The Performance of Decision Tree and Ensemble Algorithms for Classifying the Graduation Status of Undergraduate Students at Universitas Negeri Jakarta

Dian Handayani[1], Abdurrahman Malik Karim[2], Dania Siregar[1], Faroh Ladayya[1]

[1] Study Program of Statistics, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta Timur 13220, Indonesia
[2] PT Cyberindo Aditama, Cyber 2 Tower 33rd Floor, Jl. HR Rasuna Said X5 No. 13, Jakarta Selatan, 12950, Indonesia
amalik.karim@gmail.com

Abstract. A bachelor's degree in Indonesia typically takes around four years to complete. This research aims to examine the data patterns related to the graduation status of a bachelor's degree from Universitas Negeri Jakarta (UNJ). The data pattern is used to determine if an undergraduate student from UNJ will graduate on time or not on time. On time graduation occurs when the student takes four years or eight semesters to obtain their bachelor's degree, while not on time graduation occurs if the student takes more than eight semesters. The graduation status is classified using AdaBoost and Random Forest algorithms, based on grade points and total credits earned from the student's courses. The AdaBoost and Random Forest algorithms are contrasted with the simpler ML algorithm, Decision Tree. The study found that decision tree, AdaBoost, and Random Forest are effective in classifying an undergraduate student from UNJ based on their graduation status 'on time' or 'not on time' during a given semester (the first, the second, the third, the fourth, the fifth or the sixth semester). The student's classification accuracy score reaches 64%. The Random Forest, AdaBoost, and Decision Tree all had accuracy scores of 64%, 63%, and 60%, respectively. Ensemble methods (AdaBoost and Random Forest) outperform the decision tree algorithm. For each semester, the mean difference in accuracy score between ensemble methods and decision tree for classifying a student's graduation status reaches 2%-5%.

Keywords: Confusion Matrix, F1 Score, Machine Learning, Permutation Feature Importance.

1 Introduction

The length of study taken by an undergraduate student to graduate with a bachelor degree is one indicator that can reflect the success of his or her studies. According to Permendikbud No. 3 Tahun 2020, the length of study for a bachelor's degree in Indonesia is said to be normal if it takes about four years. However, in reality, many students still have not graduated although their length of study is already more than four years. There are many factors that can affect the graduation status of a student. The status

© The Author(s) 2023
N. Djam'an et al. (eds.), *Proceedings of the 5th International Conference on Statistics, Mathematics, Teaching, and Research 2023 (ICSMTR 2023)*, Advances in Computer Science Research 109,
https://doi.org/10.2991/978-94-6463-332-0_5

could be ‘on time’ or ‘not on time’. In our study, we will define the graduation status for an undergraduate student is ‘on time’ if duration of study to obtain a bachelor's degree is four years. Otherwise, the graduation status will be called ‘not on time’. Cantwell & Moore (1996) stated that students with high GPA exhibit a high level of self-control. Moreover, the students who have higher self-control tend to more concentrate on their studies. As a result, their motivation to finish their studies on time will also increase.

The machine learning method (ML) is a data analysis method for classifying and predicting. ML is a field of artificial intelligence that is developed with the idea that a system can learn from data, discover data patterns, and make some decisions with as little human interaction as possible. ML is also a method that studies computer algorithms in which the program will improve its performance automatically based on data and previous experience. The ML algorithm builds some models based on sample data (training data) to implement some activities such as prediction, classification, or decision without being given some explicit instructions (IBM Cloud Education, 2020; Koza et al., 1996; SAS Institute Inc., 2019).

A decision tree is a simple ML algorithm that separates observed data into some 'branches' like a tree. Chalaris et al. (2015), Kamil & Cholil (2020), and Mayasari (2016) employed a decision tree to classify a student will graduate ‘on time’ or ‘not on time’. The results of their research show that the accuracy score of decision tree is around 60%. Although decision trees are simple to analyze and explain to some people, they also have some limitations. For example, the parameter estimates are not consistent, and if there are some slight changes to the input data then it can have a big impact on the tree structure.

An ensemble method is a statistical method for making prediction or classification by combining two or more algorithms. The ensemble method is developed to overcome some limitations in a decision tree. An ensemble method can be categorized as bagging, stacking, or boosting. Random forest is an algorithm that is commonly used in the bagging method. Random forest can overcome overfitting in prediction or classification. The prediction or classification produced by random forest is relatively accurate. However, the interpretation of results from random forest is more difficult because it relies on the aggregation of many trees. On the other hand, Adaptive Boosting (AdaBoost) is a boosting technique that is more frequently applied. AdaBoost prioritizes some better trees and gives them more weight for getting the final choice, so that the resulting model is easier to be interpreted. Furthermore, it will improve the accuracy of decision tree algorithm (Bauer & Kohavi, 1999; Breiman, 1996; Kohavi & Kunz, 1997; Maclin & Opitz, 1997).

The purpose of our research is to examine the data patterns related to the graduation status of a bachelor's degree from Universitas Negeri Jakarta (UNJ). The data pattern is used to classify if an undergraduate student from UNJ will graduate on time or not on time by comparing the Adaboost, Random Forest, and Decision Tree algorithm. The variables that are used for classifying the status are the grade point average and the number of credits taken by the students. For the first-year students, the classification process will be based on their grade point average (GPA) and ‘satuan kredit semester/SKS’ (semester credit units) earned in semester one. The classification process for

the second-semester students will be based on the GPA and total credits taken in semester one and semester two. The classification process for the third-semester students will be based on the accumulation of GPA and the total credits taken in semester one, two, and three. The classification procedures are similar for the students at the fourth, fifth, and sixth semester. The findings of our research are expected to help some study programs at UNJ for identification the graduation status of their students as soon as possible. Early detection of a student's ability to graduate 'on time' or 'not on time' is critical so that a study program can provide some special treatments, especially for the students who are identified as 'not on time' for his graduation.

2 Methodology

2.1 AdaBoost

The AdaBoost is a machine learning technique that improves the performance of predictive models by integrating some simple prediction models (weak learners) into a more complex and accurate model (Freund and Schapire,1996). The weak learner is a classification technique whose classification results are only slightly more accurate than the random guessing method (Rokach, 2010). In each iteration, the AdaBoost will give more weight to data samples that were improperly classified in the previous iteration. As a result, it will allow the Adaboost to improve the prediction for samples which are difficult to be classified. In each iteration, a new predictive model is generated and combined with the previous models, so that it will be produced a final predictive model which is better and more effective. AdaBoost may generate some prediction models that are quite robust and provide good generalizations to the structured data (Schapire, 2013). The pseudocode for Adaboost algorithm could be described as follows (Shalev-Shwartz & Ben-David, 2014):

input:
  training set $S = (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$
  weak learner WL
  number of rounds T
initialize $\boldsymbol{D}^{(1)} = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$.
for $t = 1, \ldots, T$:
  invoke weak learner $h_t = WL\left(\boldsymbol{D}^{(t)}, S\right)$
  compute $\varepsilon_t = \sum_{i=1}^{m} D_i^{(t)} I_{[y_i \neq h_t(\boldsymbol{x}_i)]}$
  let $w_t = \frac{1}{2}\ln\left(\frac{1}{\varepsilon_t} - 1\right)$
  update $D_i^{(t+1)} = \frac{D_i^{(t)} exp\left(-w_t y_i h_t(\boldsymbol{x}_i)\right)}{\sum_{j=1}^{m} D_j^{(t)} exp\left(-w_t y_j h_t(\boldsymbol{x}_j)\right)}$ for all $i = 1, \ldots, m$
output the hypothesis $h_s(\boldsymbol{x}) = sign\left(\sum_{t=1}^{T} w_t h_t(\boldsymbol{x})\right)$.

2.2 Permutation Feature Importance

Breiman (2001) created the permutation feature significance technique to assess the importance of each feature in an ML model. This strategy works by calculating how much the model's performance changes when the value of each feature is randomized. The relative contribution of each feature to the model's accuracy can be estimated by measuring these changes. The higher the difference in accuracy upon randomization, the more significant the feature in the model. This technique is relatively straightforward and may be used in a wide range of ML models, with the results assisting in feature selection and the building of better models. The permutation feature importance technique procedure, according to Louppe et al. (2013), is given by as follows:

`input`:
fitted predictive model m
tabular training dataset D
`for` each feature j:
`for` each repetition k in $1, \ldots, K$:
`Randomly shuffle` column j of dataset D to obtain $\widetilde{D}_{k,j}$, a corrupted version of the data.
`Compute` the score $s_{k,j}$ of model m on corrupted data $\widetilde{D}_{k,j}$.
`Compute` importance i_j for feature f_j defined as:

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}$$

2.3 Research Data

The data of GPA for UNJ students from Semester 105 (2016) to Semester 114 (2021) were obtained from the Unit Pelaksana Teknis Teknologi, Informasi dan Komunikasi (UPT TIK) UNJ. Our observation subjects are the undergraduate students who enrolled at UNJ in the academic year of 2016/2017 and 2017/2018. They are considered as our observation units because we could have the information about their GPA from semester 1 until their graduation, so that we can identify them as a student who graduate 'on time' or 'not on time'. In 2016 and 2017, there were 10,215 undergraduate students at UNJ. This number also includes the students who withdrew or moved to other universities. Then, our analysis is only based on 8,295 students.

2.4 Data Analysis Procedures

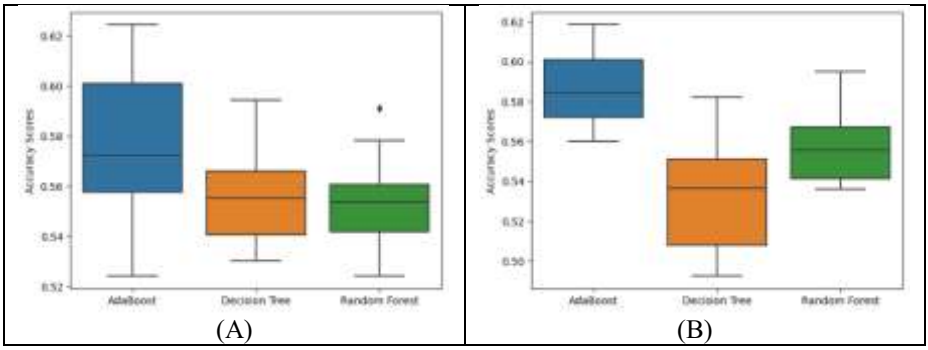
1. The data was divided into two parts: training data (up to 80% of total data) and testing data (up to 20% of total data).
2. A model was selected using 10-fold cross-validation on training data from semesters 1 to 6. For each semester, three models (AdaBoost, Random Forest, and Decision Tree) were specified. The input variables were GPA and number of credits. A total

of 18 models were built, and for each semester, the model with the highest accuracy was chosen.

3. An F-test was carried out to find out whether there was a significant difference in accuracy scores produced by the three models (AdaBoost, Random Forest, and Decision Tree). The level of significance (α) that was specified was 0.0167 (after Bonferroni correction).
4. T-tests were carried out to determine whether the pairs of models had a significant difference in accuracy scores. The level of significance (α) that was specified was 0.0167 (after Bonferroni correction). The t-tests were conducted if the results of the F-test suggested that there was a significant difference in accuracy scores among the three models. Then, based on the results of the t-tests, the model with the highest accuracy score was selected.
5. The testing data was fitted using the selected model in (4), and the final accuracy score was calculated.
6. Using the permutation feature importance technique, the percentage of importance of the factors (independent variable/input) used to classify the graduation status of UNJ students was calculated.

3 Results

The accuracy of decision trees, adaboost, and random forest algorithms for classifying whether a student will graduate 'on time' or 'not on time' are evaluated in our study. Figure 1 presents the boxplot of accuracy among AdaBoost, Random Forest, and Decision Tree algorithms from 1st semester until 6th semester.



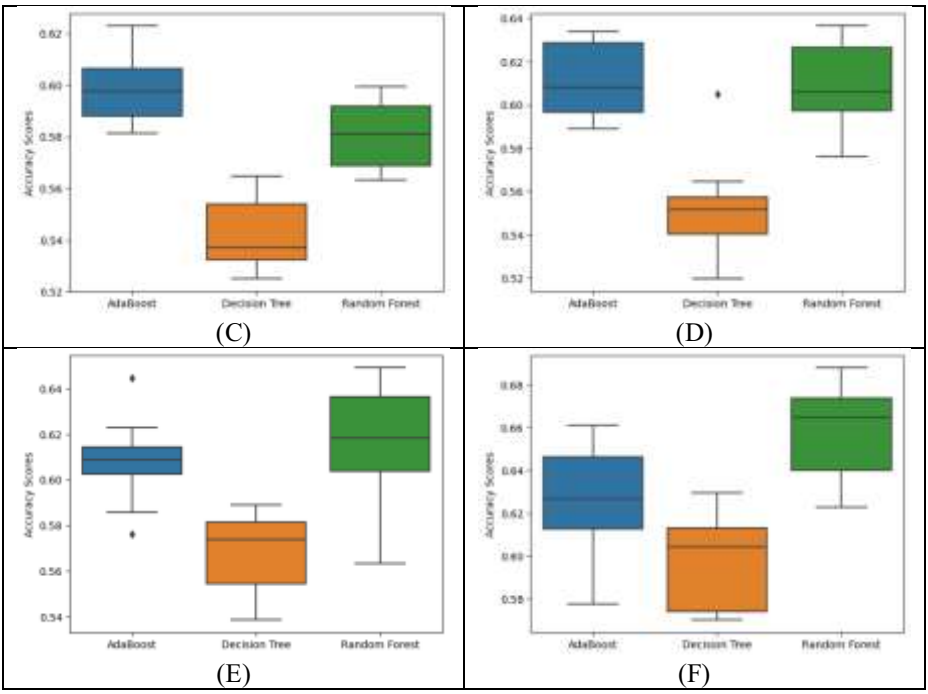


Fig. 1. The accuracy scores in the first semester (A), second semester (B), third semester (C), fourth semester (D), fifth semester (E), and sixth semester (F).

The Adaboost algorithm has the highest average accuracy (around 62%), with a substantial difference in accuracy scores among the three algorithms (p-value close to zero). It is considered the best algorithm for classifying whether a student will graduate on time or not based on student's performance in the first semester.

In the second semester, the random forest algorithm produces the most consistent classification results due to its smallest variance of accuracy scores. However, the Adaboost algorithm has the largest average accuracy score (58.7%), which is not significantly different from the random forest. This makes it the best algorithm for this semester.

For the third semester, the Adaboost algorithm has the highest average accuracy score (about 2% to 5% higher than the decision tree and random forest). The F-test results show that the decision tree, adaboost, and random forest algorithms have substantial differences in accuracy scores. Furthermore, the t-test findings reveal that adaboost is noticeably different from the other two methods.

In the fourth semester, the Adaboost algorithm has the highest average accuracy score (61.1%), with a smaller diversity of accuracy scores than random forest. This makes it the best method for classifying students in the fourth semester about their graduation time.

The random forest algorithm has the highest average accuracy score (62.5%) and is the most consistent in semester 5, with the smallest variety of accuracy values. Considering the higher average accuracy scores and lower diversity of accuracy scores, the random forest algorithm is considered the best for predicting semester 5 students' graduation time.

The random forest algorithm has the greatest average accuracy score in semester 6, with an average of 64.9%. The p-value is near zero, indicating a substantial difference between the adaboost, decision tree, and random forest methods. The t-test findings suggest that the random forest algorithm differs significantly from the other two algorithms.

Having examined the overall accuracy of the AdaBoost, Random Forest, and Decision Tree algorithms, we are now interested in understanding which variables are most important for each algorithm. Figure 2 will show the importance of each variable.

Weight	Feature
0.0506 ± 0.0135	SKS_S6
0.0321 ± 0.0196	S6
0.0181 ± 0.0099	S3
0.0072 ± 0.0108	S4
0.0061 ± 0.0044	S1
0.0047 ± 0.0103	S2
0.0017 ± 0.0097	S5

(A)

Weight	Feature
0.0558 ± 0.0204	SKS_S6
0.0282 ± 0.0184	S6
0.0140 ± 0.0115	S2
0.0065 ± 0.0169	S5
0.0035 ± 0.0187	S4
-0.0033 ± 0.0133	S3
-0.0076 ± 0.0147	S1

(B)

Weight	Feature
0.0807 ± 0.0103	SKS_S6
0.0418 ± 0.0217	S6
0.0224 ± 0.0083	S5
0.0203 ± 0.0116	S2
0.0192 ± 0.0113	S3
0.0186 ± 0.0100	S4
0.0051 ± 0.0114	S1

(C)

Fig. 2. The importance of independent variable in the AdaBoost (A), Decision Tree (B), and Random Forest (C)

Figure 2 depicts the importance level of variables based on the Permutation Feature importance criteria. By applying the three algorithms, i.e adaboost, decision tree and random forest, it can be found that the number of credits up to semester 6 and the semester 6 GPA are the variables which have the largest of importance level. The credits variable, on the other hand, has an importance level of up to 8% in the random forest algorithm, indicating that it determines nearly 10% of the prediction outcome. Figure 2 further shows that, except for semester 1's GPA, all variables in the random forest algorithm have almost the same level of importance. This contrasts with the adaboost and decision tree algorithms, where the importance level of other variables does not exceed 2% except for the 6th-semester credits and GPA variables.

4 Conclusion

In this study, the decision tree, adaboost, and random forest algorithm are evaluated to classify some undergraduate students in UNJ whether they will graduate on time or not time. The classification is carried out for the students from semester 1 through semester 6. By utilizing the GPA and the number of credits that earned for the subjects taken by

the student, an undergraduate student from semester 1 through semester 6 can be classified whether his (or her) graduation status will be on time or not on time in the end of their studies. The performance of ensemble methods (Adaboost and random forest) is better than the decision tree algorithm, with the average of accuracy scores difference ranging 2% to 5% in each semester. Furthermore, our study revealed the two key variables for classifying graduation status, namely the number of credits up to semester 6 and the GPA in semester 6. The number of credits has an importance level of roughly 5% to 8% in the ensemble methods, whereas the GPA has an importance level of around 3% to 4%. Meanwhile, in the decision tree algorithm, the number of credits has an importance level of roughly 6% and the GPA of around 3%. Other variables have a 2% importance level in the random forest method. However, they are less than or equal to 1% in the decision tree and adaboost. The final result from our classification obtains the highest accuracy score is 64%.

References

1. Bauer, E., & Kohavi, R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, 36(1), 105–139 (1999).
2. Breiman, L. Bias, Variance, and Arcing Classifier (Issue April) (1996).
3. Breiman, L. Random forests. Machine Learning, 45(1), 5–32 (2001).
4. Cantwell, R. H., & Moore, P. J. The Development of Measures of Individual Differences in Self-Regulatory Control and Their Relationship to Academic Performance. Contemporary Educational Psychology, 21(4), 500–517 (1996).
5. Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., & Lykeridou, K. Examining students' graduation issues using data mining techniques-The case of TEI of Athens. AIP Conference Proceedings, 1644, 255–262 (2015).
6. Freund, Y., & Schapire, R. E. Experiments with a New Boosting Algorithm. Proceedings of the 13th International Conference on Machine Learning, 148–156 (1996).
7. IBM Cloud Education, what is Machine Learning? https://www.ibm.com/cloud/learn/machine-learning, last accessed 2023/06/04.
8. Kamil, M., & Cholil, W. Analisis Perbandingan Algoritma C4.5 dan Naive Bayes pada Lulusan Tepat Waktu Mahasiswa di Universitas Islam Negeri Raden Fatah Palembang. Jurnal Informatika, 7(2), 97–106 (2020).
9. Kohavi, R., & Kunz, C. Option Decision Trees with Majority Votes. In Proc. 14th International Conference on Machine Learning, 1996, 161–169 (1997).
10. Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Artificial Intelligence in Design '96, 151–170 (1996).
11. Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. Understanding Variable Importances in Forests of Randomized Trees. Advances in Neural Information Processing Systems, 26 (2013).
12. Maclin, R., & Opitz, D. An Empirical Evaluation of Bagging and Boosting. In Proc. 14th National Conference on Artificial Intelligence Tools, 546–551 (1997).
13. Mayasari, N. Comparison of Support Vector Machine and Decision Tree in Predicting On-Time Graduation (Case Study: Universitas Pembangunan Panca Budi). Int. J. Recent Trends Eng. Res, December (2016).

14. Rokach, L. Pattern Classification Using Ensemble Methods. World Scientific Pub. Co (2010).
15. SAS Institute Inc., Machine Learning: What it is and why it matters. https://www.sas.com/en_us/insights/analytics/machine-learning.html, last accessed 2023/06/04.
16. Schapire, R. E. Explaining adaboost. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, 37–52 (2013).
17. Shalev-Shwartz, S., & Ben-David, S. Understanding Machine Learning from Theory to Algorithms. Cambridge University Press (2014).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.