define the hyperparameters as well as their potential values or ranges.
2. Make a grid with all conceivable hyperparameter combinations.
3. For each hyperparameter combination in the grid:
   a. Train the model on the training set using the current hyperparameters.
   b. Using a performance metric, evaluate the model on the validation or cross-validation set (CV = 5).
   c. Keep track of the performance statistic.
4. Choose the hyperparameter combination that produced the best performance measure.

Grid search optimizes the SVM parameters using a cross-validation technique as a performance metric. The goal is to identify good hyper-parameter combinations so that the classifier can predict unknown data accurately. According to Lin et. al. [11], the cross-validation technique can prevent the overfitting problem. One of the biggest problems of SVM parameter optimization is that there are no exact ranges of C and values. The wider the parameter range is, the more possibilities the grid search method has of finding the best combination parameter. Therefore, in this study, we decided to create a range of C and γ from 0.5 to 5 with a difference of 0.5.

2.5. Classification Model Evaluation

The classification process seeks to categorize all datasets based on their actual classes accurately, but in reality, classification performance cannot be considered perfect. Therefore, evaluation is needed. Accuracy is one of the evaluation criteria for the classification process. The accuracy value of the classification results can be calculated using a confusion matrix. The confusion matrix is a table that displays classifier performance information [4]. The confusion matrix is presented as follows:

Table 2.2. Confusion Matrix

Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

where True Positive (TP) is the amount of data in the model that is predicted to be positive and correctly classified; True Negative (TN) is the amount of data in the model that is predicted to be negative and correctly classified; False Positive (FP) is the amount of data in the model that is predicted to be positive but classified incorrectly; and False Negative (FN) is the amount of data in the model that is predicted to be negative and classified incorrectly. The confusion matrix calculates accuracy, recall, precision and F1-score.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{2.9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.11}$$

$$\text{F1-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.12}$$

3. RESULTS AND DISCUSSION

3.1. Data Exploration

Before further analysis is performed, data exploration is required to see the characteristics of the data. As we explained earlier, the number of students who graduated in the 2022-2023 Academic Year was 2346, consisting of 1383 (58.95%) who successfully graduated on time and 963 (41.05%) who graduated exceeded the required time limit. If we compare by gender, female students have a better chance of graduating on time than male students. If we look at the regional origin, students who live in the Bandung area and outside the Bandung area have a similar chance of graduating on time. Furthermore, students who are most likely not to graduate on time get a TOEFL score equivalent to A1, a pesantren score equivalent to B or do not have a graduation predicate. The categories of variables used along with the classes are shown in the Figure 3.1.

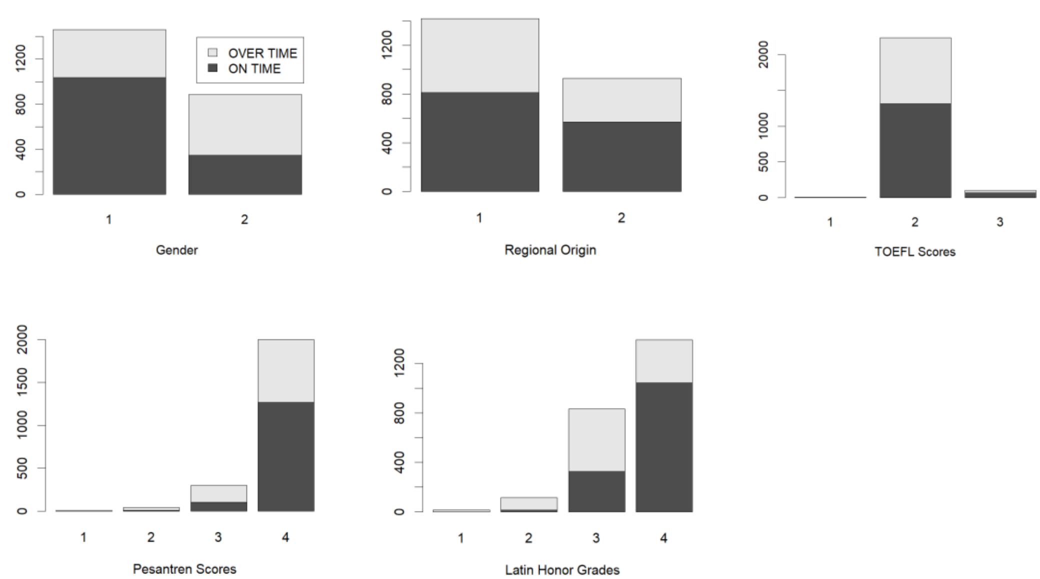

Figure 3.1. Variable Categories Based on Graduation Time

3.2. SVM Optimized with Grid Search Algorithm

In this study, we applied parameter optimization of SVM using grid search with 10-fold cross validation. Grid search is used to optimize SVM parameters in the RBF Kernel within the range of 0.5 to 5 with a difference of 0.5. The SVM parameters optimization using Grid Search experimental results are shown in Table 3.1.

Table 3.1. SVM Best Parameter using Grid Search

C	γ	Error
0.5	1	0.2618551

Table 3.1 shows that the best parameters for an SVM model with an RBF kernel are 0.5 for C and 1 for γ, which delivered the lowest error value compared to other combinations. The best C and γ parameters were then input into the SVM model for classification.

SVM can be solved with the help of a Lagrange multiplier by substituting data on the Lagrange multiplier equation using the RBF kernel function with the value of C and γ parameters that has been obtained. The value of each α can be determined with the condition $0 \leq \alpha_i \leq C$. In the case studied, 443 numbers support vectors were obtained that had α values that met the requirements to be substituted into the classification equation. The value of b obtained is 0.1087349, so the SVM model formed using the RBF kernel is as follows:

$$G(x) = \sum_{i=1}^{S} \alpha_i y_i \exp\left(-1 \parallel \boldsymbol{x}_i - \boldsymbol{x}_j \parallel^2\right) + 0.1087349$$

The SVM with the best parameters is then evaluated using a subset of test data. Performance measurements can be seen through the confusion matrix results and the other metrics such as accuracy, recall, precision, and F1-Score.

Table 3.2. Evaluation of SVM Classification with RBF Kernel

Actual	Predicted		Accuracy	Recall	Precision	F1-Score
	On Time	Not On Time				
On Time	370	141	0.736	0.892	0.724	0.799
Not On Time	45	148				

Table 3.2 shows that the number of students who graduated on time and were predicted to graduate on time through the model were 370. In addition, the number of students who did not graduate on time and were predicted not to graduate on time through a model were 148. An accuracy value of 73.6% illustrates that the model is quite good at classifying data. A precision value of 72.4% describes the level of accuracy between the requested data and the model's prediction value. A recall value of 89.2% describes the system's ability to retrieve information. An F1-score of 79.9% indicates that the model or system has a good balance between precision and recall.

Table 3.3. Comparison of Training and Testing Accuracy

Training Data Accuracy	Testing Data Accuracy
74.2%	73.6%

To determine whether there is an overfitting problem, the accuracy of the training data and the testing data can be compared. If training error is much smaller than testing error, the model may be overfitting. Table 3.3 shows that the difference in accuracy between training data and test data is very small. So, if new data is applied to the model, then the accuracy results are quite consistent and do not indicate the occurrence of overfitting. This study in line with Lin et. al. [11] that the Grid Search algorithm can prevent the overfitting problem.

4. CONCLUSION

Performance of SVM can be improved using parameter optimization. By using grid search optimization, the best parameters for C and γ are obtained that produce the smallest errors and it can prevent the overfitting problem. The classification of student graduation time using SVM shows quite good accuracy using five features: Gender, Regional Origin, Pesantren Scores, TOEFL Scores, and Latin Honors Grades. By utilizing the results of SVM modelling, UNISBA is expected to improve the quality of graduates by improving their ability to complete their studies. The risk of delays in graduation can be considered early by paying attention to the background and achievements of students.

ACKNOWLEDGEMENT

The authors thank Universitas Islam Bandung, which provided financial assistance for the dissemination of research through Research and Community Services Grants, and everyone who contributed to this study.

REFERENCES

[1] Abe, S., 2010. *Support Vector Machines for Pattern Classification* (2nd ed.). Springer-Verlag. http://www.springer.com/series/4205

[2] Agwil, W., Fransiska, H., & Hidayati, N., 2020. Analisis Ketepatan Waktu Lulus Mahasiswa Dengan Menggunakan Bagging Cart. *FIBONACCI: Jurnal Pendidikan Matematika Dan Matematika*, *6*(2), 155. https://doi.org/10.24853/fbc.6.2.155-166

[3] Ariani, N. K. D., Sumarjaya, I. W., & Oka, T. B., 2013. Analisis Faktor-Faktor Yang Memengaruhi Waktu Kelulusan Mahasiswa Dengan Menggunakan Metode Gompit (Studi Kasus: Mahasiswa Fakultas MIPA Universitas Udayana). *E-Jurnal Matematika*, *2*(3), 40–45.

[4] Bramer, M., 2007. *Principles of Data Mining*. Springer-Verlag.

[5] Fatmawati, & Rifai, N. A. K., 2023. Klasifikasi Penyakit Diabetes Retinopati Menggunakan Support Vector Machine dengan Algoritma Grid Search Cross-validation. *Jurnal Riset Statistika*, 79–86. https://doi.org/10.29313/jrs.v3i1.1945

[6] Fide, S., Suparti, & Sudarno, 2021. *Analisis Sentimen Ulasan Aplikasi Tiktok di Google Play Menggunakan Support Vector Machine (SVM) dan Asosiasi*. *10*(3), 346–358. https://ejournal3.undip.ac.id/index.php/gaussian/

[7] Gede, I., Sudipa, I., & Darmawiguna, M., 2024. *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia. https://www.researchgate.net/publication/377415198

[8] Kervanci, I. S., Akay, M. F., & Özceylan, E., 2024. Bitcoin Price Prediction using LSTM, GRU And Hybrid LSTM-GRU with Bayesian Optimization, Random Search, and Grid

Search for The Next Days. *Journal of Industrial and Management Optimization*, *20*(2), 570–588. https://doi.org/10.3934/jimo.2023091

[9] Kim, M. C., Lee, J. H., Wang, D. H., & Lee, I. S., 2023. Induction Motor Fault Diagnosis Using Support Vector Machine, Neural Networks, and Boosting Methods. *Sensors*, *23*(5). https://doi.org/10.3390/s23052585

[10] Kusumaningrum, A. P., 2017. *Optimasi Parameter Supprort Vector Machine Menggunakan Genetic Algorithm Untuk Klasifikasi Microarray Data*. Institut Teknologi Sepuluh Nopember.

[11] Lin, S. W., Ying, K. C., Chen, S. C., & Lee, Z. J., 2008. Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications*, *35*(4), 1817–1824. https://doi.org/10.1016/j.eswa.2007.08.088

[12] Martadinata, A. M., 2019. Peran Mahasiswa dalam Pembangunan Di Indonesia. *IDEA : Jurnal Humaniora*, *2*(1), 1–12. https://doi.org/10.29313/idea.v0i0.4168

[13] Rerung, R. R., 2018. Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk. *Jurnal Teknologi Rekayasa*, *3*(1), 89. https://doi.org/10.31544/jtera.v3.i1.2018.89-98

[14] Rustam, Z., & Audia Ariantari, N. P. A., 2018. Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance. *Journal of Physics: Conference Series*, *1028*(1). https://doi.org/10.1088/1742-6596/1028/1/012005

[15] Schölkopf, B., & Smola, A. J., 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Massachusetts Institute of Technology.

[16] Shams, M. Y., Elshewey, A. M., El-kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z., 2024. Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, *83*(12), 35307–35334. https://doi.org/10.1007/s11042-023-16737-4

[17] Suhada, K., Elanda, A., & Aziz, A., 2021. Klasifikasi Predikat Tingkat Kelulusan Mahasiswa Program Studi Teknik Informatika dengan Menggunakan Algoritma C4.5 (Studi Kasus: STMIK Rosma Karawang). *Dirgayama: Jurnal Manajemen Dan Sistem Informasi*, *1*(2), 14–27.

[18] Trisnawati, N. K., & Astawa, I. G. S., 2023. Klasifikasi Penyakit Hepatitis C Menggunakan Algoritma Support Vector Machine. *JNATIA*, *1*(4), 1215–1219.

[19] Triyaningsih, S. L., & Triastity, R., 2015. Perbedaan Ketepatan Kelulusan Mahasiswa ditinjau dari Gender, Indeks Prestasi Kumulatif (IPK) dan Program Studi (Survei pada Mahasiswa Fakultas Ekonomi Unisri Surakarta Tahun Kelulusan Tahun 2012-2013). *EKSPLORASI*, *27*(2).

[20] Universitas Islam Bandung., 2022. *Pedoman Akademik Universitas Islam Bandung Tahun Akademik 2022-2023*.

[21] Zeniarja, J., Salam, A., & Ma'ruf, F. A., 2022. Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa. *Jurnal Rekayasa Elektrika*, *18*(2), 102–108. https://doi.org/10.17529/jre.v18i2.24047