Journal of Applied Science, Engineering, Technology, and Education *ISSN: 2685-0591 (online)*
Vol. 4 No. 2 (2022)
https://doi.org/10.35877/454RI.asci1581

Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education

Rizal Bakri[a,b,*], Niken Probondani Astuti[c], & Ansari Saleh Ahmar[d]

[a]*Statistics Research Group, STIEM Bongaya, Jl. Mappaoudang No. 28, Makassar, 90223, Indonesia*
[b]*Department of Digital Business, Universitas Negeri Makassar, Jl. Pendidikan No. 1, Makassar, 90222, Indonesia*
[c]*Department of Management, STIEM Bongaya, Jl. Mappaoudang No. 28, Makassar, 90223, Indonesia*
[d]*Department of Statistics, Universitas Negeri Makassar, Jl. Dg. Tata Raya, Makassar, 90223, Indonesia*

Abstract

This study aims to predict a student's graduation on time (GOT) using machine learning algorithms. We applied five different machine learning algorithms, namely Random Forest, Support Vector Machine (Linear Kernel), Support Vector Machine (Polynomial Kernel), K-Nearest Neighbors, and Naïve Bayes. These algorithms were tested using 10-fold cross validation and simulated various parameter tuning values. The results show that the Random Forest algorithm produces the best accuracy and kappa statistics values, so this algorithm is suitable for modeling predictive data of students graduating on time. This predictive model is expected to be useful for higher education management in designing their strategies to assist and improve student academic performance weaknesses in order to achieve graduation on time.

Keywords: machine learning algorithms; graduation on time (GOT); parameter tuning.

Received: 17 August 2022 Revised: 29 November 2022 Accepted: 20 December 2022

1. Introduction

The increasing number of universities in Indonesia has made those that have existed for a long time try to provide the best service to increase the number of student admissions and the quality of their graduates. One of the indicators that is the main issue for measuring the quality of a university and has attracted the attention of policymakers, industrial owners, educators, and researchers in recent years is the graduation rate and timeliness of students in achieving a degree (Yue & Fu, 2017). The Indonesian Ministry of Education, through the National Accreditation Agency for Higher Education (BAN-PT), has determined that the standard for on-time graduation is at least 50% of the number of students admitted during that period (Pradipta et al., 2019). If these standards are not achieved, the accreditation of the university will decrease. Besides that, students who take longer to graduate have an impact on the university's budget since the university must spend more money on extra resources, like more classrooms to accommodate more students and lecturer salaries. Therefore, higher education needs to handle this issue brilliantly and proactively, as the university's achievement depends highly on the graduation rate. One of the solutions to handle this issue is by analyzing students' performance, since it can be an indicator to predict the students' graduation time. However, analyzing the performance of students is very complicated and tedious as it involves many data points that are continuously increasing year by year (Ogwoka et al., 2015). Alternatively, data mining can be used to perform analysis in order to solve this problem.

According to Han et al. (2023) data mining is a process of discovering interesting patterns, models, and other kinds of knowledge in large data sets. It is about analyzing and categorizing the data, as well as summarizing the knowledge of many kinds of information collected in databases and data warehouses (Reddy et al., 2010). Data mining with machine learning algorithms is now used extensively in the education system to analyze students' performance, predict when they will graduate, and address other related issues. As research has been done by Suhaimi et al. (2019) about a predictive model of graduation on time using machine learning algorithms. Their research used five machine learning algorithms, namely Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (PolyKernel), and Support

* Corresponding author.
E-mail address: rizal.bakri@stiem-bongaya.ac.id

Vector Machine (RBFKernel) with four different k-folds of 5, 10, 15, and 20. Their study showed that Support Vector Machine (PolyKernel) performed better than other classifiers, and 5 and 20 k-folds were the best for this experiment. Lagman et al. (2020) conducted research on student graduation using machine learning algorithms to improve classification algorithm accuracy for student graduation prediction using an ensemble model. Their research used four machine learning algorithms, namely Logistic Regression, Neural Network, Decision Tree, and Naïve Bayes. The results show that Logistics Regression performed better than other classifiers. Other research uses machine learning in educational data mining about predictive of time graduation have been carry out using two-level classification algorithm by Tampakas et al. (2019), Bassi et al. (2019) have been research of students graduation on time prediction model using artificial neural network, predicting time to graduation at a large enrollment American university using logistic regression and gradient boosted trees by Aiken et al. (2020), and time series prediction on college graduation using KNN algorithm by Salim et al. (2020), and there are many other studies that use machine learning algorithm.

Using machine learning algorithms, one can improve the accuracy of the educational data mining model and conduct a more in-depth analysis of the mined data. Researchers typically employ machine learning algorithms to discover patterns in data sets by allowing them to learn on their own (Asyraf et al., 2017). However, from the various studies previously mentioned, the use of machine learning algorithms by simulating parameters tuning in educational data mining has not been carried out. According to Weerts et al. (2020) the performance of many machine learning algorithms depends on their parameter tuning settings. Therefore, in this research, we apply five different machine learning algorithms, namely Random Forest, Support Vector Machine (Linear Kernel), Support Vector Machine (Polynomial Kernel), K-Nearest Neighbors, and Naïve Bayes using 10-fold cross validation and various parameter tuning techniques.

2. Methodology

This research methodology follows the Cross-Industry Process for Data Mining (CRISP-DM), with six steps associated with the required activities (Suhaimi et al., 2019). The details of each step are discussed briefly:

Business Understanding. Understanding business is the first step of this research. The issues related to students' graduation status and the factors that contribute to their timely graduation will be the primary focus of this step.

Data Understanding. Data understanding is the second step in the CRISP-DM process, and it requires the researcher to acquire the necessary data and transform it into a format that can be mined with the Data Mining Tool. The STIEM Bongaya students' database, which contains the historical data of STIEM Bongaya's students from cohorts 2013 to 2018, was used as the method for data collection in this study. The distribution and its range values were then identified by carefully examining this data. The analysis shows that the raw data from the STIEM Bongaya database has 14 attributes with 4,093 rows of data consisting of two undergraduate student departments, which are management and accounting.

Data Preparation. The research model's competence is highly dependent on the dataset's quality, making this step of the CRISP-DM process crucial. This step includes a number of activities, including data selection and cleaning, data construction, and data integration and formatting. Based on their relevance to the objectives of this research, a number of attributes from the raw dataset were chosen. 14 attributes with GOT status as the target or class label are listed in Table 1 as the list of selected attributes.

Modelling. The modeling step is where we look for useful patterns in the data. To find patterns in datasets throughout the machine learning process, modeling of the dataset is required. There are several modeling algorithms in data mining; however, not all of them are appropriate for this study topic. Before modeling, the data is divided into two parts, namely, 80% for training data and 20% for testing data. After that, the prepared data was trained on five different classifiers with 10-fold cross validation and parameter tuning, as shown in Table 2.

Evaluation. The values of the accuracy score and the Kappa statistic were used to evaluate these classifiers during the model evaluation step. The best classifier chosen for this study is the one with the highest score.

Deployment. All of the research's progress, outcomes, and findings, as well as any issues or constraints, are provided in a report form during the deployment phase. The predictive models are developed using R Studio.

Table 1. Atributes of Graduation on time (GOT)

Atribute ID	Value	Description
NCP	0 - 4	Student's Number Credit Passed
SMT4	0 - 4	Student's GPA Semester 4
SMT3	0 - 4	Student's GPA Semester 3
SMT2	0 - 4	Student's GPA Semester 2
SMT1	0 - 4	Student's GPA Semester 1
AA	16 - 46	Student's Age Admission
FS	Accounting, Financial, Marketing, Human Resource	Student's Focus Study
FI	IDR 1 - IDR 499.999	Student's Father Income
	IDR 500.000 - IDR 999.999	
	IDR 1.000.000 - IDR 1.999.999	
	IDR 2.000.000 - IDR 4.999.999	
	IDR 5.000.000 - IDR 20.000.000	
	More than IDR 20.000.000	
	Nil Income	
MI	IDR 1 - IDR 500.000	Student's Mother Income
	IDR 500.000 - IDR 999.999	
	IDR 1.000.000 - IDR 1.999.999	
	IDR 2.000.000 - IDR 4.999.999	
	IDR 5.000.000 - IDR 20.000.000	
	More than IDR 20.000.000	
	Nil Income	
SEX	Male, Female	Student's Sex
RE	with Parents, with Guardian, Boarding House, Dormitory, Others	Student's Residence
TR	Public transportation,	Student's Transportation
	Private Car,	
	Private Motorcycle	
	Walk to campus	
DEP	Management, Accounting	Department taken by student
CT	Regular Class, Executive Class	Class type taken by student
GOT Status	Yes, No	Student's graduation status

Table 2. Machine Learning Algorithm with Parameter Tuning

Algorithm	Parameter Tuning
Random Forest	mtry = (2,3,4)
Support Vector Machine (Linear Kernel)	cost = (0.01, 0.1, 1)
Support Vector Machine (Polynomial Kernel)	degree = (1,2,3); scale = (0.01, 0.1); cost = (0.25, 0.5, 1)
K-Nearest Neighbors	K = (1, 3, 5, 7, 9)
Naïve Bayes	usekernel = (T, F); adjust = (0.01, 0.1, 1); fl = (0.01, 0.1, 1)

3. Result and Discussion

The performance of five machine learning algorithms with tuning parameters validated using 10-fold cross validation will be discussed in this session, as well as the importance of features related to students' graduation on time (GOT).

3.1. Performance with Parameter Tuning

Data training of students' graduation times on STIEM Bongaya was modeled using five machine learning algorithms, namely Random Forest, Support Vector Machine (Linear Kernel), Support Vector Machine (Polynomial Kernel), K-Nearest Neighbors, and Nave Bayes, with 10-fold cross validation and various parameter tunings as shown in Table 2.

The results of the performance of the five algorithms can be seen in Figure 1. In Figure 1, the accuracy and kappa statistical values are shown in the form of intervals obtained from a 10-fold cross-validation trial that was simulated with different tuning parameter values. The results show that the machine learning algorithm with the best accuracy and Kappa statistic is the Random Forest with a mtry tuning parameter value of 4. Then the next best accuracy and Kappa statistic value is still the random forest algorithm with an mtry of 3.

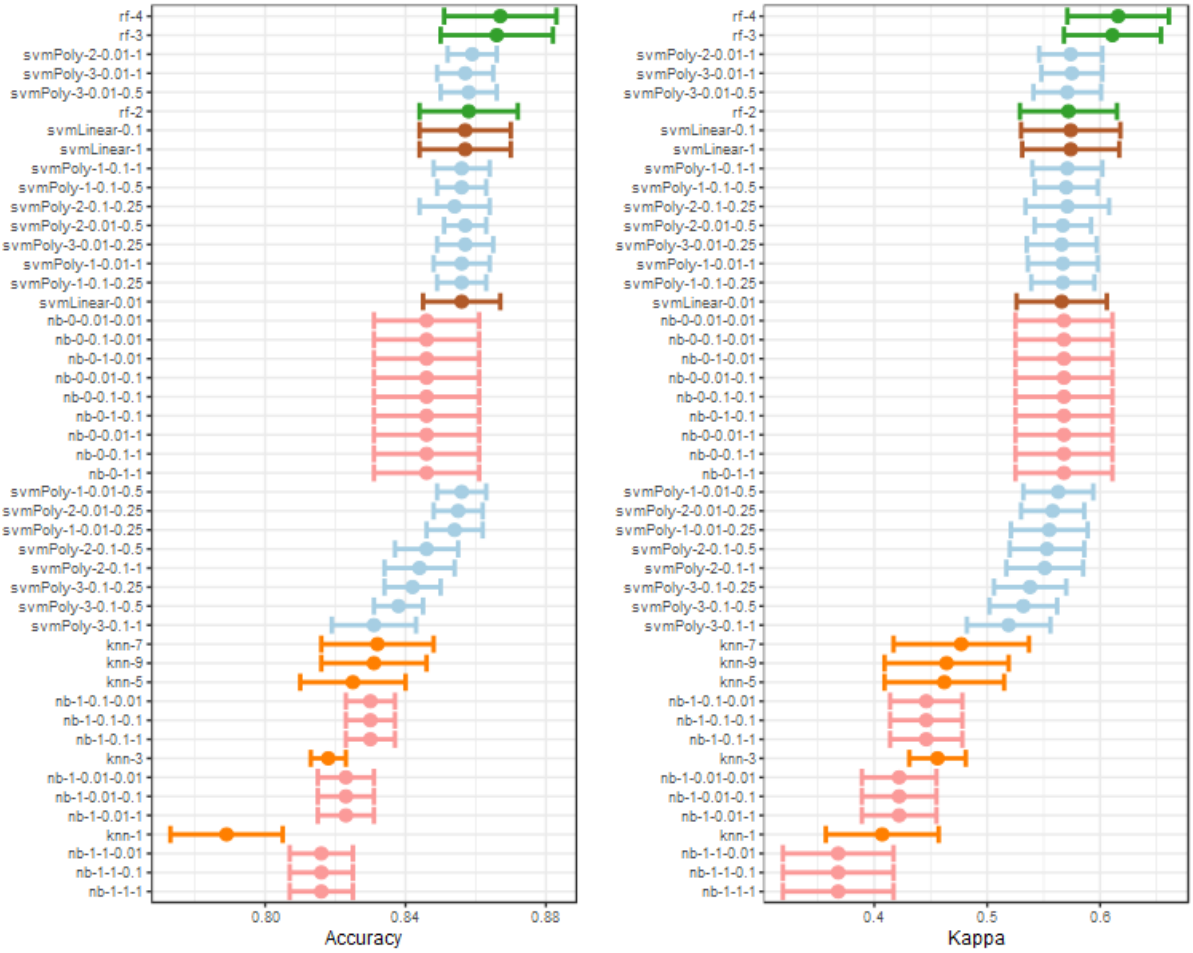

Fig. 1. Accuracy and Kappa statistics with 10-fold Cross-validation and Parameters tuning

This study is similar to previous research by Hutt et al. (2019) on the prediction of on-time graduation, which applies the random forest algorithm and achieves the best prediction performance with ROC values in the range between 0.629 and 0.694. Another study that has been conducted by Gunawan et al. (2021) regarding the prediction of graduation on time states that the random forest algorithm has the highest accuracy value compared to other algorithms, with an accuracy value of 0.726. Recently, this algorithm has produced the best accuracy values from various research fields. Figure 1 also shows the machine-learning algorithm that has the lowest accuracy and Kappa statistic, namely the Naïve Bayes algorithm. However, this algorithm for certain tuning parameter values has the same accuracy value, and some are even better than the accuracy values of the SVM Poly and KNN algorithms. The use of this method in various studies to predict graduation time has been carried out, including research conducted by Lagman et al. (2019), Gerhana et al. (2019), and Sugiharti et al. (2017), which states that the Naïve Bayes algorithm has a high accuracy value. Then in Figure 1, it also shows that the accuracy value for the K-Nearest Neighbors algorithm with various tuning parameter simulations is never better than other algorithms except that it is only better than the Naïve Bayes algorithm for certain parameter values. These results are also in line with research conducted (2019), which states that the accuracy of the K-Nearest Neighbors algorithm is better than the Naïve Bayes algorithm. On the other hand, an interesting machine learning algorithm to discuss in Figure 1 is the Support Vector Machine (Polynomial Kernel) algorithm. In Figure 1, it can be seen that only this algorithm has a small value interval compared to other interval algorithms for all the parameter

tunings that have been tried. The accuracy value produced by this method is more stable and consistent for each experiment, up to 10-fold cross-validation. So there have been several studies on graduation time predictions using this algorithm, and the results show that this method produces a high accuracy value (Ahmad Tarmizi et al., 2019; Kesumawati & Utari, 2018; Suhaimi et al., 2019; Pang et al., 2017). Therefore, each algorithm has advantages and disadvantages for predicting a student's timely graduation (GOT) in higher education. So, based on the simulation results using parameter tuning, the best accuracy value and Kappa statistic produced by each machine learning algorithm are shown in Table 3, which shows that in this study, the Random Forest algorithm produced the best accuracy value and Kappa statistic with a mtry of 4.

Table 3. Performance Statistics for Five Machine Learning Algorithms

Methods	Parameter tuning	Accuracy	Kappa Statistic
Random Forest	mtry = 4	0.867	0.616
SVM Polynomial	Degree = 2; scale = 0.01; cost = 1	0.859	0.574
SVM Linear	cost = 0.1	0.857	0.574
Naïve Bayes	usekernel = F; adjust = 0.01; fl = 0.01	0.856	0.568
K-Nearest Neighbort	K = 7	0.832	0.477

3.2. Feature Importance

In this section, we will discuss the variables that influence the prediction of student graduation on time using the Relative Importance Score method of the Random Forest algorithm.

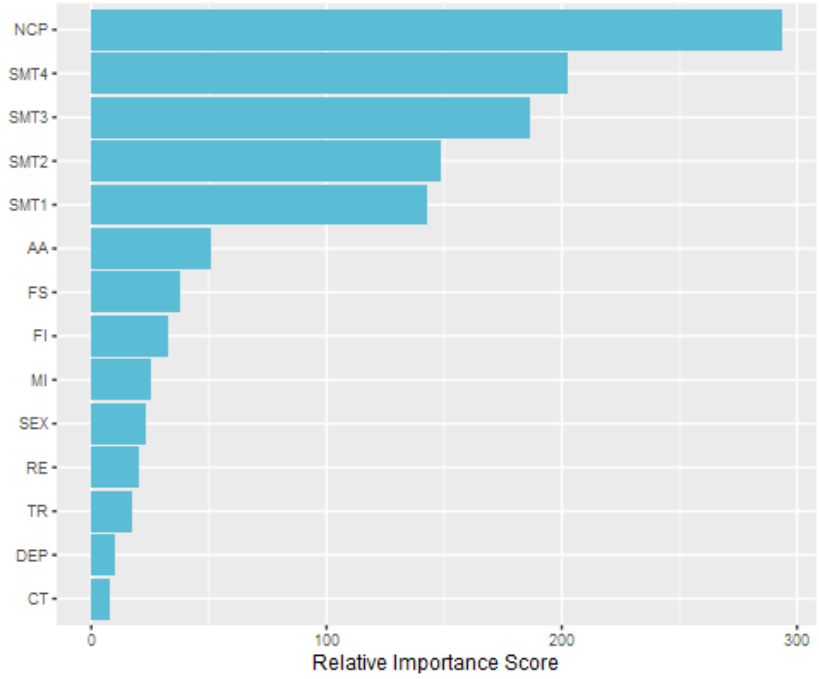

Fig. 2. Feature Importance of Graduation on time (GOT)

Figure 2 shows the relative importance scores of 14 features, which consist of two feature groups, namely, features with a continuous scale and features with a categorical scale. The graph shows that features with a continuous scale have more influence on model formation to predict a student's graduation-on-time (GOT) than features with a categorical scale. On a continuous scale, the following attributes have the highest relative importance scores: number of credits passed (NCP), student's GPA from semester 4 to semester 1 (SMT4-SMT1), and student's age at admission (AA). Then, the attributes in order of highest to lowest relative importance score on the categorical scale are: focus study (FS), father income (FI), mother income (MI), sex (SEX), student's residence (RE), student's transportation (TR), department taken

by the student (DEP), and class type taken by the student (CT). As a result, the higher education institution can develop a student study plan strategy based on this feature importance score in order to increase the student's graduation date.

4. Conclussion and Future Work

The purpose of this research, we apply five different machine learning algorithms, namely Random Forest, Support Vector Machine (Linear Kernel), Support Vector Machine (Polynomial Kernel), K-Nearest Neighbors, and Naïve Bayes using 10-fold cross validation and various parameter tuning values. The results show that the Random Forest algorithm produces the best accuracy and kappa statistics values, so this algorithm is suitable for modeling predictive data of students graduating on time. However, in the future, we will improve the performance of these five machine learning algorithms for dealing with imbalanced data problems in target data using various techniques. This research can inform higher education administrators, students, and academics about students whose performance is most likely to fail graduation on time and the solutions that can be found. Additionally, this strategy has the potential to enhance the academic quality of higher education by significantly reducing the number of students who are unable to graduate on time.

References

Ahmad Tarmizi, S. S., Mutalib, S., Abdul Hamid, N. H., Abdul-Rahman, S., & Md Ab Malik, A. (2019). A Case Study on Student Attrition Prediction in Higher Education Using Data Mining Techniques. *Communications in Computer and Information Science*, *1100*, 181–192. https://doi.org/10.1007/978-981-15-0399-3_15/COVER

Aiken, J. M., de Bin, R., Hjorth-Jensen, M., & Caballero, M. D. (2020). Predicting time to graduation at a large enrollment American university. *PLOS ONE*, *15*(11), e0242334. https://doi.org/10.1371/JOURNAL.PONE.0242334

Asyraf, A. S., Abdul-Rahman, S., & Mutalib, S. (2017). Mining textual terms for stock market prediction analysis using financial news. *Communications in Computer and Information Science*, *788*, 293–305. https://doi.org/10.1007/978-981-10-7242-0_25/COVER

Bassi, S. J., Gbenga Dada, E., Abdulkadir Hamidu, A., Dauda Elijah, M., & Author, C. (2019). *Students Graduation on Time Prediction Model Using Artificial Neural Network*. *21*(3), 28–35. https://doi.org/10.9790/0661-2103012835

Gerhana, Y. A., Fallah, I., Zulfikar, W. B., Maylawati, D. S., & Ramdhani, M. A. (2019). Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period. *Journal of Physics: Conference Series*, *1280*(2), 022022. https://doi.org/10.1088/1742-6596/1280/2/022022

Gunawan, Hanes, & Catherine. (2021). C4.5, K-Nearest Neighbor, Naïve Bayes and Random Forest Algorithms Comparison to Predict Students' On Time Graduation. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *4*(2), 62–71. https://doi.org/10.24014/ijaidm.v4i2.10833

Han, J., Pei, J., & Tong, H. (2023). Data Mining: Concepts and Techniques. In *Morgan Kaufmann*. In Morgan Kaufmann.

Hutt, S., Gardner, M., Duckworth, A. L., & D'Mello, S. K. (2019). Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. *International Educational Data Mining Society*.

Kesumawati, A., & Utari, D. T. (2018). Predicting patterns of student graduation rates using Naïve bayes classifier and support vector machine. *AIP Conference Proceedings*, *2021*(1), 060005. https://doi.org/10.1063/1.5062769

Lagman, A. C., Alfonso, L. P., Goh, M. L. I., Lalata, J. A. P., Magcuyao, J. P. H., & Vicente, H. N. (2020). Classification algorithm accuracy improvement for student graduation prediction using ensemble model. *International Journal of Information and Education Technology*, *10*(10), 723–727. https://doi.org/10.18178/IJIET.2020.10.10.1449

Lagman, A. C., Gonzales, J. G., Ramos, R. F., Calleja, J. Q., Legaspi, J. B., Solomo, M. V. S., Fernando, C. G., Ortega, J. H. J. C., & Santos, R. C. (2019). Embedding naïve bayes algorithm data model in predicting student graduation. *ACM International Conference Proceeding Series*, 51–56. https://doi.org/10.1145/3369555.3369570

Ogwoka, M. T., Wilson Cheruiyot, K., & George Okeyo, K. (2015). A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms. *International Journal of Computer Applications Technology and Research*, *4*(9), 693–697. www.ijcat.com693

Pang, Y., Judd, N., O'Brien, J., & Ben-Avie, M. (2017). Predicting students' graduation outcomes through support vector machines. *Proceedings - Frontiers in Education Conference, FIE*, *2017-October*, 1–8. https://doi.org/10.1109/FIE.2017.8190666

Pradipta, A., Hartama, D., Wanto, A., Saifullah, S., & Jalaluddin, J. (2019). The Application of Data Mining in Determining Timely Graduation Using the C45 Algorithm. *IJISTECH (International Journal of Information System and Technology)*, *3*(1), 31–36. https://doi.org/10.30645/IJISTECH.V3I1.30

Reddy, S. G., Srinivasu, R., Poorna, M., Rao, C., & Rikkula, S. R. (2010). Data Warehousing, Data Mining, OLAP And OLTP Technologies Are Essential Elements To Support Decision-Making Process In Industries. *(IJCSE) International Journal on Computer Science and Engineering*, *02*(09), 2865–2873. http://pwp.starnetinc.com/larryg/articles.html

Salim, A. P., Laksitowening, K. A., & Asror, I. (2020). Time Series Prediction on College Graduation Using KNN Algorithm. *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*. https://doi.org/10.1109/ICOICT49345.2020.9166238

Solichin, A. (2019). Comparison of decision tree, Naïve Bayes and K-nearest neighbors for predicting thesis graduation. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 217–222. https://doi.org/10.23919/EECSI48112.2019.8977081

Sugiharti, E., Firmansyah, S., & Devi, F. R. (2017). Predictive Evaluation of Performance of Computer Science Students of UNNES using Data Mining based on Naïve Bayes Classifier (NBC) Algorithm. *Journal of Theoretical and Applied Information Technology*, *28*(4). www.jatit.org

Suhaimi, M. N., Abdul-Rahman, S., Mutalib, S., Abdul Hamid, N. H., & Md Ab Malik, A. (2019). Predictive Model of Graduate-On-Time Using Machine Learning Algorithms. *Communications in Computer and Information Science*, *1100*, 130–141. https://doi.org/10.1007/978-981-15-0399-3_11/COVER

Tampakas, V., Livieris, I. E., Pintelas, E., Karacapilidis, N., & Pintelas, P. (2019). Prediction of students' Graduation time using a two-level classification algorithm. *Communications in Computer and Information Science*, *993*, 553–565. https://doi.org/10.1007/978-3-030-20954-4_42/COVER

Weerts, H. J. P., Mueller, A. C., & Vanschoren, J. (2020). *Importance of Tuning Hyperparameters of Machine Learning Algorithms*. https://doi.org/10.48550/arxiv.2007.07588

Yue, H., & Fu, X. (2017). Rethinking Graduation and Time to Degree: A Fresh Perspective. *Research in Higher Education*, *58*(2), 184–213. https://doi.org/10.1007/S11162-016-9420-4/METRICS