# Predicting outcomes in *Introductory Programming* using J48 classification

**Mewati Ayub & Oscar Karnalim**

Maranatha Christian University
Bandung, Indonesia

ABSTRACT: In a computer science (CS) major, *Introductory Programming* becomes a substantial course, which determines whether students can complete that major or not. This study evaluates the correlation between student data with the students' capacity to pass that course. Such correlation is exploited according to a data mining technique called J48. For each student, the work incorporates personal, prior education, admission and assessment data. Based on an evaluation of 41 pieces of student data, the national test score for mathematics in Indonesia presents the most promising attributes, followed by the admission test score. The results of this study are expected to provide a brief insight for CS lecturer and the university, so that they can handle emerging issues in CS education, especially, the low retention rate.

## INTRODUCTION

Introductory Programming is a course, which becomes a key factor for computer science (CS) students, especially first-year students who have just completed high school and wish to continue their journey through the CS major. After spending several weeks on the Introductory Programming class, some students realise that they can hardly understand programming logic and syntax due to their lack of computational thinking ability. Consequently, it may generate an issue in student retention in the CS major. Some students tend to give up on their CS major since they think that they have not assimilated such skills. In addition, despite the fact that most students may pass the course, they are not guaranteed to pass advanced programming courses in the next semester, if they do not pass this course with distinguished results. Recognising the difficulty of programming, a course is proportional to the assigned semester of a given course. The higher its assigned semester is, the more difficult its course material is.

There have been several pieces of research, which aimed to identify the issues experienced by students in learning programming, especially in the introductory programming course. Byrne and Lyons [1], and Bergin and Reilly [2] have studied some of the factors, which influence the success of novices in the introductory programming course; namely, previous computing experience and prior academic performance. According to Rountree et al, decision tree classifier is used to identify combinations of factors that interact to predict success or failure in the introductory programming course [3]. Also, a study by Wiedenbeck et al has shown that self-efficacy and the mental model have a direct effect on overall success in an introductory programming course [4]. A study by Wilson indicates that a formal class in programming and game playing *…promote success in an introductory computer science course, there is a significant gender difference particularly for game playing* [5].

In this study, prediction attributes that determine the success of novice students in computer programming will be explored further. Such prediction will be conducted using data mining technique, particularly the J48 classification technique. As a case study, this work incorporates student data from the Introductory Programming course, which was held in the Computer Science (Informatics) major, at Maranatha Christian University. The attributes cover four aspects, which are personal, prior education, admission and assessment data. The result of this study is expected to become a supplementary data source for handling the student retention issue. For instance, providing a more-sophisticated learning method or course syllabus based on high-valued prediction attributes. Moreover, these data can also be used by university admissions in terms of student recruitment. They can filter the students recruited, based on given prediction patterns.

## CLASSIFICATION IN DATA MINING

The data mining task can be divided into two categories, descriptive and predictive [6]. The classification method is a predictive data mining task, which is defined as a predictive method that is used to classify unseen data [6][7].

The predictive model is generated based on the analysis of a training data set. Prediction of a new data should be done using the model. There are some techniques that can be used for classification, such as decision tree induction, Bayes classification or rule based classification [6]. In this study, the authors applied decision tree induction as a classification technique; namely, J48 classification. The J48 classification is Weka's implementation of C 4.5 decision tree learner. Weka implements a later and slightly improved version, which called C4.5 revision 8 [7].

For each leaf in a decision tree derived from the J48 classification, there are two numbers (n/m), which mean that *n* instances reach the leaf, but *m* instances are classified incorrectly [7]. The percentage of correctly classified instances will determine whether the generated model is sufficiently good. When the amount of data for training and testing is limited, stratified tenfold cross validation techniques can be used to ensure the performance of the classifier [7].

Rules in the IF-THEN form can be extracted from a decision tree. Each path from the root to a leaf node can be written as one rule. The rule antecedent (IF-part) is formed by combining the splitting criterion along a given path using the *AND* connection. The leaf node which contains the class prediction forms the rule consequent (THEN-part) [6].

## METHODOLOGY

The class under study was a first-year course; namely, Basic Programming. It is based on the Python programming language in procedural style programming. All students had entered the course directly from high school. After the students have completed the course, they should be able to specify, design, code and test a computing solution. The course was composed of two sessions: theory of 150 minutes' duration and laboratory practice of 210 minutes' duration. These sessions were conducted once a week during that semester. In term of course material, Basic Programming was divided into logic and advanced-technique material.

Logic material was taught in the first seven weeks (before a mid-test). It covered data types, variables, conditional statements and looping. On the other hand, advanced-technique material was taught during the last seven weeks (after the mid-test). It covered many advanced techniques that are frequently used in programming. Such techniques include functions, arrays, searching and sorting. The assessment of this course consists of a mid-semester written examination (25% marks), a final written examination (25% marks), mid-semester and final laboratory examinations (25% marks), and 12 weekly laboratory assignments (25% marks).

In this case, the data set has been extracted from 41 students who have a minimum 75% of attendance for the Basic Programming course. Each piece of student data consists of personal, admission, prior education and assessment data. Pre-processing for these data was done by resolving inconsistencies, and transforming data to obtain quality data that are feasible for classification.

As described in Table 1, a group of attributes has been selected for student classification. These attributes consist of:

a) personal data, such as gender and student's home town (province);
b) prior education data, such as high school major, national test score for mathematics (NTM);
c) admission data, such as admission test score (ATS);
d) assessment data from basic programming course, such as final written examination score (WES) and final laboratory examination score (LES). The gender, province, major, NTM, ATS attributes will be utilised to predict WES or LES.

Table 1: Student data set.

| Attribute name | Description | Possible values |
|---|---|---|
| Gender | Student's gender | [M=male, F=female] |
| ATS | Admission test score | [E : excellent, G : good, F : fair] |
| Province | Student's home town | [J = Java, L = outside Java] |
| Major | High school major | [1 = major A, 2 = outside major A] |
| NTM | National test score for mathematics | [E : excellent, G: good, F : fair] |
| WES | Written examination score | [E : excellent, G : good, N : not passed] |
| LES | Laboratory examination score | [E : excellent, G : good, N : not passed] |

Based on the student data set, this study explored the final examination scores as class attributes against personal data, prior education data and admission data through classification technique. The experiment was performed twice, one for final written examination scores and the other for final laboratory examination scores.

## RESULTS AND DISCUSSION

In Table 2, the data set is grouped according to gender, major, province, ATS and NTM. Descriptive statistics (means and standard deviation) for each group are shown in Table 2.

Table 2: Descriptive statistics.

| Grouping | | n | WES | | LES | |
|---|---|---|---|---|---|---|
| | | | Means | SD | Means | SD |
| Gender | Male | 33 | 56.12 | 24.88 | 53.45 | 19.54 |
| | Female | 8 | 48.75 | 18.07 | 43.00 | 19.73 |
| Major | Major A | 28 | 59.00 | 20.65 | 55.00 | 17.54 |
| | Outside major A | 13 | 45.38 | 27.83 | 43.69 | 22.75 |
| Province | Java | 31 | 56.13 | 25.33 | 53.68 | 20.75 |
| | Outside Java | 10 | 50.20 | 18.07 | 44.40 | 15.23 |
| ATS | Excellent | 13 | 67.23 | 23.85 | 61.77 | 20.34 |
| | Good | 25 | 52.68 | 19.36 | 49.08 | 16.69 |
| | Fair | 3 | 17.00 | 10.58 | 26.00 | 16.37 |
| NTM | Excellent | 4 | 88.75 | 13.15 | 76.50 | 19.28 |
| | Good | 20 | 61.30 | 18.82 | 56.35 | 15.34 |
| | Fair | 17 | 38.88 | 18.44 | 39.71 | 17.20 |

To examine the differences between WES and LES for each of the group attributes, the authors used a *t*-test for gender, major and province. For example, the WES mean for male students was compared to the WES mean for female students. In carrying out the *t*-tests, assumptions of normality and equality of variances were confirmed. In each group, the *t*-tests suggested no significant differences between any of the factors and the final examination scores.

The authors inspected the differences between final examination scores for the ATS group and the NTM group using the one-way Anova test. The results of the Anova tests showed that there was a significance difference ($p$-value < 0.01) between the attributes' value in both groups. Table 3 indicates the F-value for each group with a critical F-value of 5.21. They used MaxStat Lite version 3.6 to calculate descriptive statistics, *t*-tests and Anova tests.

Table 3: F-value of Anova tests for ATS and NTM.

| Grouping | WES | LES |
|---|---|---|
| | F-value | F-value |
| ATS | 7.59 | 5.40 |
| NTM | 14.62 | 9.82 |

This study further explored the relationship between final examination scores with the group attributes using the J48 classification of the students' data set using WEKA version 3.8.1 as a data mining toolkit. The first classification was done using written examination score (WES) as the attribute class. The second classification used laboratory examination score (LES) as the class. Because of limited data, this study utilised a ten-folds cross validation to ensure the validation of the results.

The result of the first classification is shown in Figure 1 in the form of a J48 pruned tree with 70.73% accuracy in WES predicting. There are 29 correctly classified instances and 12 incorrectly classified instances. Figure 1 indicates that the NTM attribute is the most affective attribute in predicting written examination score (WES). The result of WES is accordingly determined based on NTM. If NTM is excellent, then, WES is excellent. Similarly, if NTM is fair, then, WES is not passed. However, if NTM is good, major and gender attributes also contribute to predict the WES. Table 4 resumes the rules generated from the J48 decision tree in Figure 1.

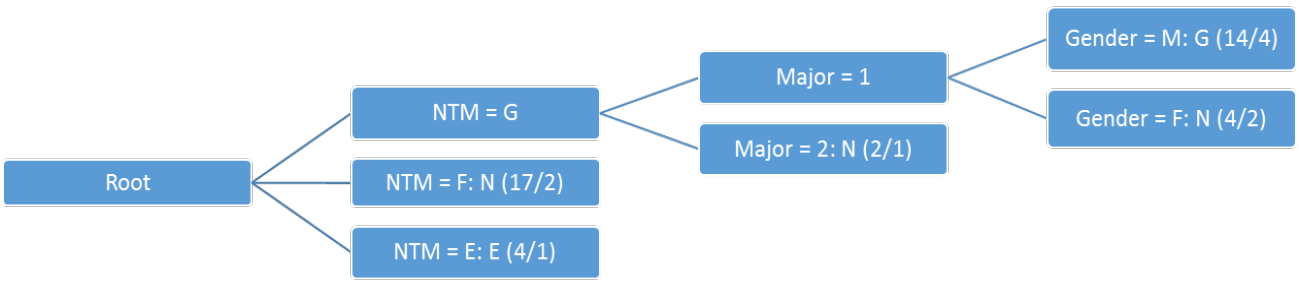


Figure 1: J48 pruned tree for WES predicting.

Table 4: Rule for WES.

| Rule # | Rule's premise | WES Percentages of instances | | |
|---|---|---|---|---|
| | | Excellent | Good | Not passed |
| 1 | IF NTM = Good and Major = 1 and Gender = Male | - | 71.43% | - |
| 2 | IF NTM = Good and Major = 1 and Gender = Female | - | - | 50% |
| 3 | IF NTM = Good and Major = 2 | - | - | 50% |
| 4 | IF NTM = Fair | - | - | 88.23% |
| 5 | IF NTM = Excellent | 75% | - | - |

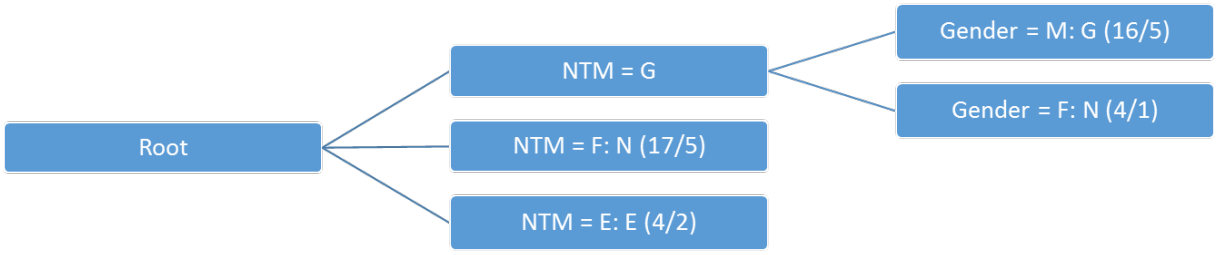


Figure 2: J48 pruned tree for LES predicting.

The result of the second classification is shown in Figure 2 in the form of the J48 pruned tree with 58.54% accuracy percentage of LES predicting. There are 24 correct classified instances and 17 incorrectly classified instances. Figure 2 indicates that the NTM attribute is also the most affective attribute in predicting the laboratory examination score (LES). The result of LES is accordingly determined based on NTM. If NTM is excellent, then, LES is excellent. Similarly, if NTM is fair, then, LES is not passed. Nearly the same as predicting WES, if NTM is good, gender attributes also contribute to predict the LES. Table 5 reports on the rules generated from the J48 decision tree in Figure 2.

Table 5: Rule for LES.

| Rule # | Rule's premise | LES Percentages of instances | | |
|---|---|---|---|---|
| | | Excellent | Good | Not passed |
| 1 | IF NTM = Good and Gender = Male | - | 68.75% | - |
| 2 | IF NTM = Good and Gender = Female | - | - | 75% |
| 3 | IF NTM = Fair | - | - | 70.59% |
| 4 | IF NTM = Excellent | 50% | - | - |

## CONCLUSIONS

Student success or failure in the Introductory Programming course will influence the continuity of their study in computer science. In this study, five attributes were used to predict outcomes in introductory programming: gender, student home town (province), high school major, national test score for mathematics (NTM) and admission test score (ATS).

Written examination score (WES) and laboratory examination score (LES) are used as class attributes. The statistical test results of this study show two predictive factors in the following order of importance: national test score for mathematics and admission test score. Using J48 classification, the authors obtained a confirmation that the national test score for mathematics becomes the most impactful prediction attribute. From the J48 decision tree, the combination of factors to predict success or failure can be resumed.

## ACKNOWLEDGMENT

The authors would like to acknowledge the financial support provided by the Maranatha Christian University Research Committee, by means of the Maranatha Christian University Grant.

REFERENCES


1. Byrne, P. and Lyons, G., The effect of student attributes on success in programming. *Proc. ITiCSE 6th Annual Conf. on Innovation and Technol. in Computer Science Educ*., New York, USA, 49-52 (2001).
2. Bergin, S. and Reilly, R., Programming: factors that influence success. *Proc. 36th SIGCSE Technical Symp. on Computer Science Educ*., Missouri, USA, 411-415 (2005).
3. Rountree, N., Rountree, J., Robins, A. and Hannah, R., Interacting factors that predict success and failure in a CS1 course. *Proc. ITiCSE Conf. on Innovation and Technol. in Computer Science Educ.*, Leeds, United Kingdom, 101-104 (2004).
4. Wiedenbeck, S., Labelle, D. and Kain, V.N.R., Factors affecting course outcomes in introductory programming. *Proc. 16th Annual Workshop of the Psychology of Programming Interest Group*, Carlow, Ireland, 97-110 (2004).
5. Wilson, B.C., A study of factors promoting success in computer science including gender differences. *Computer Science Educ.*, 12, **1-2**, 141-164 (2002).
6. Han, J., Kamber, M. and Pei, J., *Data Mining Concepts and Techniques*. Waltham: Elsevier Inc. (2012).
7. Witten, I.H., Frank, E. and Hall, M.A., *Data Mining Practical Machine Learning Tools and Techniques.* Burlington: Elsevier Inc. (2011).