CLASSIFICATION SYSTEM FOR STUDENT STUDY DURATION ON DEPARTMENT OF INFORMATION SYSTEMS AT MUSAMUS UNIVERSITY, USING ID3

Nasra Pratama Putra and Gerzon Jokomen Maulany
Department of Information Systems, Faculty of Engineering, Universitas Musamus, Merauke, Indonesia

ABSTRACT

Student graduation data is important data in the accreditation process. The data can be used for both Universities, Faculties, and Departments. Data about students who graduate can provide useful information if utilized optimally. Therefore, this study aims to produce a system that can classify the length of the student’s study period by utilizing the data of students who have graduated. The technique used in this study is data mining techniques. The data mining method used is the decision tree method, where the algorithm used is the ID3 algorithm. Some criteria used as attributes are gender, age at entry, place of birth, and GPA. The data sample used is graduate data in the Department of Information Systems at Musamus University. The total number of samples used was 131 with details of 92 training data and 39 test data. The results of this study are in the form of a system built using the PHP programming language and MySQL database. In the test, algorithm ID3 in this study produced a prediction of the length of study time for students with precision values of 66.67%, recall 46.15%, and accuracy of 74.36%.

Keywords: ID3, classification, student study duration, information systems

Cite this Article: Nasra Pratama Putra and Gerzon Jokomen Maulany, Classification System for Student Study Duration on Department of Information Systems at Musamus University, Using Id3, International Journal of Mechanical Engineering and Technology, 9(12), 2018, pp. 878–885

http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=9&IType=12

1. INTRODUCTION

The Study Program is required to have good performance. Good performance is judged by the increasing number of graduates in proportion to the number of students accepted. This situation causes the ratio of lecturers to students to be maintained. The more students accepted in higher education must be proportional to the number of graduates in each year. Students as entities or objects are important aspects of a university both operationally and strategically (Himawan, 2011). Based on the regulations presented in book II the standards and procedures for accreditation of tertiary institutions by BAN-PT (National Higher Education Accreditation...
Agency) in 2011 stated that one aspect of the accreditation assessment was students and graduates. Especially regarding the standard evaluation of students and graduates, the components assessed are the recruitment system for new students, and graduates (average study and GPA) (Muarif, 2013). Graduation data can be an evaluation program for study programs which will later become input for study programs and academic supervisors. Evaluation is done so that it can provide treatment to students who have problems. The aim of the treatment is expected to improve the performance of graduates. Because a system is needed that can process graduate data and produce an old classification of student studies. This classification can be used to help study programs in predicting and taking strategic steps.

Prediction is a scientific process to obtain knowledge systematically based on physical evidence. Predictions on student graduation can be used by majors to take anticipatory actions. One of the anticipatory actions that can be taken is the appointment of an appropriate supervisor, the intensity of the implementation of academic guidance and other anticipations. One technique for making predictions is data mining techniques. Data mining is a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and related knowledge from various large databases (Kusriini and Luthfi, 2009). Meanwhile, classification is a process of finding a model (function) that describes and distinguishes a data class or concept that aims to be used to predict classes of objects whose label class is unknown (Han et al., 2011). The classification process is based on four components, namely class, predictor, dataset training and testing dataset (Gorunescu, 2011).

Many algorithms in data mining can be done for classification, one of which is using the Decision Tree. Decision tree learning is a method that seeks to find functions that have a discrete value and are resistant to data that has errors (data noise) (Ariestya et al., 2016). This method is able to learn disjunctive expressions. Some types of decision tree learning are Iterative Dichotomiser 3 (ID3), Assistant and C4.5. Comparison of the use of other algorithms other than the decision tree to classify using various datasets has been done (Jadhav and Chanme, 2016). The study compared the performance of the K-NN method, Naïve Bayes, and decision tree. The results of this study indicate that the decision tree is the fastest performing method, producing more accurate decisions and having a low error rate.

The same study was also conducted to test the performance of using several decision tree algorithms (Pal and Pal, 2013). The study predicts student academic performance. The variables used are grades at secondary school, gender, parents' income and so on. The research concludes that ID3 decision tree algorithm which has the best accuracy and C4.5 algorithm can learn effectively to make a model in the case of student value predictions accumulated from previous years. Regarding student performance, ID3 algorithm has also been successfully used in forming decision trees for data of students who have passed lectures for one year or the first two semesters (Amalia and Naf’an, 2017). The results of the study are used to classify new students who need to get matriculation.

Based on several studies that have been done (Latuheru and Sahupala, 2018; Maulany et al., 2018; Waremra and Bahri, 2018), it can be concluded that the ID3 algorithm provides the best results. In this study, data mining techniques were used by applying algorithm ID3 to student graduation data. The results of this study are in the form of decision trees or a set of rules that can be used to predict the length of study of students by using data analysis of student graduation in the Department of Information Systems at Musamus University.

2. METHODS
The method used in developing data mining is CRISP-DM. CRISP-DM is a method formed by the European commission in 1996 that applies standards in the process of data mining. In CRISP-
DM there are six phases that will be carried out in data mining development research in accordance with the illustration in Figure 1.

1. Business Understanding

There are problems in predicting student graduation in the Department of Information Systems at Musamus University.

2. Data Understanding

The main data source used in this study is a dataset of students at the Musamus University Information System who have graduated from the 2006 class to the 2012 class. The student dataset consists of NPM attributes, Name, Department, Faculty, Place of Birth, Date of Birth, Address, No Mobile, Force, Thesis Title and GPA.

Some attributes are carried out in the data classification process, including GPA grouped into 3 categories (> 3.50, 3.00-3.50, <3.00), Birthdays are grouped into 2 categories (merauke, outside merauke), Age Ssaat Enter is grouped into 3 categories (<= 20 years, 21-30 years,> = 31 years), Graduation is grouped into 2 categories (<= 5 years,> 5 years), Classification of student attributes can be seen in table 1 below.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Man, Woman</td>
<td>Discrete</td>
</tr>
<tr>
<td>Age at entry</td>
<td>&lt;=20 years, 21 – 30 years, &gt;=31 years</td>
<td>Discrete</td>
</tr>
<tr>
<td>Place of birth</td>
<td>Merauke, Outside of Merauke</td>
<td>Discrete</td>
</tr>
<tr>
<td>GPA</td>
<td>&gt;3,50, 3,00-3,50, &lt;3,00</td>
<td>Discrete</td>
</tr>
</tbody>
</table>
1. Data Preparation

The amount of data to be used is 131 data, 92 training data and 39 test data. To get better quality data, there are several preprocessing techniques used, namely data validation and data size reduction and discretization. But in this study all data will be tested without reduction.

2. Modeling Phase

This stage is also called the learning stage. At this stage the data trained by the classification model then produces a number of rules. In this study, modeling using the ID3 algorithm.

ID3 is a classification approach in data mining by creating trees based on existing attributes to overcome a problem. A decision tree is a tree where each branch of a node represents an alternative choice and each end of the node / node represents a decision. ID3 takes the concept of information theory where the selection of attributes to form a tree is done by a statistical property called information gain. The gain value is used to measure the quality of an attribute in separating example training into the target class (Kristanto, 2013). In determining root (root) and branches, ID3 uses the largest information gain value of the existing attributes. Information gain is obtained from the formula:

\[
Gain(S,F) = Entropy(S) - \sum_{f \in \text{values}(F)} \frac{|S_f|}{|S|} Entropy(S_f)
\]  

(1)

Entropy is the amount of information contained in the attributes obtained from the formula:

\[
Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i
\]  

(2)

1. Evaluation Phase

In the evaluation phase testing of models is done to obtain accurate model information. Evaluation and validation using confusion matrix.

The confusion matrix method represents the results of model evaluations using matrix labels, if the dataset consists of two classes, the first class is considered positive and the second class is considered negative. Evaluation using confusion matrix produces values of accuracy, precision and recall.

<table>
<thead>
<tr>
<th>Correct Classification</th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>True positives (tp)</td>
</tr>
<tr>
<td></td>
<td>False negatives (fn)</td>
</tr>
<tr>
<td>-</td>
<td>False positives (fp)</td>
</tr>
<tr>
<td></td>
<td>True negatives (tn)</td>
</tr>
</tbody>
</table>

The following is the equation of the confusion matrix model:

Precision is used to measure how much the proportion of positive data classes are successfully predicted correctly from the overall positive class prediction results, which are calculated using equations:

\[
\text{precision} = \frac{tp}{tp+fp}
\]  

(3)

Recall is used to show the percentage of positive data classes that are successfully predicted correctly from the overall positive class data, which is calculated by the equation:

\[
\text{recall} = \frac{tp}{tp+fn}
\]  

(4)

Accuracy is the number of data comparisons that are correct with the total amount of data. Can be calculated using equations:

\[
\text{accuracy} = \frac{tp+tn}{tp+tn+fp+fn} \times 100\%
\]  

(5)
2. Deployment

After forming the model and analyzing and measuring it in the previous stage, then at this stage the most accurate model is also applied to determine the classification of student graduation predictions on time.

3. RESULTS AND DISCUSSION

3.1. Implementation

System implementation is based on the functional requirements listed in Table 3. Some of the tools used during implementation are using the PHP programming language and MySql database. The system is built in the form of a website because the php programming language is an open source programming language.

<table>
<thead>
<tr>
<th>No</th>
<th>SRS-ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SRS-Predik-01</td>
<td>Display all survey data (graduate data)</td>
</tr>
<tr>
<td>2</td>
<td>SRS-Predik-02</td>
<td>Partitioning training data and testing data</td>
</tr>
<tr>
<td>3</td>
<td>SRS-Predik-03</td>
<td>Make and produce decision trees</td>
</tr>
<tr>
<td>4</td>
<td>SRS-Predik-04</td>
<td>Perform and display results of performance measurements</td>
</tr>
</tbody>
</table>

The implementation of the functional displays all survey data (SRS-Predik-01) can be seen in Figure 2, next is the implementation of functional partitioning training data and test data (SRS-Predik-02) presented in Figure 3.

![ID3 CLASSIFICATION](image)

**Figure 2.** Functional implementation displays all survey data
Figure 3. Functional implementation partitioning training data and test data

Figure 4. Implementation of functional make and produce decision trees

The implementation of the functional make and produce decision tree (SRS-Predik-03) can be seen in Figure 4. Next is the implementation of the functional conducting and displaying the results of performance measurement (SRS-Predik-04) presented in Figure 5.
3.1.1. Interpretation

The pattern generated from the data mining process can be displayed in the form of a decision tree as shown in Figure 4. Rules for the decision can be taken as follows:

a) if (age _ when _ enter == <= 20 years AND GPA ==> 3.50 AND type _ sex == Male AND place _ born == outside Merauke) then> 5 years

b) if (age _ when _ enter == <= 20 years AND GPA ==> 3.50 AND type _ sex == Woman) then <= 5 years

c) if (age _ when _ entry == 21 - 30 years AND GPA ==> 3.50) then <= 5 years

d) if (age _ when _ enter == 21 - 30 years AND GPA == 3,00 - 3,50 AND type _ sex == Male AND place _ born == Merauke) then> 5 years

e) if (age _ when _ entered ==> = 31 years) then <= 5 years

3.1.2. Performance Measurement

Performance measurement is done to evaluate the decision tree that has been built previously as in Figure 4. Performance measurement is done by using confusion matrix. Confusion matrix is obtained by comparing the results of predictions from the system and actual results. To measure the performance of a decision tree, the first step is to count true positives, true negatives, false positives, and false negatives from each measurement, then enter the results into the confusion matrix.

The next step, calculate precision using equation (3), recall using equation (4), and accuracy using equation (5). Accuracy is defined as the level of closeness between the value of the predicted result and the actual value. Precision is defined as the measurement of accuracy. If the data is predicted positively, how often the prediction data is correct whereas recall is defined as measurement of completeness. From the actual amount of data that is positive, how much data is predicted to be positive

The id3 performance test was conducted on 39 data of test data, of which 13 had a decision <= 5 years, and 26 data had a decision> 5 years. From the test data, we get results like Table 4.

Table 4. ID3 Testing Results

<table>
<thead>
<tr>
<th>Amount of data</th>
<th>Classified as &lt;=5 years</th>
<th>Classified as &gt;5 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real decision :&lt;=5 years</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>Real decision : &gt;5 years</td>
<td>26</td>
<td>3</td>
</tr>
</tbody>
</table>
\[ \text{precision} = \frac{6}{6 + 3} \times 100\% = 66.67\% \]

\[ \text{recall} = \frac{6}{6 + 7} \times 100\% = 46.15\% \]

\[ \text{accuracy} = \frac{23 + 6}{13 + 26} \times 100\% = 74.36\% \]

4. CONCLUSIONS AND RECOMMENDATIONS

The conclusions that can be drawn from the results of this study are as follows:

1. From the results of the application of algorithm ID3 in the prediction of the length of study of students at the Musamus University Information System Study Program, it can be concluded that the most dominant attributes are age at entry, GPA, gender and the last is the place of birth.

2. The accuracy of data testing carried out on 39 test data is 74.36\%, with precision of 66.67\% and recall of 46.15\%.

Suggestions that can be made for further research are as follows:

1. It is necessary to do further analysis of the influential attribute data to increase the level of testing.

2. An analysis of the old study period decision needs to be done again, by adding more than 2 forms of decision.

REFERENCES


Classification System for Student Study Duration on Department of Information Systems at Musamus University, Using Id3

