# Predicting students' final passing results using the Classification and Regression Trees (CART) algorithm

3 authors:

Julianti Kasih
Universitas Kristen Maranatha
11 PUBLICATIONS   16 CITATIONS

SEE PROFILE

Mewati Ayub
Universitas Kristen Maranatha
24 PUBLICATIONS   84 CITATIONS

SEE PROFILE

Sani Susanto
Universitas Katolik Parahyangan
48 PUBLICATIONS   65 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Evaluation of Cognitive Impact of Program Visualization in Learning Programming View project

Information Technology Faculty of Maranatha University Bandung View project

# Predicting students' final passing results using the Classification and Regression Trees (CART) algorithm

**Julianti Kasih†, Mewati Ayub† & Sani Susanto‡**

Maranatha Christian University, Bandung, Indonesia†
Parahyangan Catholic University, Bandung, Indonesia‡

ABSTRACT: The aim of this research shares that of a previous research programme that involved a lecturer as an academic advisor helping students to predict their final passing results based on their performance in several subjects in early semesters of their study period [1]. Previous research used discriminant analysis, which is considered to be impractical in this instance. In this research, a more practical method is introduced and applied, based upon the classification of functions of data mining. The classification was performed using software based upon the Classification and Regression Trees (CART) algorithm.

INTRODUCTION

Ideally, this work is the fulfilment of the authors' promise to undertake research to provide academic advisors with a more practical way of predicting the final passing results of a student [1]. The passing results in the Indonesian education system are classified into three grades: Extraordinary (*Cum Laude*), Very Satisfactory and Satisfactory.

The arguments for making this kind of prediction are as follows:

- First, one of the important aims of higher education in the Republic of Indonesia is to prepare the academic participants (students) to become members of society with the academic and/or professional abilities to enable them to apply/develop/enrich the foundations of knowledge in the sciences, technology and the arts.
- Second, to achieve this aim, undergraduate students are assigned to several academic advisors (an informal translation of the Indonesian term *Dosen Wali*) throughout their years of higher educational studies.
- Third, academic advisors, who are lecturers, have as their main task the fostering of students' academic and non-academic activities. With regard to students' academic activities, one of the duties of the academic advisor is to help students in setting up their study plans for each semester.
- Fourth, setting up a study plan includes providing guidance for students regarding how many subjects, and which subjects, to undertake.
- Fifth, through this guidance, students are expected to obtain the best passing results at the end of their undergraduate study [1][2].

In the previous research, it was demonstrated that discriminant analysis helped academic advisors in the Faculty of Information Technology, University X in Bandung, West Java, Indonesia, to predict the final passing results of a student based on his/her performance in some subjects in the early stage, i.e. the first four semesters, of a higher education study programme. (Please note that for reasons of confidentiality, the full name of the institution has not been included). This facility enables academic advisors to assist students to set up their study plans for each semester, so that the students perform to their full potential [1].

This work aims to help the academic advisors with a more practical way of predicting the final passing results of a student. In this research, a data mining task called *classification* was employed. Classification is performed through a technique called Classification and Regression Trees (CART), which diagrammatically is presented in the form of *decision trees*. This kind of diagram tree serves in a more practical manner compared to the territorial map employed in the previous research [1].

OVERVIEW OF THE BACKGROUND THEORY

The *MIT Technology Review* chose data mining as one of ten emerging technologies that will change the world. Numerous definitions of data mining are available. However, Larose concisely defined data mining as *discovering knowledge in data* in the title of one of his publications [3]. Furthermore, classification is one of the six main tasks of data mining; the others being description, estimation, prediction, clustering and association.

Some important points about classification are as follows:

- In classification, there are two types of variables. The first is a *target variable*. This variable is categorical, which means it could be partitioned into several classes or categories. Second is a set of *predictor variables*. The classification task examines a large set of records, called the *training data*. Each record contains information about the target variable, as well as a set of input or predictor variables.
- The algorithm applied to carry out the classification task would, firstly, examine the training data set containing both the predictor variables and the (already classified) target variable. The algorithm *learns* which combinations of predictor variables are associated with which target variable. Then, the algorithm looks at new records, for which no information about the target variable is available. Based on the classifications in the training data set, the algorithm assigns classifications to the new records.
- There are two prominent algorithms in the area of classification: the Classification and Regression Trees (CART) algorithm and C4.5 algorithm. These two algorithms result in a *decision tree*, which is, basically, a collection of *decision nodes*, connected by *branches*, extending downward from the *root node* until terminating in *leaf nodes*.
- The CART algorithm produces a binary tree, while the C4.5 algorithm is not restricted to binary splits.
- The binary feature of the CART algorithm produces a very readable decision tree and for this reason is chosen for this research. The detailed calculation was performed using Clementine 10.1 data mining software [3].

EXPERIMENT: THE RESULT AND INTERPRETATION

As in the previous research programme, the academic transcripts from 146 alumni served as input data, which were available from the first author [1]. From this work, the students' *final passing results* are determined by the *final marks* from the following five subjects: IF102 (Introduction to Computer Application), IF103 (Introduction to Information Technology), IF104 (Algorithms and Programming), IF105 (Basic Programming) and IF202 (Linear Algebra and Matrices). The final marks of these five subjects take the role of *predictor variables*. The final marks for a subject are classified into five groups: A (High Distinction), B (Distinction), C (Credit), D (Pass) and E (Fail) with some intermediates such as B+ and C+.

Based on the passing results, the students' final passing results were classified into three groups: 1 - Extraordinary (*Cum Laude*); 2 - Very Satisfactory; and 3 - Satisfactory. The students' final passing results take the role of *target variable*.
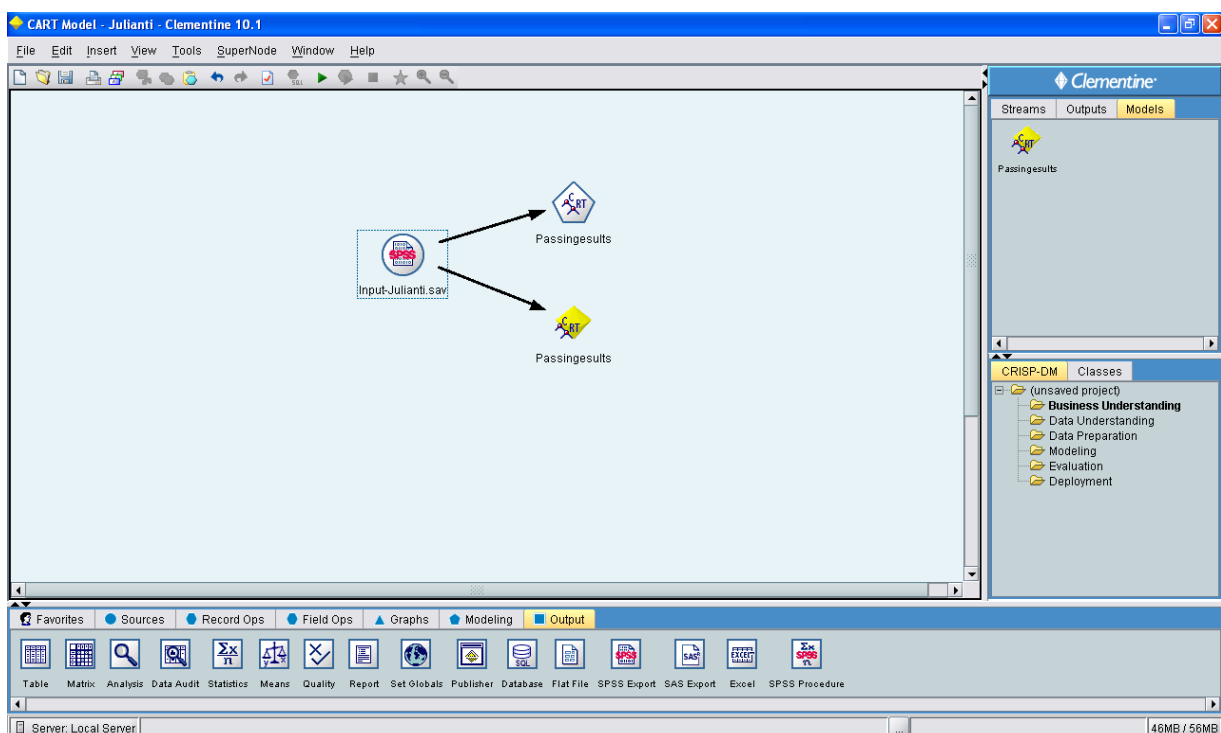


Figure 1: The classification model representation in Clementine 10.1.

The Clementine 10.1 model is represented in Figure 1. In this figure, the icon on:

- the left side describes the input data,
- the top right side describes the algorithm employed, in this case, the CART algorithm, and
- the bottom right side describes the output of the algorithm.

The input data are saved in the form of an SPSS worksheet file, a section of which is shown in Figure 2.



Figure 2: A part of the data file.

The result of this model in the form of a two-level decision tree is presented in Figure 3. An interpretation of this diagram follows.
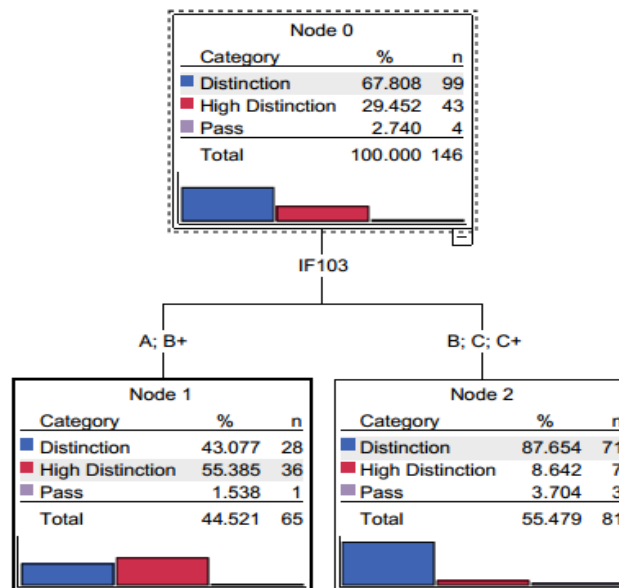


Figure 3: The two-level decision tree output.

Node-0 describes the following information: out of 146 alumni, the number of those with the final passing results High Distinction, Distinction and Pass are, respectively, 43 (29.452%), 99 (67.808%) and 4 (2.740%).

Node-1 branch represents a group of alumni who received an A or B+ grade for the subject IF 103 (Introduction to Information Technology). From the *training data*, it was predicted that future students belonging to this group would finish their studies with a High Distinction, Distinction and Pass, in their final passing results 55.385%, 43.077% and 1.538% probabilities, respectively. A similar interpretation applies to Node-2.

The decision tree can be extended into a longer version. For instance, a three-level decision tree for this research problem is given in Figure 4.
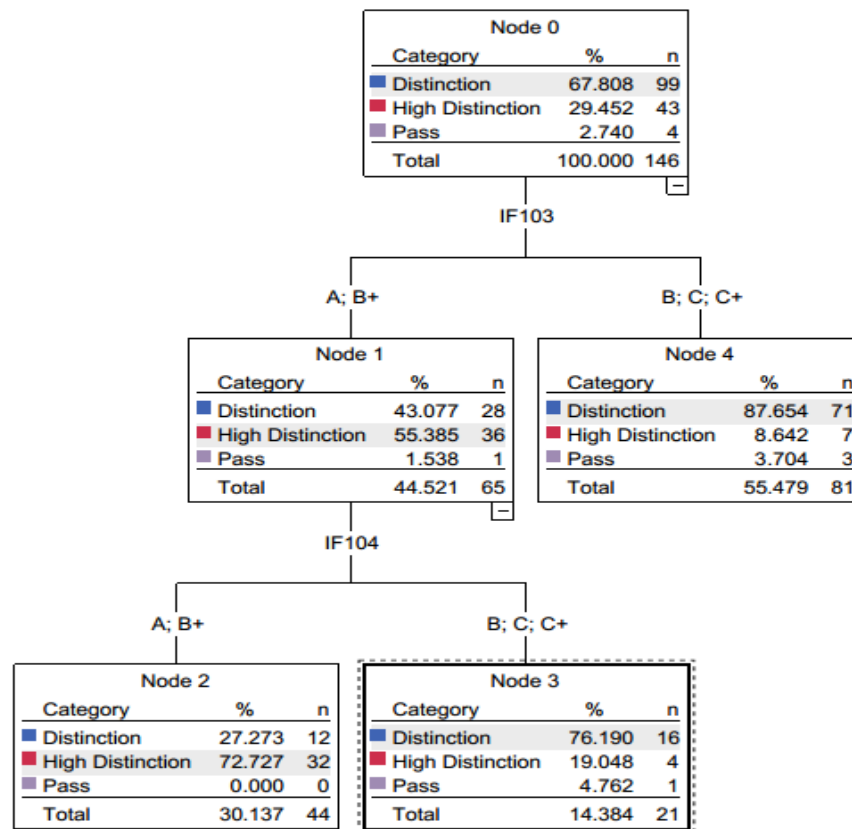


Figure 4: The three-level decision tree output.

## CONCLUSIONS AND SUGGESTION FOR FURTHER RESEARCH

This research demonstrates how academic advisors can predict the final passing results of students. Compared to the previous research, this demonstrates that the prediction can be performed in a more practical manner [1]. In the previous research, it was necessary to first generate the two discriminant functions and a territorial map, followed by the calculation of two discriminant scores for each student. Finally, the two scores were plotted on the map to find the corresponding group, which represents a group of students' final passing results. In this research, all this is represented by a tree diagram, called a decision tree.

Further research may take the form of the investigation of the prediction of the final passing results, which might be based on a data mining task called *association*. Association is performed through a technique called Market Basket Analysis. Such research will be undertaken in the not too distant future.

## REFERENCES

1. Kasih, J. and Susanto, S., Predicting students' final results through discriminant analysis. *World Transactions on Engng. and Technol. Educ.*, 10, **2**, 144-147 (2012).
2. The Republic of Indonesia. Government Regulation about Higher Education, 60 (1999).
3. Larose, D.T., *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc. (2005).