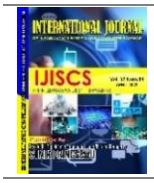

IJISCS (International Journal of Information System and Computer Science)
IJISCS, Volume 7, Number 3, December 2023
E ISSN: 2598-246X; P ISSN: 2598-0793, pp. 195-199
Accredited SINTA 4 Number 79/E/KPT/2023
https://jurnal.ftikomibn.ac.id/index.php/ijiscs

STUDENT GRADUATION TIME PREDICTION USING LOGISTIC REGRESSION, DECISION TREE, SUPPORT VECTOR, AND ADABOOST ENSEMBLE LEARNING

*Corresponding author
[1]*ardhana.desfiandi@binus.ac.id
[2]bsoewito@binu.edu

Ardhana Desfiandi[1]*, Benfano Soewito[2]
[1,2]School of Computer Science Bina Nusantara University, Jakarta, Indonesia
[1,2]Jl. KH. Syahdan No. 9, Kemanggisan, Palmerah Jakarta, Indonesia

Article history:
Received August 31, 2023
Revised September 18, 2023
Accepted October 10, 2023

Keywords:
Data mining;
Machine Learning;
Classification;
Ensemble Learning;
Adaboost

Abstract
Universities in Indonesia are working hard to improve the graduation rates of their students as it is considered a measure of success and quality in terms of accreditation. This study focuses on analyzing the effectiveness of machine learning algorithms, regression, Support Vector Machine (SVM) Decision Tree and ensemble learning, with AdaBoost wether the Computer Science students will graduate on time or not. The data used for this analysis consists of student records from 2015 to 2019. Includes 14 variables. To understand the relationships between these variables a two-dimensional visualization called a Heatmap was employed. The research findings indicate that the Support Vector Machine (SVM) and AdaBoost Decision Tree (DT) algorithm performs better than the other algorithms. The Decision Tree and AdaBoost (DT) model achieved an F1-score of 0,76 and 0,82. This research contributes towards enhancing education management by facilitating decision making to ensure timely graduation, for student

1.0 INTRODUCTION

Predicting student graduation time is an important task for universities and educational institutions to better manage student resources and plan academic programs. Graduation time prediction can help identify at-risk students who may need additional support and intervention to improve student success rates. However, predicting student graduation time is a challenging task due to the complexities of factors that influence student performance and progress, such as academic performance, personal characteristics, socioeconomic status, and other external factors. In Indonesia, the quality of higher education institutions is determined by the accreditation grade issued by the Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT). To achieve better accreditation, higher education institutions must prioritize timely graduation rates and focus on improving them [1]. Accurate measurement of the number of graduates each academic year can help institutions achieve this goal.

According to [2] Data mining is the process of identifying significant patterns and insights from extensive datasets. These datasets can originate from several sources, such as databases, data warehouses, the internet, other information repositories, or real-time streaming data. Data mining using machine learning algorithms has become increasingly used in the education

sector to analyze students' academic performance, predict student graduation, and address other related concern [3]. In recent times, the utilization of machine learning has expanded significantly in Higher Education Institutions, and it is increasingly being used in scientific research known as Educational Data Mining (EDM) [4]. One of the primary research of educational data mining is to predict student graduation [5].

Machine learning, a subfield of artificial intelligence, focuses on the development of algorithms and models that allow computers to learn and make predictions from data without being explicitly programmed. Machine learning algorithms can be trained on historical student data, enabling them to recognize patterns and relationships that can inform predictions about future student outcomes. In the realm of predicting graduation time, machine learning techniques offer a powerful toolset for constructing predictive models that can generalize from historical data and provide timely estimates for individual students

This study used several classification algorithms, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), and AdaBoost Classifier to predict student graduation. The K-Fold Validation method was used to evaluate the performance of these algorithms, and a confusion matrix was used to determine their accuracy. In this research, several classification algorithms are used for student graduation prediction such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Ada Boost Classifier. K-Fold Cross-Validation technique is used to ensure that training and test database were not coincidental.

Through the application of these data mining and machine learning techniques, this research aims to provide a comprehensive analysis of their effectiveness in predicting student graduation time. By comparing the performance of logistic regression, SVM, decision trees, and AdaBoost ensemble learning, valuable insights can be obtained regarding their strengths, limitations, and applicability in educational settings. The findings of this study have the potential to guide educational institutions in optimizing academic planning, resource allocation, and personalized interventions to improve student success rates. Overall, this research contributes to the growing body of knowledge on using data mining and machine learning approaches for predicting student graduation time. By harnessing the power of these techniques, educational institutions can enhance their ability to support students effectively and facilitate timely degree completion, ultimately fostering student success and academic achievement

Related Works.

In recent years, student's graduation predictions have attracted attention in the field of education. Researchers have explored various machine learning methods, including ensemble methods, to improve the accuracy of such predictions. This section provides an overview of related work using ensemble methods, specifically Adaboost, and Suppor Vector Machine (SVM), logistic regression, and decision trees, as base classifiers for predicting student graduation time. The variable most commonly used to predict student graudation cases according to [3], [6]–[8], are variables related to student performance, such as GPA, attendance, exams, and number of credit passed. They also use demographic variables such as gender, age, Parents educational background, Parent's income, and disability. Some research also have used variables related to extracurricular activities, such as their social interaction network [6].

Several studies have compared the performance of different machine learning algorithm for predicting student graduation. A research conducted by [6] compared the performance of Naïve Bayes, Decision Tree, Neural Network, and Support Vector Machine with four different k-folds of 5,10,15, and 20. According to their research findings, the Support Vector Machine algorithm (PolyKernel), perform more accurate than other classifiers. Furthermore, they discovered that utilizing either a 5-fold or 20-fold cross-validation method gives the best results for this particular experiment.

[7]compared Logistic Regression, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, and K nearest neighbors to predict early detection on student graduation time. The results show that Decision Tree and Random Forest shown the best performance of 93.65% accuracy. Other studies using machine learning algorithm in educational data mining to predict student graduation time have been conducted using an ensemble model. A study conducted by [5] on student graduation time using machine learning ensemble model to improve classification algorithms accuracy. The results of this study reveal that logistic

regression performed better than other classifiers that is Neural Network, Decision Tree, and Naïve Bayes with the accuracy rate of 86.82%.

While the research on prediction student graduation time conducted by [8] using ensemble model with eight classification algorithms include Logistic regression, Decision Tree Classifier, Gaussian, Random Forest Classifier, Ada Boost Classifier, Support Vector Coefficient, KNeighbors Classifier and MLP Classifier, the highest accuracy of 91.87% was obtained for the MLP Classifier and 91.64% for Support Vector Machine.

Given the numerous variations of classification algorithms, it can be challenging to determine the optimal algorithm for a particular classification task as each algorithm has its own distinct advantages and limitations [8]. According to Raschka (2015) in research [8] In order to effectively combine several classification algorithms, specific rules are used to merge the strengths of each individual algorithm through the ensemble machine learning method. This can result in better performance compared to using a single classification approach. According to [9], ensemble learning techniques has gained increased attention in the field of predictive modeling. Such techniques involve combining multiple learning algorithms, resulting in enhanced accuracy of predictions.

2.0 METHODOLOGY

This paper presents a student graduation prediction model using the ensemble method. An ensemble method is a type of learning method that integrates multiple models to solve a problem. Unlike traditional learning methods that use a single model to train data, ensemble methods use a collection of models to train data and combine the results. Ensembles generally provide more accurate predictions than individual models. The purpose of this approach is to enable an accurate assessment of the factors that influence a student's graduation time prediction. The steps of the proposed methodology are shown in Figure 1.

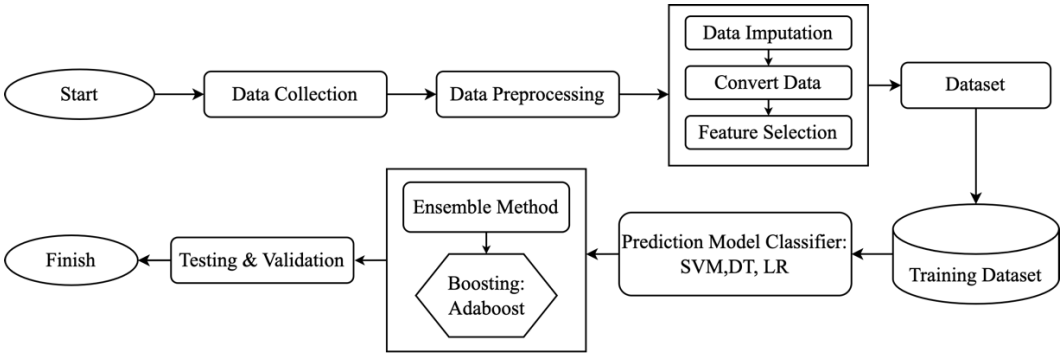

FIGURE 1. Student's Graduation Prediction Model Research Steps

This methodology starts by collecting data consists of 660 computer science students of Darmajaya Institute of Informatics and Business from 2014-2020. The dataset includes features such as name, gender, student status, enrolment year, parents education level, parents income, Educational track, student's internal and external activities, Grade Point of 1st semester until 7th semester, GPA, and graduation time. This step is followed by Data Preprocessing which included addressing missing values through data imputation using mean or mode values, converting categorical attributes into numerical form, and Lastly, feature selection was executed through *Univariate Feature Selection*, resulting in the identification of significant variables for predicting graduation time. After that, the the goal of the study is to have a refined dataset for generating highly accurate models, both individual and ensemble, by ensuring the dataset aligns with research requirements and contains clean and pertinent data. Next is Training Dataset phase, data is split through Train Validation Split..

In this study, ensemble methods are applied to provide an accurate evaluation for the features that may have an impact on student graduation time, and to improve the performance model that predicts the student's graduation time. The classification algorithms used in our comparative study: logistic regression, support vector machines (SVM), and decision trees. We then describe the ensemble learning approach used to combine the predictions of multiple models, highlighting the advantages of this technique in improving

predictive performance. Finally, we present the evaluation metrics used to assess the performance of each model, including accuracy, precision, recall, and F1-Score. Accuracy is the proportion of the total number of predictions where correctly calculated. Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases. Recall is the ratio of correctly classified cases to the total number of unclassified cases and correctly classified cases [10].

TABLE 1. Confusion Matrix

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

3.0 RESULANTS

Results of Features Correlation

TABLE 2. Features Correlation

Features	**Graduation Time**
Parent's Income	679485.684630
GPA 3rd Semester	10.338820
Educational Track	8.293161
GPA_5th Semester	5.728623
GPA 4th Semester	5.666952
GPA_2nd Semester	4.094044

Based on the correlation of the 6 features used in Table 2 above, The variable with the highest correlation to student graduation is parent's income. Students with low parent's income can motivate themselves to graduate on time because if they don't graduate on time, they will have to extend their time in college, incurring additional expenses. Other variables include GPA 2nd Semester, GPA 3rd Semester, GPA 4th Semester, and GPA 5th Semester. This is based on the common practice of new students who typically aim for a high GPA in the early semesters of their studies.

Performance Results Using Base Classifiers

TABLE 3. Result Comparison of Base Classifiers

	Performance Metrics				**Confusion Matrix**			
Classifiers Type	Accuracy	Precision	Recall	F1-Score	TP	TN	FP	FN
Logistic Regression	0.75	0.74	0.76	0.75	74	26	24	74
Decision Tree	0.70	0.77	0.55	0.64	54	84	44	16
Support Vector Machine	0.76	0.77	0.74	0.76	73	22	22	78
AdaBoost (SVM)	0.75	0.76	0.72	0.74	72	75	26	25
AdaBoost (LR)	0.74	0.74	0.73	0.74	72	75	25	26
AdaBoost (DT)	0.81	0.78	0.87	0.82	85	76	13	24

As shown in Table 3 above, The highest F1-Score values were achieved by the single algorithms Support Vector Machine (SVM) and AdaBoost Decision Tree (DT), with scores of 0.76 and 0.82, respectively. The highest accuracy values in this study were also obtained by Support Vector Machine (SVM) and AdaBoost Decision Tree (DT), which were 0.76 and 0.81, respectively. This indicates that both of these models are very effective at making predictions. The highest precision values were 0.77 and 0.78, achieved by Support Vector Machine (SVM) and AdaBoost Decision Tree (DT), respectively.

From the research results, it can be concluded that the model's ability to predict all positive values aligns with the precision values obtained. Therefore, in the case of using 6 variables to

predict the graduation time of students, it is evident that the Support Vector Machine (SVM) and AdaBoost Decision Tree (DT) algorithms have the best model performance among the four single models, which include Logistic Regression (LR), Decision Tree (DT), AdaBoost Logistic Regression (LR), and AdaBoost Support Vector Machine (SVM).

4.0 CONCLUSION

In this study, we aimed to predict student graduation at IIB Darmajaya using the Logistic Regression, Support Vector Machine, Decision Tree, and AdaBoost Ensemble Learning algorithms. The objective of this study was to determine which algorithm is the best for predicting graduation time and to provide recommendations to the institution's leadership on reducing the number of students who do not graduate on time. In this research, the dependent variable used in modeling was parent's income, GPA 3rd semester, educational track, GPA 5th semester, GPA 4th semester, and GPA 2nd semester. Meanwhile, the independent variable in this study was the graduation status, which indicates whether the student graduated on time or not. It can be concluded that the single algorithms Support Vector Machine (SVM) and AdaBoost Decision Tree (DT), are the best algorithms to use in cases like this. Support Vector Machine (SVM) and AdaBoost Decision Tree (DT) achieved the best performance metrics compared to other algorithms, especially obtaining the highest F1-Score values of 0.76 and 0.82. After conducting feature selection using univariate analysis to determine which features are most strongly correlated with student graduation time, the results revealed that parent's income, GPA 2nd semester, GPA 3rd semester, GPA 4th semester, and GPA 5th semester were the most relevant features.

REFERENCES

[1] BAN-PT, "Akreditasi Perguruan Tinggi: Instrumen Pemantauan Dan Evaluasi Peringkat Akreditasi Perguruan," Jakarta, 2022.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2022.

[3] R. Bakri, N. P. Astuti, and A. S. Ahmar, "Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 4, no. 2, pp. 259–265, 2022, doi: 10.35877/454ri.asci1581.

[4] S. K. Wanjau and G. M. Muketha, "Improving Student Enrollment Prediction Using Ensemble Classifiers," *Int. J. Comput. Appl. Technol. Res.*, vol. 7, no. 03, pp. 122–128, 2018, doi: 10.7753/ijcatr0703.1003.

[5] A. C. Lagman, L. P. Alfonso, M. L. I. Goh, J. A. P. Lalata, J. P. H. Magcuyao, and H. N. Vicente, "Classification algorithm accuracy improvement for student graduation prediction using ensemble model," *Int. J. Inf. Educ. Technol.*, vol. 10, no. 10, pp. 723–727, 2020, doi: 10.18178/ijiet.2020.10.10.1449.

[6] D. Ruete *et al.*, "Early Detection of Delayed Graduation in Master's Students," in *ASEE Annual Conference and Exposition*, 2021, pp. 1–20.

[7] N. M. Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. A. Hamid, and A. M. A. Malik, "Review on Predicting Students' Graduation Time Using Machine Learning Algorithms," *Int. J. Mod. Educ. Comput. Sci.*, vol. 11, no. 7, pp. 1–13, 2019, doi: 10.5815/ijmecs.2019.07.01.

[8] C. W. Teoh, S. B. Ho, K. S. Dollmat, and C. H. Tan, "Ensemble-Learning Techniques for Predicting Student Performance on Video-Based Learning," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 8, pp. 741–745, 2022, doi: 10.18178/ijiet.2022.12.8.1679.

[9] S. C. Agrawal, "Deep learning Based Non-Linear Regression for Stock Prediction," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1116, no. 1, pp. 1–7, 2021, doi: 10.1088/1757-899x/1116/1/012189.

[10] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, 2016, doi: 10.14257/ijdta.2016.9.8.13.