PAPER • OPEN ACCESS

A classification model for student exchange using CART algorithm

To cite this article: W B Zulfikar *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1098** 032054

View the article online for updates and enhancements.

You may also like

- Nitrogen and carbon limitation of planktonic primary production and phytoplankton–bacterioplankton coupling in ponds on the McMurdo Ice Shelf, Antarctica
Brian K Sorrell, Ian Hawes and Karl Safi
- Forecasting Student Graduation With Classification And Regression Tree (CART) Algorithm
A Maesya and T Hendiyanti
- On the inversion of the Radon transform on a generalized Cormack-type class of curves
G Rigaud

244th Electrochemical Society Meeting
October 8 – 12, 2023 • Gothenburg, Sweden
50 symposia in electrochemistry & solid state science
Deadline Extended!
Last chance to submit!
New deadline: April 21
submit your abstract!

This content was downloaded from IP address 202.80.212.22 on 19/04/2023 at 09:52

IOP Conf. Series: Materials Science and Engineering 1098 (2021) 032054 doi:10.1088/1757-899X/1098/3/032054

A classification model for student exchange using CART algorithm

W B Zulfikar*, I Taufik, A R Atmadja and R P Rahayu
Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia

*wildan.b@uinsgd.ac.id

Abstract. The university has a cooperative relationship with other universities including abroad. The cooperation program covers various fields, one of which is the academic field. Students have the opportunity to exchange information, science, and culture through student exchange programs. However, not all students are eligible to join this program because there are terms and conditions that must be met from various aspects such as academics, attitudes, and even financial conditions. The purpose of this study is to analyse and preprocess the training data and then model it in the form of classification using CART. Based on the test results, the proposed model provides satisfactory results with an accuracy percentage of 90%.

1. Introduction

Student Exchange is organized by certain parties for providing opportunities for students to study abroad within a certain period. This has many benefits, especially for students who are participants in the student exchange program. This program is effective for challenging students in developing a global perspective [1–4].

Currently, tertiary institutions in Indonesia are organizing student exchange programs. Based on observations, the discovery of facts that occur in the process of acceptance of student exchange participants. There were also unilateral resignations by prospective participants. As for the cause, the participants were not prepared technically and non-technically as well as about funding. The purpose of this study is to minimize participants who are not eligible to join the student exchange program. There are not all students in a university are entitled to participate in student exchange programs. The organizer usually provides a limited amount of quota. However, every student has the same opportunity to participate in the selection process. There are, many conditions that must be fulfilled by students including GPA scores, making English essays, interviews using English as well as reading the Qur'an test (optional).

Data mining has several function such as classification, clustering, and association [5,6]. CART (Classification And Regression Tree) is one classification method of data mining [7–11]. It will produce a classification tree if the response variables are categorical, and produce a regression tree if the response variables are continuous [7,12]. The main purpose of CART is to obtain an accurate group of data as a characteristic of a classification. The distinctive feature of the CART algorithm is that the decision node is always two-pronged or binary forked. In its implementation, a record will be classified into one of the many classifications available on the destination variable based on the values of the predictor variables.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

The proposed model uses CART to classify the feasibility of prospective student exchange program participants. Based on previous research, this method has advantages namely, the accuracy is very qualified [13–18] and this method widely used on several field [19–21].

2. Research methods

CART is the algorithm used in this study. CART is an algorithm of a data exploration technique, which is a decision tree technique. CART is a nonparametric statistical methodology developed for the topic of classification analysis, both for categorical and continuous response variables. The following right and left branch candidates will be used to make a decision tree shown in table 1.

Table 1. Left and right branch candidates.

No	Left Brach Candidate	Right Brach Candidate
1	GPA <3,3	GPA >=3,3
2	GPA <=3,5	GPA >3,5
3	Essay <70	Essay >=70
4	Essay <85	Essay >=85
5	Interview <85	Interview >=85
6	Qur'an <20	Qur'an >=20
7	Qur'an <35	Qur'an >=35
8	Income = Low	Income = Medium
9	Income = Medium	Income = High
10	Expertise Certificate = No	Expertise Certificate = Yes

The calculation of PL (Prior Left) and PR (Prior Right) is an implementation of equations (1) and (2).

$$P_l = \frac{The\ number\ of\ notes\ on\ the\ left\ candidate}{Number\ of\ notes\ in\ the\ training\ data} \quad (1)$$

$$P_r = \frac{The\ number\ of\ notes\ on\ the\ right\ candidate}{Number\ of\ notes\ in\ the\ training\ data} \quad (2)$$

$$P\ T_l = \frac{Number\ of\ records\ categorized\ on\ the\ left\ candidate}{The\ number\ of\ notes\ in\ the\ decree\ t} \quad (3)$$

$$P\ T_r = \frac{Number\ of\ records\ categorized\ on\ the\ right\ candidate}{The\ number\ of\ notes\ in\ the\ decree\ t} \quad (4)$$

In table 2 it can be seen that for the PL calculation results that are obtained from the amount of data that meets the criteria of the left branch candidate in the total amount of data. Then, PR is obtained from the amount of data that meets the criteria of the right branch candidate divided by the total data. The results of these calculations are presented in table 2.

Table 2. Result of PL and PR.

No	Left Branch Candidate	Right Branch Candidate	PL	PR
1	GPA <3,3	GPA >=3,3	0,067	0,933
2	GPA <=3,5	GPA >3,5	0,267	0,733
3	Essay <70	Essay >=70	0,067	0,933
4	Essay <85	Essay >=85	0,667	0,333
5	Interview <85	Interview >=85	0,467	0,533
6	Qur'an <20	Qur'an >=20	0,200	0,800
7	Qur'an <35	Qur'an >=35	1,000	0
8	Income Low	Income Low	0,067	0,933
9	Income Medium	Income High	0,800	0,200
10	Expertise Certificate =No	Expertise Certificate = Yes	0,400	0,600

P (j | tL) and P (j | tR) are calculated using formulas (3) and (4). Table 3 shows the calculation results of P (j | tL) and P (j | tR).

Table 3. Calculation of P(j|tL) and P(j|tR).

No	Left Branch Candidate	Right Branch Candidate	P(L|tL)	P(TL|tL)	P(L|tR)	P(TL|tR)
1	GPA <3,3	GPA >=3,3	0	1	0,571	0,429
2	GPA <=3,5	GPA >3,5	0,25	0,75	0,636	0,364
3	Essay <70	Essay >=70	0	1	0,571	0,429
4	Essay <85	Essay >=85	0,3	0,7	1,000	0,000
5	Interview <85	Interview >=85	0	1	1,000	0,000
6	Qur'an <20	Qur'an >=20	0,333	0,667	0,583	0,417
7	Qur'an <35	Qur'an >=35	0,533	0,467	0	0
8	Income Low	Income Low	0,000	1,000	0,571	0,429
9	Income Medium	Income High	0,500	0,500	0,667	0,333
10	Expertise Certificate =No	Expertise Certificate = Yes	0,333	0,667	0,667	0,333

Tables 2 and 3 can be seen that the results of the calculation of P (j | tL) with L status are obtained from the amount of data that meets the left branch candidate and the status L is divided by the total amount of data that meets the left branch candidate criteria. Likewise P (j | tL) with the status of TL divided by the entire amount of data that meets the criteria of the left branch candidate. For P (j | tR) L status is obtained from the amount of data that fulfils the criteria of the right branch candidate L status divided by the total amount of data that meets the criteria of the right branch candidate, as well as P (j | tR) TL status.

Table 4. Calculation of goodness.

No	Left Branch Candidate	Right Branch Candidate	i(tL)	i(tR)	Gini Index	Goodness
1	GPA <3,3	GPA >=3,3	0	0,490	0,498	0,041
2	GPA <=3,5	GPA >3,5	0,37	0,463	0,498	0,058
3	Essay <70	Essay >=70	0	0,490	0,498	0,041
4	Essay <85	Essay >=85	0,42	0,000	0,498	0,218
5	Interview <85	Interview >=85	0	0,000	0,498	0,498
6	Qur'an <20	Qur'an >=20	0,444	0,486	0,498	0,020
7	Qur'an <35	Qur'an >=35	0,498	0	0,498	0
8	Income Low	Income Low	0,000	0,490	0,498	0,041
9	Income Medium	Income High	0,500	0,444	0,498	0,009
10	Expertise Certificate =No	Expertise Certificate = Yes	0,444	0,444	0,498	0,053

Based on table 4, it can be seen that branch number 5 has the greatest goodness value. Then, candidate branch 5 will be the first branch in the decision tree and so on.

3. Results and discussion

The testing phase is done to find out how the algorithm works and what the results of the algorithm's process are like. In the testing process, the training data used were 171 as many as 30% of the training data. The data has been classified and is original data obtained from the authorities of the program organizing university. This test is done by calculating the recall value to get a percentage of the ability of the algorithm to find information back with the following formula:

$$Recall = \frac{Amount\ of\ data\ classified\ correctly}{The\ actual\ amount\ of\ data} \quad (5)$$

Testing the quality of the algorithm in this study was carried out three times by using a varied dataset distribution as described below:

- Scenario A (Training data sharing and Data Testing by percentage 70% dan 30%)

$$Recall = \frac{47}{52}\ x\ 100\% = 90\% \quad (6)$$

- Scenario B (Training data sharing and Data Testing by percentage 50% dan 50%)

$$Recall = \frac{82}{86}\ x\ 100\% = 95\% \quad (7)$$

- Scenario C (Training data sharing and Data Testing by percentage 40% dan 60%)

$$Recall = \frac{90}{104}\ x\ 100\% = 86\% \quad (8)$$

Based on the three test scenarios above, it can be seen that the highest accuracy lies in the second test by dividing the percentage of 50% and 50%. While the smallest percentage is in the third test with the percentage of training data 40% and 60% testing data. It can be concluded that the amount of data entered either the amount of training data or testing data affects the results of accuracy. The final rule is detailed in figure 1.

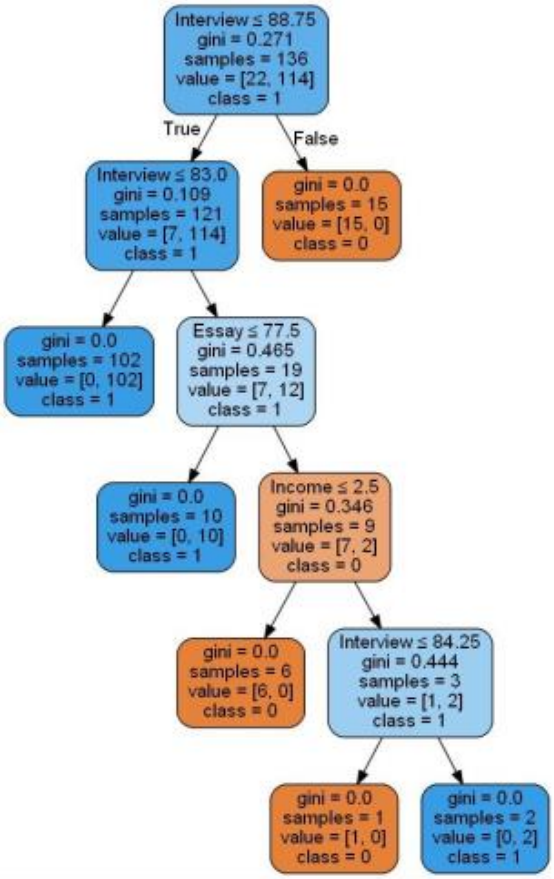

Figure 1. Final rule.

4. Conclusion

This proposed model can be used well in providing the prediction of student exchange participants. One of the variables or determinant factors used is GPA, income, essays, interviews, and others that can provide fairly accurate results. In the testing phase, all test scenarios show positive results. Further works, we suggest to improve this proposed model with any positive determinant factor that is adjustable to any condition and policy. This model needs to compare with several classification methods such as C45, Support Vector Machine, J48, and so on.

References

[1] Leutwyler B and Lottenbach S 2011 Reflection on normality: The benefits of international student exchange for teacher education *Pains and gains of international mobility in teacher education* (Brill Sense) pp 59-77
[2] Bevis T B 2019 The Student Exchange Boom Following World War II *A World History of Higher Education Exchange* (Cham.: Palgrave Macmillan) pp 135-169
[3] Messer D and Wolter S C 2007 Are student exchange programs worth it? *Higher Education* **54**(5) 647-663
[4] Matuk C and Linn M C 2018 Why and how do middle school students exchange ideas during science inquiry? *International Journal of Computer-Supported Collaborative Learning* **13**(3) 263-299
[5] Zulfikar W B, Wahana A, Uriawan W and Lukman N 2016 Implementation of association rules with apriori algorithm for increasing the quality of promotion *2016 4th International Conference on Cyber and IT Service Management* (IEEE) pp 1-5
[6] Gerhana Y A, Zulfikar W B, Ramdani A H and Ramdhani M A 2018 Implementation of Nearest Neighbor using HSV to Identify Skin Disease *IOP Conf. Ser. Mater. Sci. Eng* **288**(1) 012153
[7] Boerstler H and de Figueiredo J M 1991 Prediction of use of psychiatric services: Application of the CART algorithm *The journal of mental health administration* **18**(1) 27-34
[8] Batra M and Agrawal R 2018 Comparative analysis of decision tree algorithms *Nature inspired computing* (Singapore: Springer) pp 31-36
[9] Bhawiyuga A, Kartikasari D P, Amron K, Pratama O B and Habibi M 2019 Architectural design of IoT-cloud computing integration platform *Telkomnika* **17**(3) 1399
[10] Zulfikar W B, Irfan M, Alam C N and Indra M 2017 The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (IEEE) pp 1-5
[11] Setiawati D, Taufik I, Jumadi J and Zulfikar W B 2016 Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile *Jurnal Online Informatika* **1**(1) 24-27
[12] Bhargava N, Dayma S, Kumar A and Singh P 2017 An approach for classification using simple CART algorithm in WEKA *2017 11th International Conference on Intelligent Systems and Control (ISCO)* (IEEE) pp 212-216
[13] Zimmerman R K, Balasubramani G K, Nowalk M P, Eng H, Urbanski L, Jackson M L ... and Malosh R E 2016 Classification and Regression Tree (CART) analysis to predict influenza in primary care patients *BMC infectious diseases* **16**(1) 503
[14] Jing R, Zhang Q, Wang B, Cui P, Yan T and Huang J 2019 CART-based fast CU size decision and mode decision algorithm for 3D-HEVC *Signal, Image and Video Processing* **13**(2) 209-216
[15] Bar-Hen A, Gey S and Poggi J M 2015 Influence measures for CART classification trees *Journal of Classification* **32**(1) 21-45
[16] Lemon S C, Roy J, Clark M A, Friedmann P D and Rakowski W 2003 Classification and regression tree analysis in public health: methodological review and comparison with logistic regression *Annals of behavioral medicine* **26**(3) 172-181

[17] Li M 2017 Application of CART decision tree combined with PCA algorithm in intrusion detection *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (IEEE) pp 38-41
[18] Yan H, Hu H and Yu P 2019 A Study on Push Technology of Intelligent Agriculture Service Information Based on CART Algorithm *2019 International Conference on Robots & Intelligent System (ICRIS)* (IEEE) pp 258-260
[19] Li Y, Khan M Y A, Jiang Y, Tian F, Liao W, Fu S and He C 2019 CART and PSO+ KNN algorithms to estimate the impact of water level change on water quality in Poyang Lake, China *Arabian Journal of Geosciences* **12**(9) 287
[20] Rai S, Khandelwal N and Boghey R 2020 Analysis of Customer Churn Prediction in Telecom Sector Using CART Algorithm *First International Conference on Sustainable Technologies for Computational Intelligence* (Singapore: Springer) pp 457-466
[21] Bianco S, Ciocca G and Cusano C 2009 Color constancy algorithm selection using CART *International Workshop on Computational Color Imaging* (Berlin, Heidelberg: Springer) pp 31-40