A Two-Stage Early Prediction Model to Monitor the Students' Academic Progress

Publisher: IEEE

Susana Limanto ; Joko Lianto Buliali ; Ahmad Saikhu **All Authors**

Abstract: The high dropout rate and the low percentage of undergraduate students who graduate on time are some of the problems at higher education institutions. Various research ha... **View more**

Metadata

Abstract:

The high dropout rate and the low percentage of undergraduate students who graduate on time are some of the problems at higher education institutions. Various research has been conducted to overcome these problems, one of which is predicting student performance. Predicted results of academic performance at the beginning of the class can be used as an early warning for undergraduate students and lecturers to take the necessary actions to succeed well at the end of the semester. However, not all significant predictor variables can be obtained at the beginning of the class. The purpose of this research is to develop a two-stage prediction model to determine the passing of each student from the courses undertaken. The first stage of the prediction model was developed at the beginning of the class while the second was developed at the end of the eighth week by adding two predictor variables, namely mid-term test score and the number of presence in class. Four different methods are used: Decision Tree (able to generate rule that are easy to interpret), Random Forest (able to handle high dimensional data), Support Vector Machine (able to eliminate overfitting problem), and Logistic Regression (able to make good prediction of success in a course. In testing, all performance measures of Logistic Regression method were superior to other methods both in the first and second stages. It is also seen that the addition of predictor variables was able to increase the performance measures of all prediction models by up to 7%.

Published in: 2022 10th International Conference on Information and Communication Technology (ICoICT)

Date of Conference: 02-03 August 2022

INSPEC Accession Number: 22184752

Date Added to IEEE *Xplore*: 21 October 2022

DOI: 10.1109/ICoICT55009.2022.9914882

ISBN Information:

Publisher: IEEE

Conference Location: Bandung,

Contents

I. Introduction

The quality of a higher education can be measured by the percentage of students who graduate on time and dropout rates [1], [2]. Usually, the higher the percentage of students who graduate on time and the lower the dropout rate, the better the quality of the higher education in educating the students. The quality of the higher education can affect the views of the community so that it has an impact on people's interest in studying at the higher education [1]. Therefore, until now the dropout rate and the percentage of students graduating on time have become one of the main concerns in various higher educations.

Authors

Figures

References

Keywords

Metrics


Telkom University
UNIVERSITAS GADJAH MADA
MMU
MULTIMEDIA UNIVERSITY
2022
THE 10th
ICoICT
International Conference on Information and Communication Technology
"INTELLIGENT AND AUTONOMOUS E-SERVICES TOWARD A HYBRID INTERACTION ERA"
Telkom UNIVERSITY
G20 INDONESIA 2022
IEEE INDONESIA SECTION
Hak tek nas 27
IEEE Signal Processing Society Indonesia Chapter
Fakultas Informatika
School of Computing
Telkom University

Home Program Table of Contents TPC Committees Authors Other reviewers

2022 10th International Conference on Information and Communication Technology (ICoICT) | 978-1-6654-8165-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOICT55009.2022.9914868

Greetings from the General Chair

On behalf of the Organizing and Program Committees, we cordially invite you to the 10th International Conference on Information and Communication Technology (ICoICT) 2022, held on August 2-3, 2022. The 10th ICoICT 2022 is co-organized by Telkom University Indonesia, Multimedia University Malaysia, and Universitas Gadjah Mada Indonesia, in collaboration with the IEEE Indonesia Section and IEEE Signal Processing Society Chapter. Previous ICoICT conferences have successfully acted as a venue for bringing together various people from academia and industry to exchange and present the most recent challenges and advancements in information and communication technology (ICT). Papers from the preceding ICoICT 2013 - 2021 have been published in IEEE Xplore and are indexed in Scopus.

ICoICT 2013: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6569393

ICoICT 2014: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6908150

ICoICT 2015: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7203317

ICoICT 2016: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7565234

ICoICT 2017: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8054654

ICoICT 2018: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8509820

ICoICT 2019: https://ieeexplore.ieee.org/xpl/conhome/8825728/proceeding

ICoICT 2020: https://ieeexplore.ieee.org/xpl/conhome/9162155/proceeding

ICoICT 2021: https://ieeexplore.ieee.org/xpl/conhome/9527366/proceeding

The technical agenda for the 10th ICoICT 2022 includes five keynotes, three tutorials, and sixteen parallel sessions of five tracks on "Intelligent and Autonomous E-services toward a Hybrid Interaction Era".

In this event, a competition for the Best Paper Award is organized. The 10th ICoICT 2022 had 190 paper submissions from eighteen nations, of which ninety-four were accepted - 49.2% of the acceptance rate. All submitted papers have a rigorous peer-review procedure that assesses their importance, originality, and technical excellence. At least three professional reviewers independently examined each paper.

The 10th ICoICT 2022 is conducted virtually from the Telkom University campus. The organizing team has labored diligently to design a virtual conference that will be beneficial and interesting for both presenters and attendees. The conference format combines pre-recorded and asynchronous participation with Question-and-Answer (Q&A) and live video conversations.

Colleagues, friends, and organizations have collaborated to arrange the 10th ICoICT 2022 conference. We would like to thank everyone who has participated and supported our efforts in a variety of ways, as well as everyone who helped make this event feasible and successful. We also would like to express our gratitude to the Indonesian Ministry of Education, Culture, Research, and Technology, Bank Indonesia, and Yayasan Pendidikan Telkom, for entrusting their support to the ICoICT 2022 event. Also, the members of the Organizing Committee and Technical Committee, as well as our Telkom University colleagues, for their assistance with the organization and execution of this conference. We also want to express our appreciation to the reviewers who labored so diligently in evaluating the articles and making ideas for their enhancement. We extend our heartfelt appreciation to the Keynote and Tutorial Speakers. Lastly, we would like to acknowledge our gratefulness to the authors who have submitted their wonderful papers to this conference, as well as to all the attendees. We appreciate your virtual participation in the 10th ICoICT 2022 conference. We hope that the keynote talks, technical workshops, and parallel sessions stimulate your future research and provide you with an enjoyable experience.

The 10th International Conference on Information and Communication Technology (ICoICT) 2022 Important Dates:

Call for Paper: November 01, 2021

Paper Submission Deadline: January 7, 2022

Notification of Papers Acceptance: February 15, 2022

Paper Submission Deadline Round 2: April 10, 2022

Paper Submission Deadline Final Round: May 8, 2022

Notification of Papers Acceptance: June 11, 2022

Submission of Camera-Ready Papers and Author Registration Deadline: June 25, 2022

Conference Date: August 02-03, 2022

Dr. Vera Suryani

General Chair ICoICT 2022

2022 10th International Conference on Information and Communication Technology (ICoICT)
Organized by Telkom University, Multimedia University & Gajah Mada University
Prepared by EDAS Conference Services.

2022 10th International Conference on Information and Communication Technology (ICoICT) | 978-1-6654-8165-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOICT55009.2022.9914894

Table of Contents

2022 10th International Conference on Information and Communication Technology (ICoICT)

Digital Health Services

Emerging ICT and e-Services

ICT-based Education Services

Intelligent & Autonomous System e-Services

Network & Infrastructure for e-Services

Digital Health Services

Emerging ICT and e-Services

Emerging ICT and e-Services

Intelligent & Autonomous System e-Services

Intelligent & Autonomous System e-Services

Emerging ICT and e-Services

Emerging ICT and e-Services

Intelligent & Autonomous System e-Services

Intelligent & Autonomous System e-Services

Network & Infrastructure for e-Services

Digital Health Services

2022 10th International Conference on Information and Communication Technology (ICoICT) | 978-1-6654-8165-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOICT55009.2022.9914886

Committees

Steering Committee

A Adiwijaya (Telkom University, Indonesia)

Maman Abdurohman (Telkom University, Indonesia)

Siong Hoe Lau (Multimedia University, Malaysia)

Kok-Seng Wong (VinUniversitty, Vietnam)

Kim-Kwang Raymond Choo (University of Texas at San Antonio, USA)

Abdi T Abdalla (University of Dar es Salaam, Tanzania)

Sultan Alamri (SEU, Saudi Arabia)

Kiki Maulana Adhinugraha (La Trobe University, Australia)

Ari Moesriami Barmawi (Telkom University, Indonesia)

Parman Sukarno (Telkom University, Indonesia)

Rina Pudjiastuti (Telkom University, Indonesia)

Organizing Committee

General Chair

Vera Suryani (Universitas Telkom, Indonesia)

General Co-Chair

Imelda Atastina (Telkom University, Indonesia)

Ong Thian Song (Multimedia University, Malaysia)

TPC Chair

Ade Romadhony (Telkom University, Indonesia)

Track Chair

Didit Adytia (School of Computing, Telkom University, Indonesia)

Dody Qori Utama (Telkom University, Indonesia)

Tee Connie (Multimedia University, Malaysia)

Kusuma Ayu Laksitowening (Telkom University, Indonesia)

Ying Han Pang (Multimedia University, Malaysia)

Mardhani Riasetiawan (Universitas Gadjah Mada, Indonesia)

Publication Chair

Dawam Dwi Jatmiko Suwawi (Telkom University, Indonesia)

Rifki Wijaya (Telkom University, Indonesia)

Secretariat Chair

Siti Karimah (Telkom University, Indonesia)

Dita Oktaria (Telkom University, Indonesia)

Shih Yin Ooi (Multimedia University, Malaysia)

Ramanti Dharayani (Telko, Indonesia)

Finance Chair

Andrian Rakhmatsyah (Telkom University, Indonesia)

Siew Chin Chong (Multimedia University, Malaysia)

Aniq Atiqi (Telkom University, Indonesia)

Indwiarti Indwiarti (Telkom University, Indonesia)

Event and Logistic Chair

Fazmah Arif Yulianto (Telkom University, Indonesia)

Mira Kania Sabariah (Telkom University, Indonesia)

Muhammad Arief Nugroho (Telkom University, Indonesia)

Gamma Kosala (Telkom University, Indonesia)

Publicity Chair

Mahmud Dwi Sulistiyo (Telkom University, Indonesia & Nagoya University, Japan)

Mohd Fikri Azli Abdullah (Multimedia University, Malaysia)

Chia Sui Ong (Multimedia University, Malaysia)

Agung Toto Wibowo (Telkom University - Indonesia, Indonesia)

Prasti Eko Yunanto (Telkom University, Indonesia)

Rizka Reza Pahlevi (Telkom University, Indonesia)

Tutorial and Special Session Chair

Putu Harry Gunawan (Telkom University, Indonesia)

Sponsorship Chair

Kemas Wiharja (Telkom University, Indonesia)

Webmaster

Rahmat Yasirandi (King Mongkuts Institute of Technology Ladkrabang, Thailand & Telkom

University, Indonesia)

2022 10th International Conference on Information and Communication Technology (ICoICT)
Organized by Telkom University, Multimedia University & Gajah Mada University
Prepared by EDAS Conference Services.

2022 10th International Conference on Information and Communication Technology (ICoICT) | 978-1-6654-8165-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOICT55009.2022.9914840

Author index

2022 10th International Conference on Information and Communication Technology (ICoICT)
Organized by Telkom University, Multimedia University & Gajah Mada University
Prepared by EDAS Conference Services.

A Two-Stage Early Prediction Model to Monitor the Students' Academic Progress

Susana Limanto
Dept. of Informatics
Institut Teknologi Sepuluh Nopember
7025211020@mhs.its.ac.id
Dept. of Informatics Engineering
Universitas Surabaya
Surabaya, Indonesia
susana@staff.ubaya.ac.id

Joko Lianto Buliali
Dept. of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
joko@if.its.ac.id

Ahmad Saikhu
Dept. of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
saikhu@if.its.ac.id

***Abstract*—The high dropout rate and the low percentage of undergraduate students who graduate on time are some of the problems at higher education institutions. Various research has been conducted to overcome these problems, one of which is predicting student performance. Predicted results of academic performance at the beginning of the class can be used as an early warning for undergraduate students and lecturers to take the necessary actions to succeed well at the end of the semester. However, not all significant predictor variables can be obtained at the beginning of the class. The purpose of this research is to develop a two-stage prediction model to determine the passing of each student from the courses undertaken. The first stage of the prediction model was developed at the beginning of the class while the second was developed at the end of the eighth week by adding two predictor variables, namely mid-term test score and the number of presence in class. Four different methods are used: Decision Tree (able to generate rule that are easy to interpret), Random Forest (able to handle high dimensional data), Support Vector Machine (able to eliminate overfitting problem), and Logistic Regression (able to make good prediction of success in a course. In testing, all performance measures of Logistic Regression method were superior to other methods both in the first and second stages. It is also seen that the addition of predictor variables was able to increase the performance measures of all prediction models by up to 7%.**

Keywords—student performance, prediction, higher education, classification

I. INTRODUCTION

The quality of a higher education can be measured by the percentage of students who graduate on time and dropout rates [1], [2]. Usually, the higher the percentage of students who graduate on time and the lower the dropout rate, the better the quality of the higher education in educating the students. The quality of the higher education can affect the views of the community so that it has an impact on people's interest in studying at the higher education [1]. Therefore, until now the dropout rate and the percentage of students graduating on time have become one of the main concerns in various higher educations.

The number of students who enroll at a higher education is generally more than the number of students who graduate [3]. Statistics of Indonesia State Higher Education 2020 states that the dropout rate in Indonesia in 2019 is 7% (602,208 of 8,483,213 registered students) [4]. In addition, based on the preliminary sampling of 75 courses taught in the academic year 2018-2019 to 2020-2021 at a higher education in Surabaya, the average percentage of passing course was 75% and only 9.3% of courses have a passing rate of 100%. While the results of the sampling of 471 students from the 2017-2018 class at the same higher education, showed that about 16% of students graduated on time (without taking into account temporary study stops) and there were around 14% of students dropping out of college. This condition shows that the dropout rate and the number of students who graduate beyond the normal study period are still high.

There are two kinds of efforts to reduce the dropout rate and the percentage of students who graduate beyond the normal study period, namely screening prospective students so that students accepted by higher education are excellent students and increasing the passing rate for higher education students. To increase course passing rates, higher educations can make various efforts to help students, especially those with less ability, by providing academic advisors [5], guardian lecturers, assistance, and tutoring from lecturers outside of class hours [6]. In order that the efforts made by the higher education are on target and run optimally, it is necessary to have supporting information, such as academic data. However, the number of students with different backgrounds and abilities causes a lot of time to process the existing data into the useful information that can support decision making in helping students who have less ability.

Various studies have been conducted by utilizing academic data and technology to help provide the data needed by higher education. One form of existing research is to build a predictive model to provide the information needed to screen, monitor and improve student academic performance [7]. Research related to the prediction of student passing from the evaluation of first year studies [8], [9] is useful for screening excellent prospective students. Meanwhile, research related to the prediction of passing a course [10], [11] and prediction of the score of a course [3], [12], [13] can be used to support decision making in determining appropriate efforts to help students, especially those with less ability.

The results of predicting student performance can be useful for preparing actions to help students who are indicated to experience failure if they can be obtained as early as possible. Therefore, some researchers predict student performance at the beginning of the class [3], [13]. However, not all data used as predictive variables that have the potential to significantly influence can be obtained at the beginning of the class, for example test scores. Usually some test scores can be obtained after several classes.

Most researchers make a one-time prediction, namely after getting all the data from the predictor variables used.

Reference [14] conducted a research to evaluate whether student academic performance can be predicted in the early weeks of college. Predictions are made at the end of the 3rd, 6th, 9th, 12th, and 14th weeks using the same predictor variables but with increasing numbers according to those collected at the end of the week. Prediction results at the end of the third week showed that the model succeeded in predicting "unsuccessful" students with 74% accuracy and at the end of the fourteenth week gave the highest accuracy, 89%. The results also showed a tendency that the more data used (the longer the week) the higher the accuracy except at the end of the 6th week [14].

This research developed a predictive model consisting of two stages to determine the passing of each student from the courses taken. The first stage of prediction was carried out before the class start by using demographic data and academic history. While the second stage of prediction was made after the mid-term test by adding two predictor variables, the mid-term score and the number of attendance in the learning class. Mid-term test is usually held in the eighth week of sixteen meetings. In the two stages, four different method are used, namely: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR).

Testing was carried out to find out whether (1) the academic performance of students in a course can be predicted from the start or not, (2) the addition of predictor variables in the second stage could improve the performance of the first stage prediction results or not. The results of the first stage of prediction are useful as an early warning for students, lecturers, and institutions to prepare strategies, so that students who are predicted to "fail" can succeed well at the end of the class. While the prediction results of the second stage can be to be used to monitor student academic progress as well as prepare further strategies, especially for students who are predicted to fail.

The rest discussion is arranged in the following order. In the second part, a literature review is carried out on research related to student performance predictions that have been carried out by other researchers. The third section will discuss the methodology used in carrying out the research including the methods used to build the model and evaluate the prediction results. The results of the research will be discussed in the fourth section. Finally, the fifth section concludes the research results.

II. LITERATURE REVIEW

Educational Data Mining (EDM) is a process to explore the information and patterns needed from academic data in educational institutions that are stored in large quantities [15], [16]. Currently, EDM is still widely used by researchers to predict student performance [15]–[17] in order to improve the quality of student performance and the teaching and learning process [1]. This is because the prediction results can significantly be used to support decision making [15]. The main focus of existing research is to identify predictor variables that greatly affect student performance and to obtain predictive methods that can provide high predictive accuracy [17].

A. Predictor Variable

Various kinds of predictor variables are used by researchers to get the results of predicting student performance with high accuracy. Predictor variables that are usually used are academic data [3], [9]–[11], [13], [15], [16], [18]–[21], demographics [8]–[10], [13], [17]–[19], [21], social [10], managerial [10], and psychological [10], [15].

The most widely used academic data for predicting student performance are:

1. Grades (both the results of course assessments [1], [11], [19] and the grades of courses that have been taken [3], [13], [16], [20]).
2. Grade Point Achievement (GPA) [1], [9], [10].
3. Total credits taken [1], [10].
4. Presence [1], [18].
5. Frequency of accessing Learning Management System (LMS) [18], [19].
6. The value of the prerequisite courses [20].
7. The length of the prerequisite courses taken [20].
8. The load of credits taken [10], [17].

Academic and demographic data are the most widely used predictor variables. This is reinforced by [2] who states that academic data from previous education levels and demographics are the two predictor variables most often used by researchers to make predictions. This is because the two types of predictor variables are considered to have a significant influence on the prediction results [22]. However, [2] states that demographic data is a divergent factor in the research, in the sense that there are researchers who claim to be influential but there are also researchers who claim to have no effect, especially gender.

B. Prediction Method

In addition to the predictor variables, the prediction method is one of the important components that determine the accuracy of the prediction results. There are several prediction methods that are often used [22], namely DT [1], [8], [11], [15]–[19], [21], Naïve Bayes (NB) [1], [14], [15], [17]–[19], [21], k-Nearest Neighbor (k-NN) [1], [13], [14], [17], Rule based [8], Artificial Neural Network (ANN) and its development [10], [12], [15], [17], RF [14], SVM [14], [15], [17], and the Regression group [17].

Reference [21] developed a model to predict the academic performance of first-year students using the DT and NB methods. DT method provided greater accuracy than the NB method for both balanced datasets and imbalanced datasets. In the research conducted by [6], the DT method also provided the best accuracy, which was 86% compared to the kNN, ANN, and NB methods.

The predictive model of academic performance of prospective new students was developed by [15] using four different methods: ANN, DT, SVM, and NB. The results showed that the ANN method was able to provide the greatest accuracy (79%) but the DT method was able to provide the largest recall (80%) and F-measure (81%) [15]. This shows that the DT method is better in dealing with imbalanced data.

Reference [19] conducted research on the effect of student involvement on academic performance using the DT, JRIP, J48, Gradient-Boosted Trees (GBT), Classification and Regression Tree (CART), and NB methods. J48, DT, JRIP, and GBT showed better performance in terms of accuracy, kappa, and recall.

Reference [23] developed a model to predict at-risk students in order to prepare for immediate intervention. The model was developed using the ANN, SVM, and LR methods. The ANN method was superior to other methods with an accuracy of 84%-93%. Meanwhile, LR achieved an accuracy of 79.82% - 85.60% and the SVM achieved 79.95% - 89.14%.

The model to predict the employability of IT graduates was developed by [24] using nine predictor variables which include demographic data, academic scores and study tracer data. The LR method provided the best accuracy, which was 78.4% compared to the DT and NB methods.

DT is the most widely used prediction method [2], [22], [25]. Some of the reasons that underlie the use of the DT method are that this method can be applied to both numerical and categorical data [15], [19], has a simple working method [15], and the resulting rules is easy to understand and interpret [15], [17], [19].

The advantages of the SVM method are that it is considered suitable to be applied to small datasets [7], [25], works faster than other methods [15], [25], and eliminate the problem of overfitting [26]. The RF method is considered to be more accurate, the time taken to develop the model is very less and can handle high dimensional data [26]. Meanwhile, LR method can be applied to predictor variables with nominal, ordinal, interval or ratio measurement scales [26] and good for prediction of success in a course [27].

Evaluation of model performance is calculated using various measures, namely accuracy [1], [9]–[11], [15], [19], [21], Recall [1], [8]–[10], [18], [19], Precision [1], [8]–[10], [18], [21], F1-measure [8], [9], [18], Mean Square Error (MSE) [3], Mean Absolute Error (MAE) [3], [12], Area under the Curve (AUC) [18], [19], [21], and Kappa [17]–[19]. The most widely used measure is accuracy. According to [22] the three prediction methods that provide the best accuracy are Neural Network (97.00%), RF (96.01%), and DT (92.80%).

III. RESEARCH METHOD

The research began with collecting the required data from the Academic Information System of higher education. After the data were collected, then pre-processing was carried out to clean the data from irrelevant data and perform data transformation. Data transformation is a process to eliminate differences in the dataset, so that the dataset has a format that is suitable for the algorithm for data mining [2].

In the next stage, the prediction model was built using the DT, RF, SVM, and LR methods. The performance of the prediction model that had been built was tested and then evaluated using a confusion matrix. The methodology used for the formation of the prediction model can be seen in Fig. 1.

A. Data Collection

The dataset used was taken from data of 309 participants from six courses that were opened in the 2020-2021 academic year at a higher education in Surabaya. Course participants come from five programs from the 2015-2016 to 2020-2021 academic years. The six courses used in this research are compulsory subjects in semester 2 to semester 4. Four out of six courses are in the area of mathematics and statistics, while the remaining two courses are in the area of programming. One of the courses was opened in four parallel classes, taught by different lecturers.

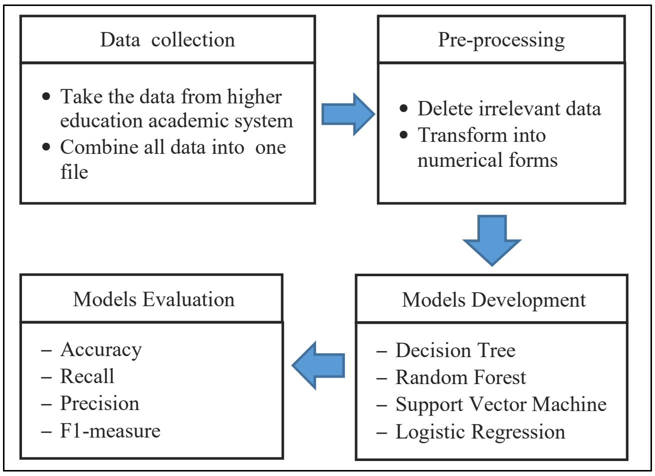

Fig. 1. Research method

There were thirteen predictor variables used for prediction in the first stage, namely X_1 to X_{13}. Meanwhile, in the second stage, fifteen predictor variables were used, namely X_1 to X_{15}. The list of predictor variables used in the research can be seen in Table I. The output of the prediction model was divided into two types of labels, namely "Pass" and "Fail". The dataset used for prediction consisted of 63 instances labeled “Fail” and 246 instances labeled “Pass”. This condition indicated that the dataset used was imbalanced.

The predictor variables used in this study were taken as a response from previous research and from the characteristics of students who were thought to have a significant influence on the prediction results. The predictor variables taken as a response from previous research can be seen in Table I, the fifth column.

B. Pre-processing

The first step was to delete irrelevant data, namely semester and academic year, student ID, parallel course and class ID, and lecturer ID. The process was continued by doing the transformation, which included changing all the data into numeric form. The transformation was done so that the data could be processed according to the algorithm used.

C. Models Development

Prediction models were formed using four different methods, namely: LR, DT, SVM, and RF. The formation of the prediction model was carried out through a training and testing process. However, the dataset used in this research was an imbalanced dataset. Imbalanced means that the number of samples of an output label differs greatly from the number of samples of other labels. The existence of imbalance data can have a negative impact on the performance of data mining algorithms [2]. To overcome this, the training and testing process was carried out using the ten folds stratified cross validation method. The data were divided into ten equal parts with the number of each label in each section being balanced. After that, one part was used for testing and nine parts were used for training. This process was be repeated ten times with different testing data for each iteration.

TABLE I. PREDICTOR VARIABLE USED

Predictor Variable	Code	Description	Scale	Reference
Study program	X_1	Study program taken by student	Nominal	[8], [13], [21]
Sex	X_2	Gender (M, F)	Nominal	[1], [8], [21]
Year of entry	X_3	Year of student entry	Interval	[13]
Cumulative credits	X_4	Credits having been collected from the first semester, including E score	Ratio	[1]
CGPA	X_5	Cumulative GPA including E score (0-4)	Interval	[1]
Number of courses taken	X_6	Number of courses taken this semester	Ratio	[21]
Number of participants	X_7	Number of participants in class	Ratio	[21]
Previous course grade	X_8	The score if it had been taken before (A, AB, B, BC, C, D, E, T). T means never been taken before.	Ordinal	
Length of time repeat	X_9	How many semesters it has been since the course was taken (0, 1, 2, …)	Ratio	
Position difference	X_{10}	Difference between the student position at current semester and the position of the course in the curriculum (0, 1, 2, ..)	Ratio	
Pass percentage	X_{11}	The percentage of passing of this course in previous semester by the same lecturer (0-100)	Ratio	
Prerequisite grade	X_{12}	Score of pre-requisite course taken in the latest.	Ordinal	[20]
Prerequisite time taken	X_{13}	How long has the pre-requisite course been?	Ratio	[20]
Mid-term test	X_{14}	Mid-term test score (0-100)	Ratio	[1], [11]
Number of attendance	X_{15}	Number of attendance before mid-term test (0-6)	Ratio	[1], [10]

D. Evaluation

Evaluation of the performance of the prediction model was carried out using a confusion matrix [6]. The performance measure was based on the confusion matrix that is usually used is accuracy. Accuracy is the ratio between the number of correctly predicted results compared to the number of data [1], [6]. Calculation of accuracy can be done using (1).

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

True Positive (TP) is the number of students who are predicted to "Fail" from students who "Fail" [1], [14]. True Negative (TN) is the number of students who are predicted to "Pass" from students who "Pass" [1], [14]. False Positive (FP) is the number of students who are predicted to "Pass" from students who "Fail" [1], [14]. False Negative (FN) is the number of students who are predicted to "Fail" from students who "Pass" [1], [14].

However, if the dataset is imbalanced, then the accuracy measure alone can be misleading. Therefore, it is necessary to use other measures, namely Recall, Precision, and F1-measure [14]. Recall is the probability that the model can classify students who "fail" correctly [14], [18]. Recall calculation can be done using (2). Precision is the probability that the model can classify students who are predicted to "fail" correctly [14], [18]. Precision calculation can be done using (3). Using the Recall and Precision measures separately can lead to biased comments in the evaluation of the prediction model [14]. F1-measure is the average of the harmonics between Recall and Precision [8], [14], [18]. The F1-measure calculation can be done using (4).

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1 - measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

IV. RESULTS AND DISCUSSIONS

Training and testing at each stage was carried out on 309 existing data using ten folds stratified cross validation. The results of testing in the first stage can be seen in Table II. Table II shows that the LR method is superior to other methods in all performance measures except the precision value. The precision value of the LR method is 3% lower than the precision value of the DT method. The model formed by the LR method was able to correctly predict 91.59%. Even though the dataset was imbalanced, the model was able to predict students who "fail" well. This can be seen from the Recall and the F1-measure values generated from the prediction using the LR method of at least 0.76. The performance shown by the LR method in this research is better than the results in [23], [24], [27], which showed the performance of the LR method ranged from 0.697 to 0.856. This proves that the LR method is good for the prediction of success in a course as stated by [27].

TABLE II. TESTING RESULT IN THE FIRST STAGE

	Accuracy	Recall	Precision	F1-measure
DT	91.58%	0.71	**0.86**	0.77
RF	90.29%	0.72	0.81	0.75
SVM	87.4%	0.45	0.79	0.57
LR	**91.59%**	**0.76**	0.83	**0.79**

The model using the DT and RF methods produce a performance that did not differ much when compared to the LR performance, which was between 1.3% to 5%. These results are in line with the literature review conducted by [22] which states that the DT and RF methods are two of the three prediction methods that are able to provide high accuracy. While the classification performance of the model formed by the SVM method was the lowest value compared to other methods. However, the accuracy of the prediction results using the SVM method was actually quite good, 87.4% even though the Recall value generated from the SVM method was 0.45. This showed that the SVM method was only not so good at predicting the minority class from the dataset used in this research.

All methods provide a minimum prediction accuracy of 87.4% and a minimum Precision of 0.79. It can be concluded that the academic performance of students in a course can be predicted at the beginning of the class. By using the data collected in this research, similar characteristics (particularly the CGPA, the suitability of taking courses with the curriculum, and mid-term test score) of the participants in the area of mathematics and statistics and in the area of programming show the similarity of students' academic performance. Therefore, the academic performance of the participants in the area of mathematics and statistics and in the area of programming can be predicted before the class starts based on the performance of the participants in the previous period based on the similarity of the student characteristics. Even though the predictions were made before class started, the resulting accuracy (91.59%) was better than [14] which made predictions starting in the third week of lectures (74%).

Since predictions can be made before class starts, then it provides opportunities for students, lecturers, and institutions to prepare strategies and take neccesary actions so that students who were predicted to fail can succeed at the end of the class. Therefore it will increase the course passing rate.

The results of testing in the second stage can be seen in Table III which shows that for all performance measures, the LR method still gives better results than other methods. However, the performance of the model formed by the DT and RF methods does not differ much from the LR performance. When viewed from the value of accuracy and precision, all methods provide a minimum accuracy of 88.69% and a minimum precision of 0.8. This showed that all methods were able to predict well. In addition, there was also an increase in the classification performance of the SVM method. So, it can be used to monitor student academic progress as well as to prepare the necessary actions.

TABLE III. TESTING RESULST IN THE SECOND STAGE

	Accuracy	**Recall**	**Precision**	**F1-measure**
DT	92.23%	0.71	0.90	0.77
RF	90.94%	0.75	0.82	0.77
SVM	88.69%	0.5	0.8	0.61
LR	**93.55%**	**0.78**	**0.9**	**0.83**

If the results of the first stage of performance are compared with the results of the second stage of performance, it can be seen that there was an increase in scores in all performance measures or at least the same in the second stage. Increases ranged from 0% to 7%. Only the Recall and F1-measure from the DT method did not increase from the first stage to the second stage. This shows that the performance of the model in the second stage is better than in the first stage. The tendency to increase the recall value in the second stage shows an increase in the ability of the model to make predictions for the minority class.

Prediction in the second stage was done by adding two predictor variables, namely X_{14} and X_{15}. Based on Table II and Table III, it can be seen that there was trend of increasing the value of each measure of model performance in the second stage. It showed that the addition of predictor variables can improve the performance of the prediction model, especially accuracy. This is in line with research conducted by [14], the performance of the prediction model will increase with the addition of data.

Based on the model formed using the DT method, it can be seen that the most significant predictor variables in the first stage prediction (taken from tree level zero to level two) were X_5, X_{10}, X_6, and X_{11}. While in the second stage (taken from tree level zero to level 2), the predictor variables that have a significant effect were X_5, X_{14}, X_6, and X_{10}. X_{14} was an additional predictor variable in the second stage. The position of X_{10} (tree level 1) and X_{11} (tree level 2) as predictor variables that have a significant effect in the first stage were replaced by X_{14}. This supports the previous conclusion which states that the addition of predictor variables can improve the performance of the prediction model. While X_5 remains the most influential predictor variable (tree level 0) both in the first and second stages. The tree of the prediction model formed by the DT method for the first stage can be seen in Fig. 2 and for the second stage it can be seen in Fig. 3.

```
|--X5 <= 1.50
|   |--X10 <= 7.50
|   |   |--X6 <= 1.50
|   |   |   |--class: 0
|   |   |--X6 >  1.50
|   |   |   |--X1 <= 0.50
|   |   |   |   |--class: 1
|   |   |   |--X1 >  0.50
|   |   |   |   |--X1 <= 1.50
|   |   |   |   |   |--class: 0
|   |   |   |   |--X1 >  1.50
|   |   |   |   |   |--class: 1
|   |--X10 >  7.50
|   |   |--class: 1
```

```
|--X5 >  1.50
|   |--X5 <= 2.50
|   |   |--X11 <= 2.50
|   |   |   |--X3 <= 5.50
|   |   |   |   |--class: 0
|   |   |   |--X3 >  5.50
|   |   |   |   |--X1 <= 0.50
|   |   |   |   |   |--class: 0
|   |   |   |   |--X1 >  0.50
|   |   |   |   |   |--class: 0
|   |   |--X11 >  2.50
|   |   |   |--X6 <= 6.50
|   |   |   |   |--X12 <= 5.50
|   |   |   |   |   |--class: 0
|   |   |   |   |--X12 >  5.50
|   |   |   |   |   |--class: 0
|   |   |   |--X6 >  6.50
|   |   |   |   |--class: 0
|   |--X5 >  2.50
|   |   |--class: 0
```

Fig. 2. Model formed with DT method in the first stage

```
|--X5 <= 1.50
|   |--X14 <= 2.50
|   |   |--X6 <= 1.50
|   |   |   |--X9 <= 2.00
|   |   |   |   |--class: 0
|   |   |   |--X9 >  2.00
|   |   |   |   |--class: 1
|   |   |--X6 >  1.50
|   |   |   |--X8 <= 0.50
|   |   |   |   |--X13 <= 1.00
|   |   |   |   |   |--class: 0
|   |   |   |   |--X13 >  1.00
|   |   |   |   |   |--class: 1
|   |   |   |--X8 >  0.50
|   |   |   |   |--class: 1
|   |--X14 >  2.50
|   |   |--X10 <= 4.50
|   |   |   |--X6 <= 5.00
|   |   |   |   |--X1 <= 1.50
|   |   |   |   |   |--class: 0
|   |   |   |   |--X1 >  1.50
|   |   |   |   |   |--class: 1
|   |   |   |--X6 >  5.00
|   |   |   |   |--class: 0
|   |   |--X10 >  4.50
|   |   |   |--class: 1
```

```
|--X5 >  1.50
|   |--X5 <= 2.50
|   |   |--X14 <= 4.50
|   |   |   |--X11 <= 2.50
|   |   |   |   |--X3 <= 5.50
|   |   |   |   |   |--class: 0
|   |   |   |   |--X3 >  5.50
|   |   |   |   |   |--class: 0
|   |   |   |--X11 >  2.50
|   |   |   |   |--X6 <= 6.50
|   |   |   |   |   |--class: 0
|   |   |   |   |--X6 >  6.50
|   |   |   |   |   |--class: 0
|   |   |--X14 >  4.50
|   |   |   |--class: 0
|   |--X5 >  2.50
|   |   |--class: 0
```

Fig. 3. Model formed with DT method in the second stage

The results of the evaluation showed that students' academic performance can be predicted well at the beginning of the class using academic history data. This can be seen from the minimum predictive accuracy value of 87.4% (Table II and Table III). The results of predicting student academic performance at the beginning of the class provide opportunities for students, lecturers, and institutions to prepare strategies so that students who are predicted to "fail" can succeed well at the end of the class so that it will increase the course passing rate. Efforts made at the beginning of the class by considering the prediction results in the first stage can be evaluated using the prediction results in the second stage. The evaluation results provide additional opportunities for the students who are predicted to fail to prepare better for the rest of the semester. Lecturers can provide personal assistance to such students (especially to students who are predicted to fail in both the first and second stages) with additional classes or other actions. So the results of these two-stage predictions can be used by lecturer to monitor student academic progress as well as to prepare the necessary actions.

V. CONCLUSION

In this research, a two-stage prediction model was built to predict the passing of each student from the courses undertaken using four different methods: DT, RF, SVM, and LR. The first stage of prediction was carried out before the class starts, while the second stage was carried out after mid-term test, by adding two predictor variables, namely the mid-term test score and the number of presence in class. The accuracy of the prediction results for the first and second stages showed that all methods were able to predict student academic performance well, but the LR method gave the best performance.

Based on the comparison of the classification performances of each model in the first stage and the second stage, it was seen that there was an increase or at least the same value from each model performance measure in the second stage. This was confirmed by the tree formed from the DT method. The tree showed that in the second stage, mid-term test becomes one of the most influential predictor variables in determining the probability of a student passing or not from the courses taken. So the results of the second stage of prediction can be used to monitor student academic development, after actions are taken based on the results of the first stage of prediction.

The lowest accuracy and precision values obtained from this study were 0.79 while the maximum recall value was 0.78. This recall value can be increased by overcoming the imbalanced data condition, such as by using hybrid of the heuristic method with Synthetic Minority Oversampling Technique, which will be our future research.

REFERENCES

[1] N. Hutagaol and Suharjito, “Predictive modelling of student dropout using ensemble classifier method in higher education,” Adv. Sci. Technol. Eng. Syst., vol. 4, no. 4, pp. 206–211, 2019, doi: 10.25046/aj040425.

[2] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: literature review and best practices,” Int. J. Educ. Technol. High. Educ., vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.

[3] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, “Machine Learning Based Student Grade Prediction: A Case Study,” pp. 1–22, 2017, [Online]. Available: http://arxiv.org/abs/1708.08744.

[4] PDDikti, “Higher Education Statistics 2020,” pp. 81–85, 2020, [Online]. Available: https://pddikti.kemdikbud.go.id/publikasi.

[5] O. Daramola, O. Emebo, I. Afolabi, and C. Ayo, “Implementation of an Intelligent Course Advisory Expert System,” Int. J. Adv. Res. Artif. Intell., vol. 3, no. 5, pp. 6–12, 2014, doi: 10.14569/ijarai.2014.030502.

[6] I. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, “An artificial intelligence approach to monitor student performance and devise preventive measures,” Smart Learn. Environ., vol. 8, no. 1, 2021, doi: 10.1186/s40561-021-00161-y.

[7] L. M. Abu Zohair, “Prediction of Student’s performance by modelling small dataset size,” Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0160-3.

[8] T. Fahrudin, J. L. Buliali, and C. Fatichah, “Predictive modeling of the first year evaluation based on demographics data: Case study students of Telkom University, Indonesia,” Proc. 2016 Int. Conf. Data Softw. Eng. ICoDSE 2016, pp. 0–5, 2017, doi: 10.1109/ICODSE.2016.7936158.

[9] R. G. Santosa, Y. Lukito, and A. R. Chrismanto, “Classification and Prediction of Students’ GPA Using K-Means Clustering Algorithm to Assist Student Admission Process,” J. Inf. Syst. Eng. Bus. Intell., vol. 7, no. 1, p. 1, 2021, doi: 10.20473/jisebi.7.1.1-10.

[10] A. K. Hamoud and A. M. Humadi, “Student’s Success Prediction Model Based on Artificial Neural Networks (ANN) and A Combination of Feature Selection Methods,” J. Southwest Jiaotong Univ., vol. 54, no. 3, 2019.

[11] S. A. Kumar, “Efficiency of Decision Trees in Predicting Student ’ S Academic Performance,” pp. 335–343, doi: 10.5121/csit.2011.1230.

[12] Y. Lee et al., “Prescribing Deep Attentive Score Prediction Attracts Improved Student Engagement,” no. Edm, pp. 743–748, 2020, [Online]. Available: http://arxiv.org/abs/2005.05021.

[13] A. Elbadrawy and G. Karypis, “Domain-aware grade prediction and top-n course recommendation,” RecSys 2016 - Proc. 10th ACM Conf. Recomm. Syst., pp. 183–190, 2016, doi: 10.1145/2959100.2959133.

[14] G. Akçapınar, A. Altun, and P. Aşkar, “Using learning analytics to develop early-warning system for at-risk students,” Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0172-z.

[15] H. A. Mengash, “Using data mining techniques to predict student performance to support decision making in university admission systems,” IEEE Access, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[16] M. A. Al-Barrak and M. Al-Razgan, “Predicting Students Final GPA Using Decision Trees: A Case Study,” Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/ijiet.2016.v6.745.

[17] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” Comput. Electr. Eng., vol. 89, no. October 2019, pp. 1–10, 2021, doi: 10.1016/j.compeleceng.2020.106903.

[18] S. Helal et al., “Predicting academic performance by considering student heterogeneity,” Knowledge-Based Syst., vol. 161, pp. 134–146, 2018, doi: 10.1016/j.knosys.2018.07.042.

[19] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, “Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores,” Comput. Intell. Neurosci., vol. 2018, 2018, doi: 10.1155/2018/6347186.

[20] J. Xu, T. Xing, and M. Van Der Schaar, “Personalized Course Sequence Recommendations,” IEEE Trans. Signal Process., vol. 64, no. 20, pp. 5340–5352, 2016, doi: 10.1109/TSP.2016.2595495.

[21] M. A. Yehuala, “Application Of Data Mining Techniques For Student Success And Failure Prediction The Case Of DebreMarkos University,” Int. J. Sci. Technol. Res., vol. 4, no. 4, pp. 91–94, 2015.

[22] E. Tjandra, S. S. Kusumawardani, and R. Ferdiana, “Student Performance Prediction in Higher Education : A Comprehensive Review,” no. 2019.

[23] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, “Predicting academic performance of students from VLE big data using deep learning models,” Comput. Human Behav., vol. 104, 2020, doi: 10.1016/j.chb.2019.106189.

[24] K. C. Piad, M. Dumlao, M. A. Ballera, and S. C. Ambat, “Predicting IT employability using data mining techniques,” 2016 3rd Int. Conf. Digit. Inf. Process. Data Mining, Wirel. Commun. DIPDMWC 2016, no. January 2014, pp. 26–30, 2016, doi: 10.1109/DIPDMWC.2016.7529358.

[25] A. M. Shahiri, W. Husain, and N. A. Rashid, “A Review on Predicting Student’s Performance Using Data Mining Techniques,” Procedia Comput. Sci., vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[26] A. Salini, U. Jeyapriya, S. M. College, and S. M. College, “A Majority Vote Based Ensemble Classifier for Predicting Students Academic Performance,” Int. J. Pure Appl. Math., vol. 118, no. 24, pp. 1–11, 2018.

[27] L. Yan and Y. Liu, “An Ensemble Prediction Model for Potential Student Recommendation Using Machine Learning,” Symmetry (Basel)., vol. 12, no. 5, pp. 1–17, 2020, [Online]. Available: https://doi.org/10.3390/sym12050728.

Telkom University

UNIVERSITAS GADJAH MADA

MMU MULTIMEDIA UNIVERSITY

CERTIFICATE

OF APPRECIATION

This is to certify that

Susana Limanto

has participated as

PRESENTER

in

The 10th International Conference on Information and Communication Technology (ICoICT 2022)

"Intelligent and Autonomous E-services Toward a Hybrid Interaction Era"

Bandung, 2nd – 3rd August 2022

Dr. Vera Suryani
General Chair ICoICT 2022
Telkom University

Prof. Ts. Dr. Ong Thian Song
General Co-Chair of ICoICT 2022
Multimedia University

IEEE INDONESIA SECTION

IEEE Signal Processing Society Indonesia Chapter