PREDICTING STUDENT PERFORMANCE USING MACHINE LEARNING FOR STUDENT MANAGEMENT IN UNIVERSITY

Berlit Deddy Setiawan, Dermawan Wibisono
Institut Teknologi Bandung, Indonesia
E-mail: berlit_deddy@sbm-itb.ac.id, dwibisono@sbm-itb.ac.id

Abstract

Higher education institutions play an important role in providing quality education and producing skilled human resources. In Indonesia, the demand for higher education is increasing due to population growth and increasing awareness of the importance of higher education. ABC University, which is currently ranked 46-50 in the Indonesia Uni Rank 2023, faces challenges in ranking. To thrive in this competitive landscape, universities must be selective in admitting quality students and ensuring an effective academic development process. Machine learning capabilities can be leveraged to predict potential student academic performance, facilitate timely interventions, and support for improving learning outcomes. However, currently there is no research that focuses on building predictive models that integrate student profiles with academic achievement. This study aims to establish a relationship between the theory of the Random Forest algorithm and the prediction of potential student achievement. The aim is to develop an accurate and efficient method for managing student affairs at ABC University. This study uses both quantitative and qualitative approaches, with a focus on numerical data analysis and produces classification predictions. The research process begins with a thorough analysis of the business situation to understand the university environment and determine research topics. The researcher then defines research boundaries, prioritizes key issues, and develops a research framework. The study analyzed the profiles of students who graduated in 2016-2017, combined with academic achievement data. Regression test was conducted to determine the effect of 18 attributes on performance. Random Forest Machine Learning was compared with other techniques to identify the most accurate predictive model for student academic performance. ABC University, the Random Forest model achieves a prediction rate of 89.60%.

Keywords: Higher education institutions, Predicting students' potential academic performance, Random Forest algorithm, Factors influencing student performance

Introduction

Higher education institutions are academic educational institutions that play a role in providing relevant and high-quality higher education, capable of producing qualified human resources that meet the qualifications demanded by the job market (Tien et al., 2020). That higher education refers to the university level, consisting of various faculties that offer academic education in specific disciplines (Barthos, 1992). In Indonesia, there is a growing demand for higher education due to population growth and increased awareness of the importance of education (Ajisuksmo, 2017). However, it is worth noting that as of 2022. Indonesia has 4,593 universities, but only 20 campuses are included in the world rankings (Lukman & Said, 2022). Of the 20 campuses, only five are included in the top 500. Below is a table of the top five universities in Indonesia, including the ranking of ABC University (Rosser, 2019).

Table 1
Top 5 University in Indonesia, UniRank version 2023

Rank	University	City
1	Universitas Indonesia	Depok
2	Universitas Gajah Mada	Sleman
3	Institut Teknologi Bandung	Bandung
4	Universitas Brawijaya	Malang
5	Universitas Bina Nusantara	Jakarta
46-50	ABC University	-

Based on the table 1 above, 4 out of 5 top universities are public universities and one is a private university. ABC University is ranked 46-50, which is still quite far behind. There are 10,157,323 senior high school and vocational high school students in Indonesia. both public and private Those who are active in the 12th grade or 3rd grade in 2023 have the potential to become university students (Ramadhan & Megawati, 2023). Below is a table of the top five regions in Indonesia with the highest number of high school and vocational school students.

Table 2
Top 5 regions with the highest number of high school and vocational school students in Indonesia

No	Regions	Senior High School	Vocational School
1	West Java Province	792.478	1.066.366
2	East Java Province	535.577	761.539
3	Central Java Province	437.985	783.805
4	North Sumatra Province	381.827	304.346
5	South Sulawesi Province	228.913	123.12

Based on tables 1 and 2, the competition among prospective students is spread across the island of Java, with West Java province being the largest market, which also includes the market for ABC University. The competition among universities to recruit high-quality students has become increasingly intense (Musselin, 2018). Therefore, each university is expected to strive to provide the best quality education to succeed in this competitive environment (Hemsley-Brown et al., 2016). The success of a higher education institution is often measured by the quality of students who pursue education at that institution (Dumford & Miller, 2018). Conversely, the failure to produce high-quality students is often seen as a lack of management capability on the part of the university in delivering the teaching and learning process (Markova et al., 2017). Below is a map showing the distribution of universities in Indonesia (Negoro et al., 2021).
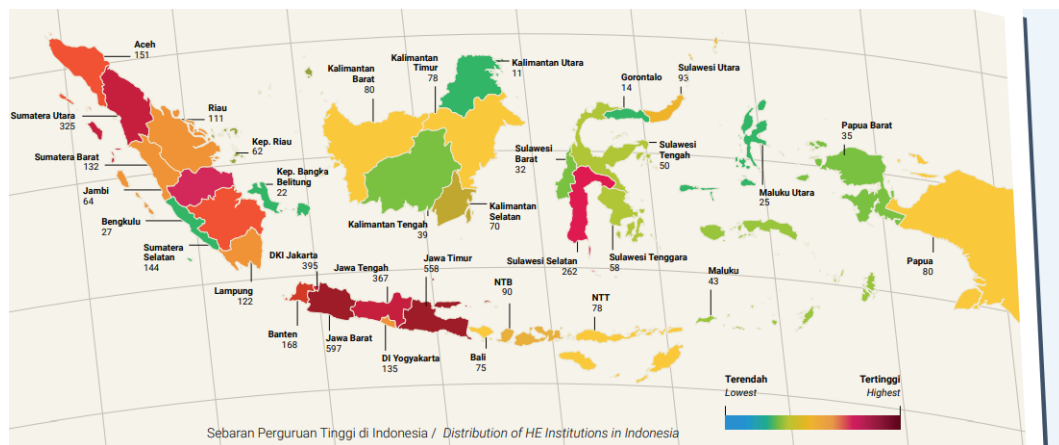

Figure 1
Distribution of High Education Institutions in Indonesia 2020

Based on Figure 1, the highest distribution of universities in Indonesia is on the island of Java to obtain high-quality student outcomes, universities ABC need to be selective in accepting qualified prospective students and are also expected to ensure the academic development process of students, providing them with a high chance of success in their education. One crucial step in the selection and management of the education process is predicting the potential academic performance of students (Alyahyan & Düştegör, 2020).

Predicting the potential academic performance of students is essential because by analyzing student data and indicators, it is possible to assess whether a student is likely to succeed and perform well in their education or not (Casillas et al., 2012). This prediction will utilize machine learning capabilities to determine predictions and identify the most significant attribute indicators that influence student performance. this enables timely interventions and the necessary support to improve learning.

This prediction serves as one of the considerations for assessing the potential academic performance and serves as a warning for students with declining or poor

academic performance potential. This prediction will greatly help the university ABC to remain in the strategy of becoming a university with the best quality and can compete against other universities, with indicators of the success of students who excel and manage education with quality. Below is a table of the study duration classification for undergraduate programs at ABC University.

Tabel 3
Classification of Length of Study bachelor's degree in ABC University

No	Length of Study	Classification
1	3 Years	Fast
2	3.5 – 4 Years	On Time
3	4.5 – 6 Years	Lated

Based on tables 3 the prediction of students' academic performance will determine whether students with specific indicators can graduate earlier, graduate on time, or graduate late. The following is a graph analyzing the "Five Whys" technique at ABC University.
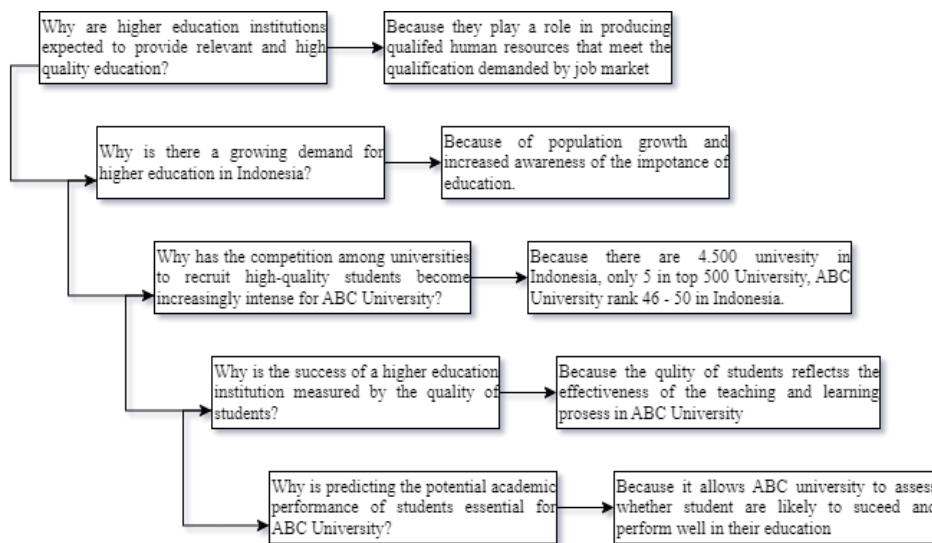

Figure 2
five why analysis ABC University

Based on Figure 2 above, it can be inferred that predicting the potential academic performance is crucial for ABC University.

The prediction results are highly beneficial for the University in determining which students require academic assistance to enhance their academic performance. Therefore, ABC University needs a method for predicting students' academic performance to remain competitive with other universities.

Based on the background that has been described, this study aims to create a predictive model for potential student academic achievement that can be used in the Directorate which is responsible for managing students at the university level.

Research Methodology

The research methodology relates to the steps and procedures that will be carried out to achieve the objectives and obtain answers to research problems (Hidayat & Alifah, 2022b). These steps and procedures embody the research framework. This study aims to link the Random Forest algorithm theory with the method of predicting student performance potential. The researcher intends to find an accurate and efficient performance method in improving student management in universities (case study: ABC University). This study uses two types of models, namely a quantitative approach and a qualitative approach (Hidayat & Alifah, 2022a). The data originally used in this study were categorical data and then converted into numerical data. The converted data was then analyzed using the Decision Tree method to produce predictions for the potential classification of student academic achievement. The main research topics and research framework established are supported by a literature review, which helps identify research gaps to be explored and provides an understanding of existing research. Researchers also conducted Focus Group Discussions (FGD) with program managers to consider the factors that affect student performance and the required predictive output. The next step involves data collection, where data is collected according to research needs and stored for processing. The second stage is the data modeling stage using the implementation of the Decision Tree algorithm and the k-Folds cross validation technique.

The population in this study consisted of students at ABC University who would be the subject of research on predicting student academic achievement. The sample data to be used includes academic data and student profiles from 2016 to 2017 in all study programs. The primary data for this study were obtained from the ABC University academic database, supplemented by Focus Group Discussions (FGD). Interviews were conducted with experts who are competent in the field to determine the factors that influence student performance which will be predicted in this study. Data on factors that affect the academic achievement of ABC University students include taking credits, debt, passing credits, grade point average (GPA), majors, mother's income, income, father's income, mother's education, father's education, gender, mathematics, father Job. This secondary data is the ABC University Academic Database. The method of data analysis in this study consists of attribute identification. Identification of research attributes is carried out through a process of reviewing the literature on previous research as a basis for a preliminary survey to obtain relevant attributes for modeling and analysis.

Results and Dicsussion

This chapter will present the study's results and findings. To examine the data distribution and analyze the regression lines of independent and dependent variables, the researcher utilized IBM SPSS Statistics 27 software. The results and discussion will be

divided into several subchapters: Data Distribution, Implementation of Data Processing using Random Forest and Decision Tree, and Training and Testing Results.

A. Analysis

During the analysis phase, several steps will be undertaken. These include examining the data distribution, analyzing the regression lines of independent and dependent variables, utilizing IBM SPSS Statistics 27 software, selecting relevant features, and implementing Decision Tree methodology.

B. Data Distribution

The purpose of presenting the data distribution in this section is to provide a comprehensive overview of the distribution of research data. This section includes the percentage distribution of data for each input variable, as well as the output variable in the study. The data distribution of the input attributes from the Science and Technology or Social Humanities Grade 2016 and 2017 datasets is presented. The analysis is performed using IBM SPSS Statistics 27, and the descriptive statistics and frequencies are displayed in the following table.

Table 4
Data distribution of input variable(s)

Descriptive Statistics						
	N	Range	Minimum	Maximum	Mean	Std. Deviation
Gander	495	1	1	2	1.53	.500
Admission	495	1	1	2	1.09	.293
School Type	495	1	1	2	1.23	.420
Mathematic	495	3	1	4	2.86	.774
English	495	2	2	4	3.25	.564
Father Job	495	5	1	6	2.56	1.341
Mother Job	495	5	1	6	4.18	1.860
Father Education	495	6	1	7	4.31	1.491
Mother Education	495	6	1	7	3.88	1.338
Father Income	495	5	1	6	4.22	1.099
Mother Income	495	5	1	6	2.25	1.620
Debt	495	1	1	2	1.33	.469
Length of Study	495	2	1	3	2.06	.622
GPA	495	2	2	4	3.68	.474
Department	495	1	1	2	1.73	.446
Credit Take	495	89	114	203	131.11	18.880
Credit pass	495	60	114	174	127.20	14.672
%_Attendance	495	38.32	37.46	75.78	48.741	7.837884

Valid (listwise)	N	495

From Table 4 above, there are 18 attributes that will serve as the main data for building the prediction model. These data will be subjected to linear regression testing in order to determine the most influential attributes. Here is an example of data distribution:

Table 5
Data distribution of input variable(s)-Gender

Gander		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	233	47.1	47.1	47.1
	Male	262	52.9	52.9	100.0
	Total	495	100.0	100.0	

From Table 5 above, there are 495 student data, consisting of 233 female data, which accounts for 47.1% of the data distribution, and 262 male data, which accounts for 52.9% of the total data distribution. The following is a graph representing the data distribution in Table 5.
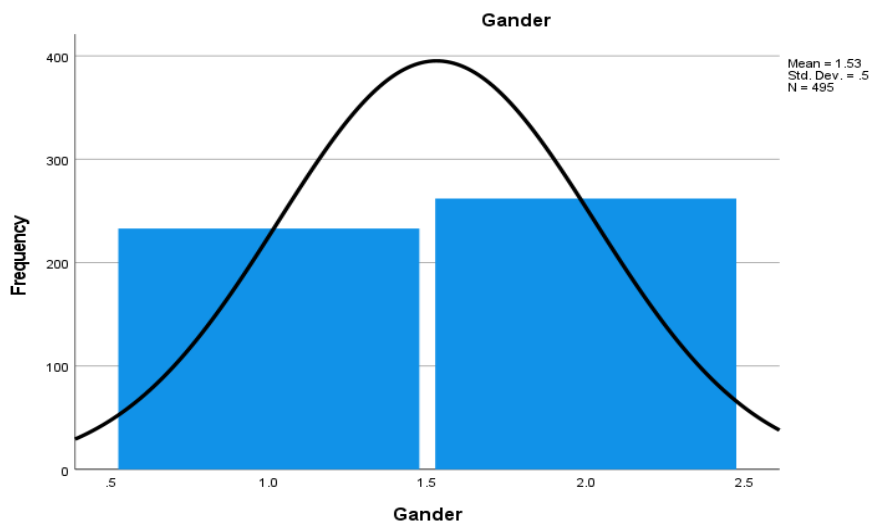

Figure 3
Data distribution Gender

From Figure 3 above, the gender distribution at ABC University exhibits a right-skewed curve because the number of males is greater than the number of females.

Table 6
Data distribution of input variable(s)-School Type

School Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	State	382	77.2	77.2	77.2
	Private	113	22.8	22.8	100.0
	Total	495	100.0	100.0	

From Table 6 above, there are 495 student data, consisting of 382 school type is state, which accounts for 77.2% of the data distribution, and 113 School type is private, which accounts for 22.8% of the total data distribution. The following is a graph representing the data distribution in Table IV.3.
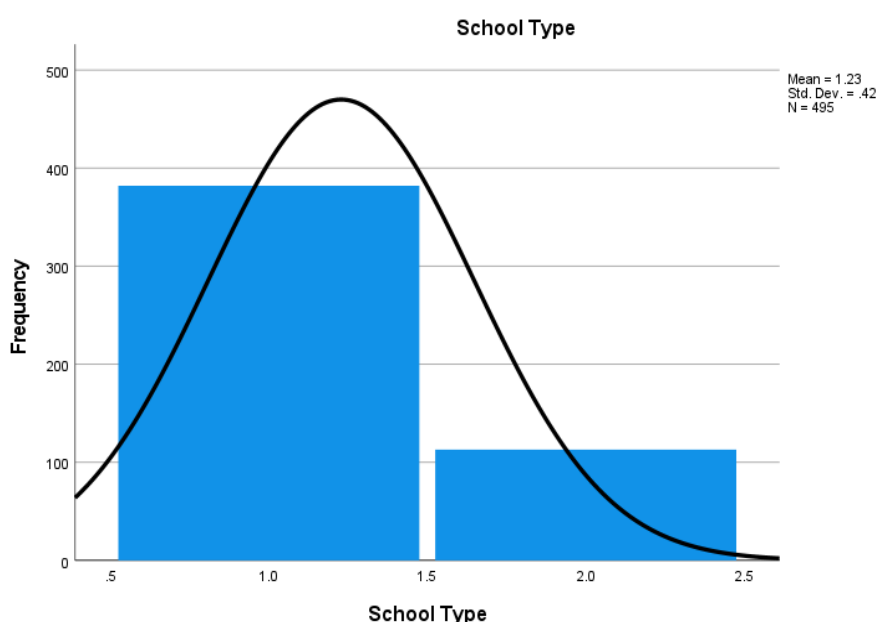

Figure 4
Data Distribution School Type.

From Figure IV.2 above, the school type distribution of students at ABC University exhibits a left-skewed curve because there are more students from public schools compared to private schools.

C. Business Solution

Based on the distribution of data, it comprises profiles of students who graduated in the academic years 2016-2017, which are then combined with achievement or academic performance data. A regression test will be conducted to determine which of the 18 attributes influence student performance and can be used as independent variables. Following the analysis of independent variables, the Random Forest Machine Learning process will be compared with other Machine Learning techniques to identify the model with the highest accuracy. This accuracy result will serve as the predictive model for

students' academic performance. The data analysis process will be divided into two research subjects: 1. Science and Technology study programs and 2. Social Humanities study programs.

D. Regression Liner Variable.

Based on the distribution of this data, a regression test will be conducted to determine which of the 18 attributes influence student performance and can be utilized as independent variables. The linear regression analysis of the variables will be performed using IBM SPSS Statistics 27 Software, employing Automatic Linear Modeling to enhance model accuracy through boosting. Throughout the analysis, the following attributes will be generated.
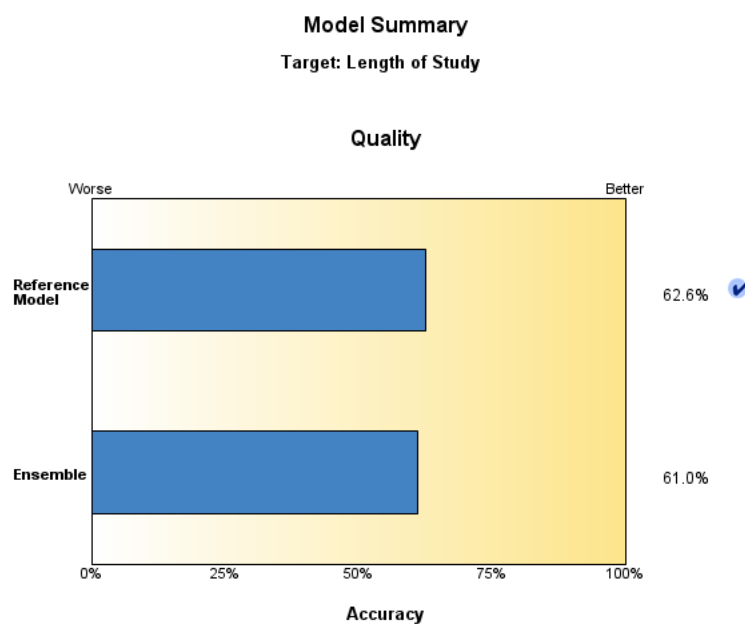
1. Science and Technology study programs

Figure 5

Model Summary Automatic Linear Regresstion – Science and Technology study programs

From Figure 5 above, the model summary of the regression results for the Science and Technology study program attribute indicates an accuracy of 62.6%. The accuracy of the reference model is better than the accuracy of the ensemble data.

Table 7

Automatic Linear Modeling - enhance model accuracy (boosting) - Science and Technology study programs

No.	Nodes	Importance	Importance	V4	V5
1	Fahter_Edu	0.0085	0.0085	Father Education	0.0085
2	Math	0.0139	0.0139	Mathematic	0.0139
3	Father_Job	0.0158	0.0158	Father Job	0.0158

4	School	0.0248	0.0248	School Type	0.0248
5	Admission	0.0254	0.0254	Admission	0.0254
6	Father_Income	0.0422	0.0422	Father Income	0.0422
7	GPA	0.1323	0.1323	GPA	0.1323
8	Credit_Pass	0.1655	0.1655	Credit pass	0.1655
9	Debt	0.262	0.2620	Debt	0.2620
10	Credit_Take	0.3096	0.3096	Credit Take	0.3096

From the table above, it can be observed that out of the 18 attributes tested, 10 attributes significantly influence the Science and Technology study programs. The most influential attribute is "Credit_take," with a predictor importance value (V5) of 0.3096. On the other hand, the attribute with the least influence is "Father's education," with a predictor importance value (V5) of 0.0085. Below is a table of the distribution of attributes with good regression values:
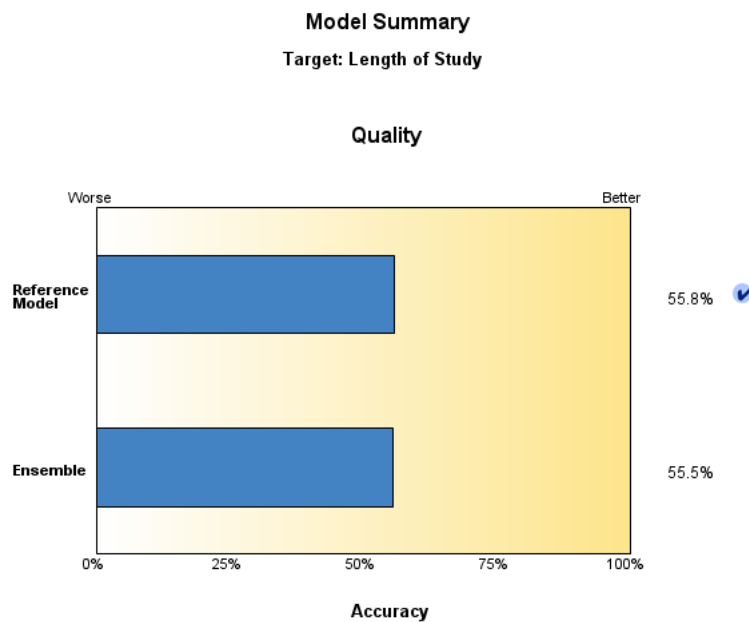
2. Social Humanities study programs

Figure 6
Model Summary Automatic Linear Regresstion – Social Humanities study programs.

From Figure 6 above, the model summary of the regression results for the social humanities study program attribute indicates an accuracy of 55.8%. The accuracy of the reference model is better than the accuracy of the ensemble data. Below is a table of the distribution of attributes with good regression values:

Table 8

Automatic Linear Modeling - enhance model accuracy (boosting) – Social Humanities study programs

No	Nodes	Importance	Importance	V4	V5
1	Father_Job	0.0076	0.0076	Father Job	0.0076
2	Admission	0.01	0.0100	Admission	0.0100
3	Math	0.0129	0.0129	Mathematic	0.0129
4	Father_Income	0.0166	0.0166	Father Income	0.0166
5	Gender	0.021	0.0210	Gander	0.0210
6	Mother_Income	0.0224	0.0224	Mother Income	0.0224
7	GPA	0.109	0.1090	GPA	0.1090
8	Credit_Pass	0.2217	0.2217	Credit pass	0.2217
9	Debt	0.2709	0.2709	Debt	0.2709
10	Credit_Take	0.3035	0.3035	Credit Take	0.3035

From the table above, it can be observed that out of the 18 attributes tested, 10 attributes significantly influence the Science and Technology study programs. The attribute with the highest influence is "Credit_take," which has a predictor importance value (V5) of 0.3035. On the other hand, the attribute with the least influence is "Father's job," with a predictor importance value (V5) of 0.0076.

Through the process of Automatic Linear Modeling - enhanced model accuracy (boosting), 12 important attributes were identified for predicting the academic performance at ABC University. These attributes include:

Tabel 9
The Attribute Result Regression

No.	Attribute	Science and Technology study programs	Social Humanities study programs
1	Father Education	v	x
2	Mathematic	v	v
3	Father Job	v	v
4	School Type	v	x
5	Admission	v	v
6	Father Income	v	v
7	GPA	v	v
8	Credit pass	v	v
9	Debt	v	v
10	Credit Take	v	v
11	Gander	x	v
12	Mother Income	x	v

From the table above, the regression analysis resulted in 12 important attributes for Science and Technology study programs. Gender and mother's income were not included in the analysis. Similarly, for Social Humanities study programs, father's education and school type were not used. These attributes will undergo further analysis using the Random Forest method and will be compared with other machine learning models. The analysis will focus on the Random Forest machine learning model.

From the regression analysis conducted on the 12 data attributes, further analysis will be performed using random forest machine learning. Additional machine learning tests will be conducted to determine which method yields the highest prediction accuracy. The machine learning analysis will be executed using Orange Data Mining Software, and the resulting analysis will be as follows.

3. Science and Technology study programs

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.945	0.889	0.889	0.896	0.889	0.786
Tree (1)	0.904	0.867	0.865	0.889	0.867	0.757
Logistic Regression	0.934	0.852	0.852	0.852	0.852	0.703
SVM	0.850	0.763	0.762	0.763	0.763	0.525

Figure 7
Result Mechine Learning Science and Technology study programs

Based on the machine learning testing conducted on the Science and Technology study programs at ABC University, it was found that the Random Forest model achieved a prediction rate of 0.896, an F1 score of 0.889, and a recall of 0.889. The decision tree structure is as follows:
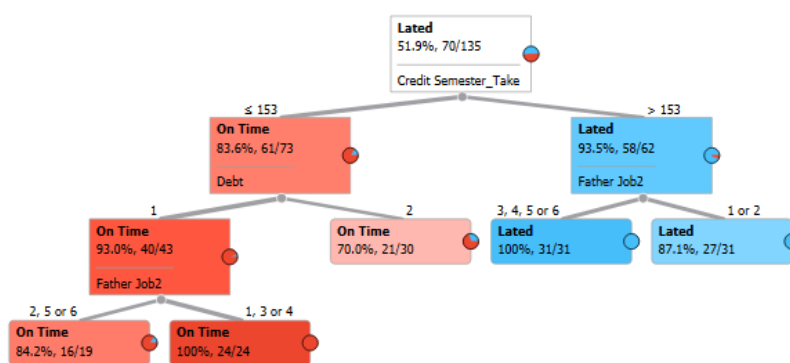

Figure 8
Decision Tree Science and Technology study programs

From Figure 8 above, the decision tree results for Science and Technology study programs yield the following prediction model: study duration is determined

by Credit Semester take. If a student wants to complete their studies on time, they should take more than 153 Credit Semester. If they have academic debt, the determination will be based on the father's occupation.The Random Forest structure is as follows:
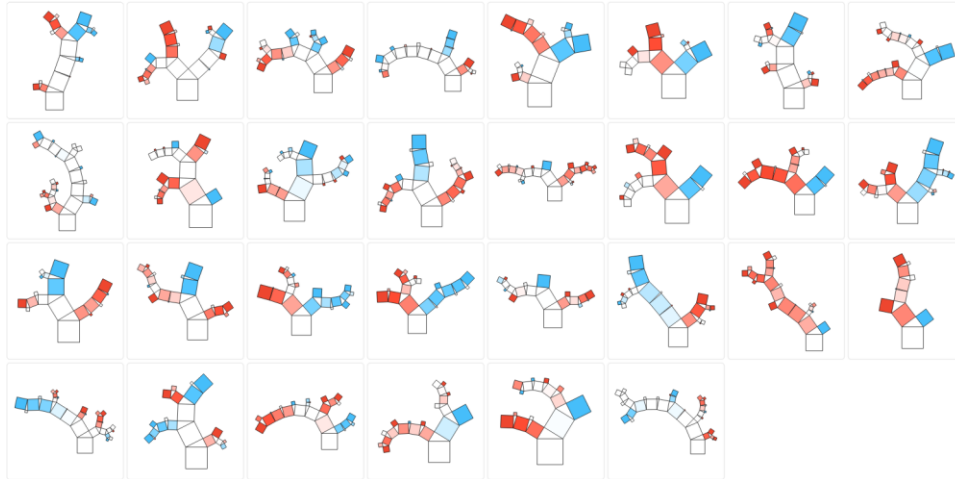

Figure 9
Random Forest Science and Technology study programs

4. Social Humanities study programs

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.882	0.786	0.780	0.780	0.786	0.556
Tree (1)	0.848	0.756	0.754	0.753	0.756	0.511
Logistic Regression	0.857	0.725	0.717	0.715	0.725	0.421
SVM	0.790	0.686	0.624	0.674	0.686	0.243

Figure 10
Result Mechine Learning Social Humanities study programs

Based on the machine learning testing conducted on the Science and Technology study programs at ABC University, it was found that the Random Forest model achieved a prediction rate of 0.780, F1 score of 0.780, and recall of 0.786. The decision tree structure is depicted below:
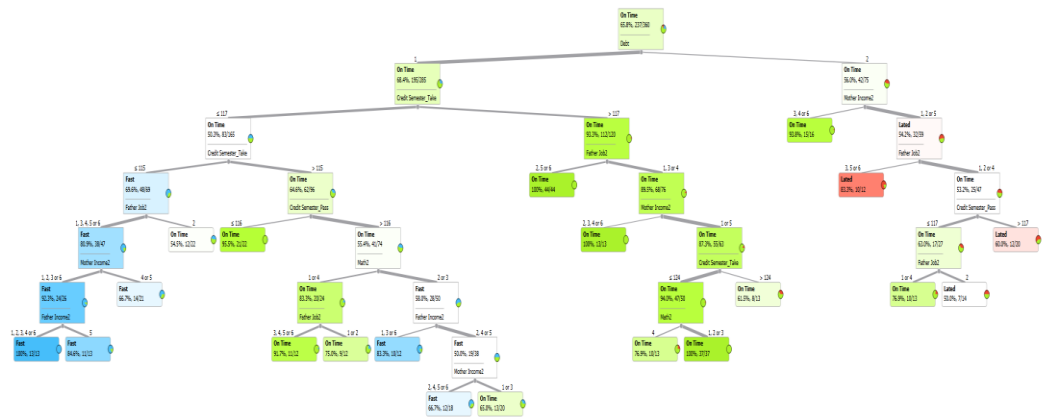
Figure 11
Decision Tree Social Humanities study programs

From Figure 11 above, the decision tree results for Science and Technology study programs yield the following prediction model: study duration is determined by Academic Debt, Credit Semester Taken, Father's Occupation, Mother's Occupation, and Father's Income Type. For example, if a student has academic debt, the prediction will consider the mother's income. If the income type is 1, 2, or 5, the prediction will be based on the father's occupation. If the father's occupation is 1, 2, or 4, the prediction will then consider the Credit Semester Taken. If the credit semester taken is less than 117, the student will be predicted to graduate late. And the Random Forest structure is as follows:
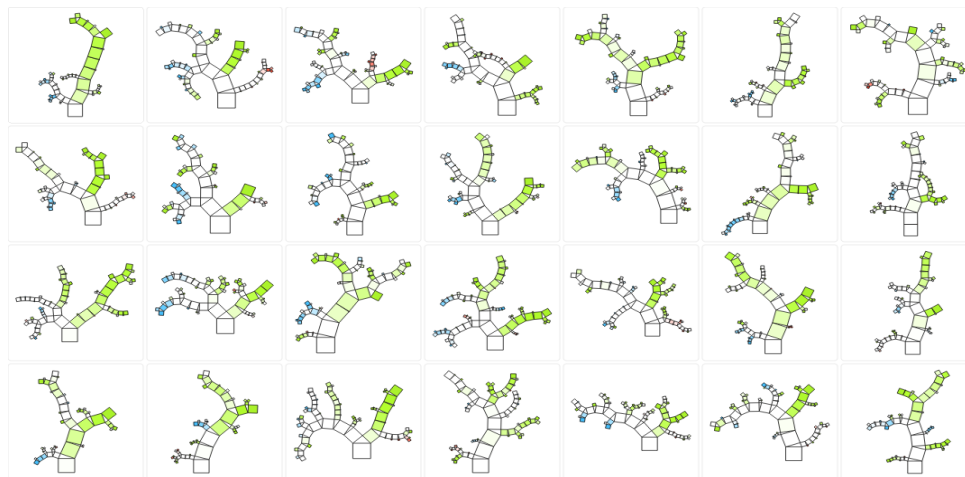

Figure 11
Random Forest Social Humanities study programs

E. Cross Validation

Based on the machine learning testing results, it is evident that random forest is the most suitable model for making predictions at ABC University. Subsequently, a 10-fold Cross Validation test will be conducted to determine the average prediction results.

The Cross Validation analysis will be performed using Orange Data Mining Software, and the analysis results are presented below.

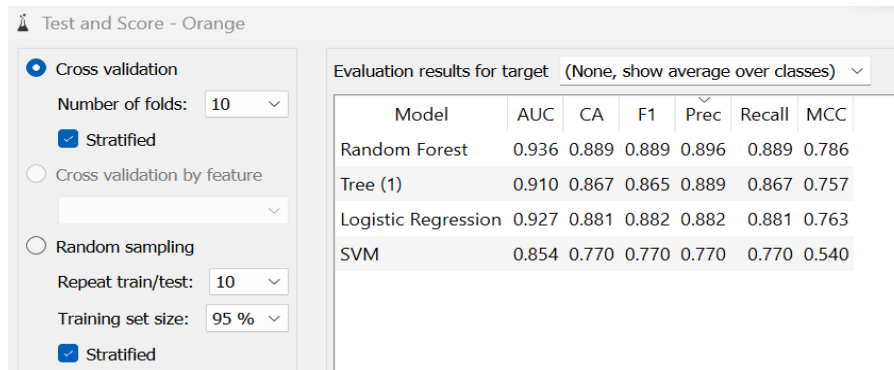1. Science and Technology study programs

Figure 12
Result 10-folds Cross Validation Science and Technology study programs
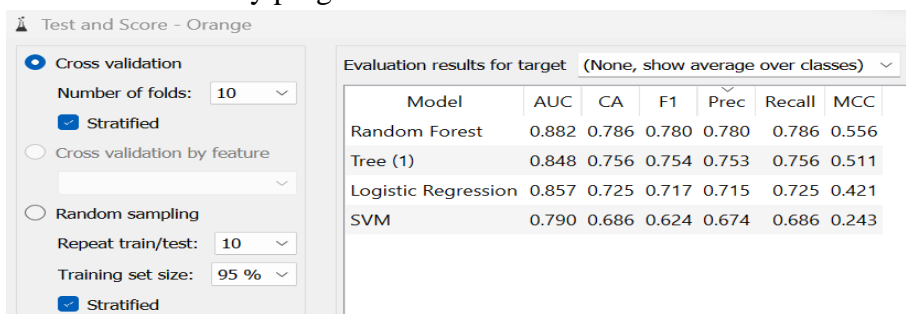
2. Social Humanities study programs

Figure 13
Result 10-folds Cross Validation Social Humanities study programs

Based on the above results, it can be observed that the prediction was conducted using 10-fold cross-validation, yielding the following average results.

Tabel 10
fold cross-validation

Model	F1	Prediction	Recall
Science and Technology study programs	0.889	0.896	0.889
Social Humanities study programs	0.754	0.780	0.756

From the table above, for the Science and Technology study program model, the prediction accuracy reaches 89.6%. As for the Social Humanities study program, the prediction accuracy reaches 78.00%.

F. Implementation Plan
1. Implementation Plan

This academic performance prediction model can be implemented promptly with the approval of the academic vice chancellor and the three directorate offices of ABC University. Currently, ABC University is in the process of developing an early warning system, and the prediction model will serve as a valuable reference for its functioning.
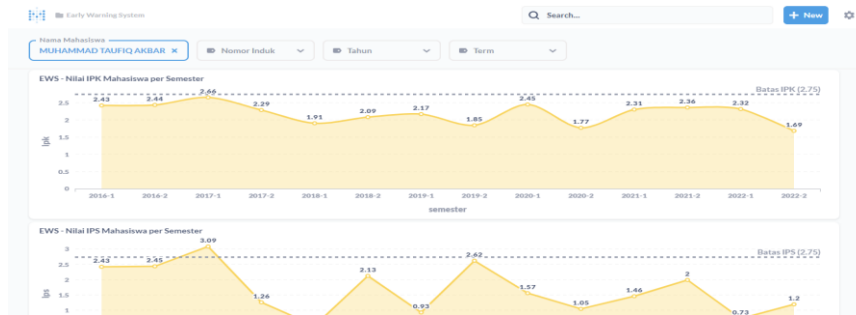

Figure 14
Early Warning System ABC University

The proposed framework can be implemented within the next three months, aligning with the planned timeline outlined in the following work breakdown structure.

Tabel 11
Implementation Plan

No	Activity	July				Aug				Oct			
		1	2	3	4	1	2	3	4	1	2	3	4
1	Evaluation data early warning system	▮	▮	▮	▮								
2	Implementation of prediction model in early warning system					▮	▮						
3	Testing 1							▮					
4	Monitoring Prediction model								▮				
5	Testing 2									▮			
6	Monitoring Prediction model										▮		
7	Validation Model											▮	▮

Conclusion

Based on the issue that occurred at ABC University, from the research results we can know that: (1) Out of the 18 attributes tested to predict the length of study for students at ABC University using linear regression, 12 independent attributes (input variables) were found to be significant. It is worth noting that these attributes differ from those examined in previous studies. (2) Based on the analysis conducted using Random Forest on the selected 12 attributes, the following prediction scores were obtained. (3) Based on the research findings, it is evident that random forest is the suitable machine learning model for creating prediction models.

Result Mechine Learning Science and Technology study programs to implement the proposed model prediction, several considerations should be considered, which include: (1) Determining the information, data, and reports that need to be supported and how the academic performance prediction model can complement each other. (2) Ensuring that the new student management system is effectively communicated to all directorates responsible for student performance. (3) Analyzing the benefits and cost

implications associated with each activity related to the implementation of the academic performance prediction model. (4) Providing an informative and communicative interface accessible to all employees at ABC University.

Berlit Deddy Setiawan, Dermawan Wibisono

BIBLIOGRAPHY

Ajisuksmo, C. R. P. (2017). Practices and challenges of inclusive education in indonesian higher education. *ASEACCU Conference on "Catholic Educational Institutions and Inclusive Education: Transforming Spaces, Promoting Practices, and Changing Minds". Assumption University of Thailand, Bangkok August*, 21–27.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*, 1–21.

Barthos, B. (1992). Perguruan tinggi swasta di Indonesia: proses pendirian, penyelenggaraan, dan ujian. *(No Title)*.

Casillas, A., Robbins, S., Allen, J., Kuo, Y.-L., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology*, *104*(2), 407.

Dumford, A. D., & Miller, A. L. (2018). Online learning in higher education: exploring advantages and disadvantages for engagement. *Journal of Computing in Higher Education*, *30*, 452–465.

Hemsley-Brown, J., Melewar, T. C., Nguyen, B., & Wilson, E. J. (2016). Exploring brand identity, meaning, image, and reputation (BIMIR) in higher education: A special section. In *Journal of Business Research* (Vol. 69, Issue 8, pp. 3019–3022). Elsevier.

Hidayat, A. R., & Alifah, N. (2022a). Analysis of The Basis of The Creative Economy in The Development Strategy of Economic Innovation. *Asian Journal of Social and Humanities*, *1*(3), 95–104.

Hidayat, A. R., & Alifah, N. (2022b). Reading for Students in English Language Education Programs. *International Journal of Social Health*, *1*(2), 57–63.

Lukman, L., & Said, I. M. (2022). Strategi Kesantunan Pemain Game dalam Saluran Youtube "Jess No Limit." *Jurnal Onoma: Pendidikan, Bahasa, Dan Sastra*, *8*(1), 63–76.

Markova, T., Glazkova, I., & Zaborova, E. (2017). Quality issues of online distance learning. *Procedia-Social and Behavioral Sciences*, *237*, 685–691.

Musselin, C. (2018). New forms of competition in higher education. *Socio-Economic Review*, *16*(3), 657–683.

Negoro, Y. A. T., Marthanty, D. R., & Soeryantono, H. (2021). Analysis of the green infrastructure implementation to the enhancement of environmental support capacity (Case study: Watershed outside University of Indonesia). *IOP Conference Series: Materials Science and Engineering*, *1098*(2), 22050.

Ramadhan, S., & Megawati, S. (2023). Implementasi Kebijakan Merdeka Belajar Kampus Merdeka Dalam Meningkatkan Kualitas Pendidikan Mahasiswa di Universitas Negeri Surabaya. *Publika*, 1581–1592.

Rosser, A. (2019). Big ambitions, mediocre results: Politics, power and the quest for world-class universities in Indonesia. *Transformations in Higher Education Governance in Asia: Policy, Politics and Progress*, 81–99.

Tien, N. H., Jose, R. J. S., Mai, N. P., Long, N. T., & Hai, T. V. (2020). Current State of Human Resource in International Universities of Vietnam. *International Journal of Multidisciplinary Research and Development*, 7(7), 22–27.