PREDICTING STUDENTS’ PERFORMANCE IN BASIC ALGORITHMS PROGRAMMING IN AN E-LEARNING ENVIRONMENT USING DECISION TREE APPROACH

Jonas de Deus Guterres, Kusuma Ayu Laksitowening, Febryanti Sthevanie
Telkom University, Indonesia
Email: jonasddeus@student.telkomuniversity.ac.id, ayu@telkomuniversity.ac.id, sthevanie@telkomuniversity.ac.id

Abstract
Predicting the performance of students plays an important role in every institution to protect their students from failures and leverage their quality in higher education. Algorithm and Programming is a fundamental course for the students who start their studies in Informatics. Hence, the scope of this research is to identify the critical attributes which influence student performance in the E-learning Environment on Moodle LMS (Learning Management System) Platform and its accuracy. Data mining helps the process of preprocessing data in a dataset from raw data to quality data for advanced analysis. Dataset set is consisting of student academic performance such as grades of Quizzes, Mid exams, Final exams, and Final projects. Moreover, the dataset from LMS is considered as well in the process of modeling, in terms of constructing the decision tree, such as punctuality submission of Quizzes, Assignments, and Final Projects. Regarding the Basic Algorithm and Programming course, which is separated into two subjects in the first and second semester, thus the research will predict the student performance in the Basic Algorithm and programming course in the second semester based on the Introduction to programming course in the first semester. Decision Tree techniques are applied by using information gain in ID3 algorithm to get the important feature which is the PP index has the highest information gain with value 0.44, also the accuracy between ID3 and J48 algorithm that shows ID3 has the highest accuracy of modeling which is 84.80% compared to J48 82.34%.

Keywords: prediction, student performance, e-learning, data mining, decision tree

Introduction
Predicting student performances is common as institutional desire to alter their method of teaching in class and anticipate the failure of the students in the final exams as well as university rank (SN Vivek Raj and SK Manivannan, 2020). It is an early-detection concept that helps the researcher identify the critical point of some cases in the past that require improvement of the performances from machine learning in order to anticipate the risk that might occur in the future and reduce. Regarding (Siegel, 2016)
said, “Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions.” Hence, this technique is effective to minimize the problem (error) regularly once the implementation of testing occurs (Siegel, 2016). However, it has a big challenge for the lecturers to find the factors that impact a student's performance in the class whether excellent or poor (Ramesh, Parkavi, & Ramar, 2013). In the Physical Learning Environment (Offline Class), the lecturer can observe and identify the student's performance easily because of direct contact among them, such as student participation and activation in class (Ng, 2021).

In research (Alturki, 2016) observe that the percentage of drop out in every course get more than 50% in computer science majors due to soft skills in programming when the student undervalued the demand of requirement needed in manual guidance. Moreover, it affirms that the student performance in introductory programming is based on competence in coding, self-determination, and the condition of study between students and lecturers. Especially in the pandemic situation of Covid-19, which is learning online as E-learning, students should have more self-dedication and concentration in programming courses because of its complexity and acquire the willingness of students to reach the goals needed from the university (Mbunge, Fashoto, & Olaomi, 2021). E-learning is a conventional method of educational process in the technology era which engages the lecturers and students in different sites apart from the classroom to conduct teaching activities (Adisurya, Santos, Fadhilah, & Lawanto, 2020). Likely face-to-face classes, learning in distance via platform room channels, by utilizing electronic and software tools such as laptop and E-learning websites, also remains effective to facilitate both the learners, as said in (Adisurya et al., 2020), to adjust this environment and learn intensively, and the lecturers to deliver the material concisely. E-learning has become the primary option for many institutions in Indonesia, after the spread of COVID-19 in the world, to assist the student study from home (Abidah, Hidaayatullah, Simamora, Fehabutar, & Mutakinati, 2020). Besides, it dramatically changes the environment and implies a negative impact for some universities that have not experienced an online academic system yet (Al Karim, Ara, Masnad, Rasel, & Nandi, 2021). Therefore, it influenced student performance due to the lack of technical knowledge and academic performance that they acquire. So, e-learning has replaced traditional learning since the outbreak of COVID-19, and lately, it has become a necessary method for students to incline with this environment (Moreno-Guerrero, Aznar-Díaz, Cáceres-Recche, & Alonso-García, 2020).

The academic institution supports their student in every possibility to guarantee the quality of student instead of failure (Altujjar, Altamimi, Al-Turaiki, & Al-Razgan, 2016). Thus, Telkom University has established the platform of learning management system (LMS) to run the online class and make the students comfortable to study with this condition. Hence, in Informatics Faculty, it provides several features which enable it to handle the activity of online class such as a code editor for a student to run the lab session, forum for discussing and forward a question to the lecturers, calendars, assignment, quizzes, and external resources as additional materials for them to be
equipped. Apart from it, there is a platform of LMS that does not involve the video conference feature to hold the online class. So, they will attend the class by utilizing another platform or application, such as Google Meet, Zoom, and Microsoft Teams, in order to attain the target.

![Figure 1](image)

**Figure 1**
The performance of a student in Basic Algorithm and Programming Course in curriculum 2016

Based on the source from Telkom University which shows in Figure 1, most students failed in the Basic Algorithm and Programming Course with index E during the previous curriculum of 2016. Meanwhile, research (Salim, Laksitowening, & Asror, 2020) proved that the time series prediction of students who graduate not on time in Telkom University using the K-nearest Neighbor algorithm generates 100% accuracy of student failure in Final Task 2, Final Defense Course, which become a tremendous problem for student due to inability of analysis in algorithm and result that mostly contains in this course. Related to this problem, this paper will focus more on algorithm courses to find the factor that influences the student performance especially in virtual learning methods for inexperienced students in an E-learning Environment (Full Online) which pertains to their performance in their study.

So, the scope of this paper is to purpose to the university by discovering the critical feature and predict the performance of students, during the pandemic situation of online class learning environment, in freshman year by using two algorithms of Decision Tree namely ID3 and J48. Moreover, the outcome could help the lecturers to improve the method of teaching and address the limitation of students through offering more personalized dedication to reduce student pressure (Chen, Luo, & Song, 2019). The data is collected from the faculty of Informatics, including grades and activities from E-learning Platform that used in Telkom University called LMS (Learning
Management System), to find a model in the training set of Introduction of Programming Course by using the ID3 and J48 algorithm to get the relevant information that can avoid the student from the risk of failing the course. The data from two subjects are combined into one training dataset and the second course becomes a label of the training set.

The reason for using the Decision Tree (DT) is because it is understandable and efficient to build due to a reasonable process that is simply converted into IF-THEN statements (Khakata, Omwenga, & Msanjila, 2019). Decidedly, DT is suitable to implement for this case in order to predict the student’s performance as well as a smaller dataset that will fulfill this final project. In addition, (Ababneh, 2019) research emphasizes the strength and weaknesses of those techniques include the Decision Tree Algorithm. Although Decision Tree (DT) is inefficient in implementing many variables, it can well-visualize the model of small data in a branch of a tree (Sunday, K., Ocheja, P., Hussain, S., Oyelere, S. S., Samson, B. O., & Agbo, 2020).

Method

This research method section uses the following process starts from data preparation, followed by selection and transformation of data, and end with Classification Model construction by identifying the critical feature.

Result and Discussions

This Phase explains the result and analysis of the algorithm testing using the ID3 and J48 algorithm that had evaluated by Information Gain and Entropy. The analysis aimed to find out the critical feature of attributes that affected student performance in the courses and compare accuracy between two algorithms given. Here, we construct two models of Decision Tree based on two scenarios which first include the feature of second-semester courses, second reduce the dimensionality by excluding the attributes from the Basic Algorithm Course in the second semester. The dataset consists of 487 student records that have been mined in data preprocessing.

A. Test Result

From Table 1, there is an improvement of students who passed increased to 77% compared to the previous curriculum in Telkom University. However, the number of students in the fail and risk category has above 20% which means the institute needs more effort to push the pass percentage to reach the overall rate of pass rate which is 82% (Bennedsen & Caspersen, 2019).

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Frequency of Grade Occurance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Index Grade</td>
</tr>
<tr>
<td>Pass</td>
<td>375</td>
</tr>
<tr>
<td>Fail</td>
<td>60</td>
</tr>
<tr>
<td>Risk</td>
<td>52</td>
</tr>
<tr>
<td>Total</td>
<td>487</td>
</tr>
</tbody>
</table>
Apart of it, this result gave the output of the important feature by using the entropy to calculate the gain information for finding the root of the tree which defines as a critical attribute. In the following was the process of setting the information gain $A$ relative to $S$ of 487 data comprising:

$$\text{Entropy}(S) = -p_{\text{pass}} \log_2(p_{\text{pass}}) - p_{\text{risk}} \log_2(p_{\text{risk}}) - p_{\text{fail}} \log_2(p_{\text{fail}})$$

$$= -\frac{375}{487} \log_2\left(\frac{375}{487}\right) - \frac{52}{487} \log_2\left(\frac{52}{487}\right) - \frac{60}{487} \log_2\left(\frac{60}{487}\right) = 1.0071$$

Afterward, the root of the tree is determined by applying the information gain and entropy method using the subset of FSG in Table 3.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Subset of FSG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSG</td>
</tr>
<tr>
<td>Class</td>
<td>A</td>
</tr>
<tr>
<td>Pass</td>
<td>234</td>
</tr>
<tr>
<td>Risk</td>
<td>14</td>
</tr>
<tr>
<td>Fail</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>255</td>
</tr>
</tbody>
</table>

$$\text{Gain}(S, FSG) = \frac{|S_E|}{|S|} \text{Entropy}(S_E) - \frac{|S_B|}{|S|} \text{Entropy}(S_B)$$

$$= \frac{|S_E|}{|S|} \text{Entropy}(S_E) - \frac{|S_B|}{|S|} \text{Entropy}(S_B)$$

$$\frac{|S_E|}{|S|} \text{Entropy}(S_E) = \frac{15}{487} \left( -\frac{9}{15} \log_2\left(\frac{0}{15}\right) - \frac{1}{15} \log_2\left(\frac{1}{15}\right) - \frac{14}{15} \log_2\left(\frac{14}{15}\right) \right) = 0.0109$$

Other attributes of FSG were also calculated, then continued to Gain $(S, FSG)$ formula as mentioned above (can use Equation 3) to generate the final gain value of every subset in Table 2 and the result in Table 3.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Information Gain Value of ID3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain</td>
<td>Value</td>
</tr>
<tr>
<td>FSG</td>
<td>0.4433</td>
</tr>
<tr>
<td>FP</td>
<td>0.3337</td>
</tr>
<tr>
<td>PRET</td>
<td>0.3178</td>
</tr>
<tr>
<td>POSTT</td>
<td>0.2994</td>
</tr>
<tr>
<td>CE</td>
<td>0.3333</td>
</tr>
</tbody>
</table>
Here, the result of the modeling of the Decision Tree using ID3 and J48 algorithm that shown in Table 4.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Include Attribute of SSG</th>
<th>Exclude Attribute of SSG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>ID3</td>
<td>0.8222</td>
<td>0.707</td>
</tr>
<tr>
<td>J48</td>
<td>N/A</td>
<td>0.811</td>
</tr>
</tbody>
</table>

B. Result Discussion

In the result illustrates the different critical features based on calculation in Table 3 which shows the FSG has the highest gain value, with 0.4433, that influenced student performance. On one hand, in Table 4 ID3 has better accuracy approximately 88% which includes the attribute SSG like lab works and quizzes. In contrast, J48 has lower accuracy of 0.81 when adding the dimensionality, and the score for precision and f-measure does not show due to risk class having 0 correctly classified.

Figure 2

Decision Tree From Forecasting Student Performance Classification.

Figure 4 (a) (exclude second courses attribute) shows that the student was at the risk if they obtained index C in FSG attribute because, in order to pass the second subject, at least the pre-test should be Good in Index C. Else it could be led to failure. However, in Figure 4 (b), there is no risk class instead of the passed and failed because of incorrect classification. Therefore, the tree without second-course attribute can be applied in the institution because it gives high accuracy and the root node between ID3 was similar even though ID3 has the highest accuracy by adding the attributes of SSG.
Conclusions

The method of forecasting student performance is essentially important to detect an early failure in the institution despite having excellent methods and solutions for students in an online environment. The result of this study has demonstrated the importance of un HIDING information that can be summarized. Hence, this study supports Telkom University to assist the performance of students because it proved by giving good accuracy 84% and creating a good model to do the prediction for the student in next semester. The following research will collect more attributes as well as other courses related to programming in order to maximize the accuracy and detect deeply students’ behaviors, therefore the university can handle the student’s issues that interrupt them to develop.

Notation

\( n, v \): The number of class and attribute

\( i, j \): The value of i-th and the weight of the partition.

\( p \): Probability of nonzero.

\( A \): Node of partition branch.

\( S \): Entropy of S.
BIBLIOGRAFI


Siegel, E. (2016). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Google Scholar
