



ARTICLE



<https://doi.org/10.1057/s41599-022-01413-z>

OPEN

# Evaluation of college admissions: a decision tree guide to provide information for improvement

Ying-Sing Liu <sup>1</sup>✉ & Liza Lee<sup>1</sup>

This study uses decision trees to analyze the admissions and enrollment of Taiwan's 5-year junior colleges to explore the reasons that students might fail in an exam-free admissions process, propose methods for improvement, and view the implementation of the pedagogical theory of multiple intelligences. The college admissions system may produce confusion in Taiwan. Schools in metropolitan areas retain an advantage for screening talent across multiple abilities, and colleges in agricultural counties may unintentionally marginalize people, resulting in insufficient enrollment or an inverse selection of talent. It has been suggested that increasing the number of schools in metropolitan areas will reduce the rates of enrollment failure and improve the compulsory education environment that many are forced to attend.

<sup>1</sup> College of Humanities & Social Sciences, Chaoyang University of Technology, 168, Jifeng E. Road, Wufeng District, Taichung 413310, Taiwan.  
✉email: [liuyingsing@yahoo.com.tw](mailto:liuyingsing@yahoo.com.tw)

## Introduction

To alleviate the pressures of pursuing further education, facilitate the adaptive development of students, balance urban and rural education and improve the academic training of the people, the 12-year basic national education program was implemented in Taiwan. This education program is based on the multiple intelligence theory, which provides an adaptive approach to the admissions system, affirming learners' right to education and equal opportunities in admission (Ministry of Education, 2009). In the 5-year junior college program, by promoting the assessment of learning abilities of students as the basis for evaluation through a variety of admission systems, schools can focus on the potential diversified development of students while at the same time creating a fair and reasonable admission system that expands the free admission process since 2011.

The exam-free admissions system considers the candidate's choice for Taiwan's 5-year junior college programs. When the number of applicants exceeds the number of available seats, the admission system recommends more than the allocated quota, using criteria such as "multiple learning performance," "technically and artistically gifted," "disadvantaged status," "balanced learning performance," and "counseling competency". Multiple learning performances during secondary school, as well as the results of the "comprehensive assessment program" and other results, are listed and sorted according to the criteria set by each school. Under this admissions system, applicants can choose their own colleges and departments according to their own conditions and interests, while the colleges can recruit the appropriate talent according to their standards. During the enrollment process, applicants may also choose to forgo the opportunity to enroll. The educational beliefs on which this system is based are fair and testable.

However, when the number of enrollments far exceeds the number of seats available at the college, a majority is generated of students who failed to pass the entrance requirements. Furthermore, if enrollees have a preference for particular colleges and departments, other schools may become underenrolled. When these phenomena occur, it is necessary to explore whether this exam-free admission registration system can maintain expected educational objectives or cause popular colleges to retain the advantage of blatantly screening outstanding students. Marginalized rural and suburban colleges produce trends of insufficient enrollment or weak competition, showing two extremes within the college enrollment phenomenon.

Decision tree analysis is a common data-mining method that can be used as a tool for supervised feature extraction and description (Berry and Linoff, 1997). High-dimensional data can be quickly learned, a hierarchical tree structure can be established, and the results obtained can be transformed into easy-to-understand rules, generally for exploration and prediction purposes. At the same time, decision trees have the advantage of reducing unnecessary variables and sorting the importance of independent variables. In this study, the classification rules (factors) for enrollment failures/successes will be easily obtained during the classification process through the decision tree method, making these rules clear, easy to understand and easy to explore for applicants (students), admission-based colleges and education experts. This study uses statistical tests and decision tree analyses to assess possible factors for enrollment failure and success, reviews the consistency of the admissions system with stated educational philosophy and objectives, and makes specific recommendations.

The content of this research article is as follows: section "Literature review" discusses the literature review, section "Data and research methods" presents the data and research methods,

section "Empirical results and discussion" shows the empirical results and discussion, and section "Conclusions" presents the conclusions.

## Literature review

Data mining is the analytical step of the knowledge discovery in databases (KDD) process (Fayyad et al., 1996). It is a technique that combines statistics, algorithms, artificial intelligence, and machine learning and uses databases for calculations and analyses to explore specific rules or models from large, cluttered data. KDD is useful for finding potentially valuable information to solve problems. This rapidly developing new approach and technology is being widely used across a broad range of applications (Aulck et al., 2019; Han, 2022; Kiss et al., 2019; Nagy and Molontay, 2018; PhridviRaj and GuruRao, 2014; Rastrollo-Guerrero et al., 2020). The decision tree is one of the most common methods used in data-mining technology and is essentially a simple classifier (Kingsford and Salzberg, 2008), which produces a kind of supervised learning that can be used as an analytical data and prediction tool. Compared with statistical methods assessing the parameters of the data, decision tree analysis is a more reliable and intuitive method (Park and Dooris, 2020) that has been widely used to solve problems in the field of education (Delibasic et al., 2013; Lee and Liu, 2021; Yao et al., 2022).

Kirby and Dempster (2014) used decision tree analysis to identify the variables that best describe student achievement, using college student achievements to identify challenges and remediation opportunities during the course selection process and the course itself. Park and Dooris (2020) explored assessments and evaluations in higher education using decision tree analysis to predict student evaluation of teaching. Križanić (2020) applied data-mining techniques to educational data in higher education institutions, exploring student behavior interacting with course materials through a decision tree to better analyze how students learn. Asif et al. (2017) believed that data mining results can provide timely warning and support to underachieving students while providing advice and opportunities to high-achieving students. Singer et al. (2020) used decision trees to predict the stability of the academic behavior of students with learning disabilities (LD) with and without accommodation factors and found that the rendered models were excellent in predicting performance.

Amburgey and Yi (2011) used the principles of business intelligence to explore first-year data from master's students in private colleges with decision tree analysis, neural network analysis, and multiple regression analysis to develop models to predict the average score (GPA) for each student. Lin et al. (2013) used the decision tree to build a Personalized Creative Learning System (PCLS) as an important predictor of university professional and learning paths. Howard et al. (2018) used eight predictive models as an early warning system to assess both students' potential learning and poor learning in the course, resulting in a preference for the Bayesian additive regression trees (BART) as the best predictive model. Lynch (2017) offers insight into the issues that can be exacerbated while applying big data analytics to education and argues that while big data can lead to personalized learning, deep student modeling, and true vertical learning, its application requires in-depth and continuous monitoring of students, classes, and teachers, as well as an invasion of privacy, potential interference with educational effectiveness, and other ethical issues.

Oranye (2016) focuses on the importance of college admissions. In past studies, data mining has been applied to predict school enrollment to provide useful information for the effective

improvement and achievement of enrollment goals. For example, Tanna (2012) developed a decision support system that enables students to choose the right university based on their entrance exam scores. Zeng et al. (2014) used decision tree construction models to predict the popularity of universities in various regions of China, and the results showed the practical feasibility of decision tree modeling. Ragab et al. (2012) proposed a hybrid recommendation for a university admission system based on data-mining techniques and knowledge discovery rules to address college enrollment forecasting issues and recommend appropriate tracking channels for students to enroll successfully. Maltz et al. (2007) used computers to develop a decision-supporting system to improve responsiveness and real-time management capabilities in the college admission process, significantly increasing the effectiveness of the process and achieving enrollment business goals.

Finally, the theory of multiple intelligences argues that learners should be empowered, not limited, in the way they learn (Gardner, 2011). In particular, presentations of intelligence vary from person to person, so each learner should have a different learning course, and it is recommended that learners should not all be measured by the same criteria for learning effectiveness. Waterhouse (2006) casts doubt on the multiple intelligence theory, arguing that the lack of sufficient empirical support implies that it should not be the basis of educational practice. Chou (2009) questioned the fairness of allocating educational resources in Taiwan and the right of all citizens to be taught. The presence of an exam-free admission registration system does not reduce the long-term pressure on candidates, nor does it slow down the teaching of exam leaders.

**Data and research methods**

This study explores the enrollment of 5-year junior colleges in Taiwan in 2016, which is 5 years after the implementation of the new education reform system in 2011, to understand how the new system is deployed in practice. The information collected comes from the committee of joint admissions. These data are filled in by the participating applicants (students), the necessary personal achievements are submitted, and the personal information and data are checked and verified by the checker before submitting a formal registration in the computer database of the joint admission registration committee. In these 5-year junior colleges, 6013 applicants used the exam-free admissions system; only 2294 people were successfully enrolled, and the remaining 3719 people failed, accounting for 61.85% of the total applicants. Please refer to Table 1 for the differences and characteristics of the four enrolling colleges in terms of location, enrollment scale, and admissions department. This study analyzes the data at the time of registration and organizes them in the form of variables based on Table 2, which contains the definition and description of the study variables indicating the data patterns and definitions for the 21 candidate-based independent variables and 1 dependent variable of the enrollment results (Y1).

Data mining is based on combined algorithms of statistical analysis, which use rapid computing abilities to analyze big data and find useful knowledge as a decision analysis tool or predictive technology (PhridviRaj and GuruRao, 2014). Vialardi et al. (2011) describe the data-mining analysis process, which consists of six steps: business understanding, data understanding, data dating, modeling, evaluation, and deployment. Some of the methods of data mining (e.g., traditional artificial neural network models) should be applied with caution because these may not have the ability to automatically filter variables. When many candidate independent variables cannot be proven to have a significant impact on the dependent variable, the variables might be screened first with decision tree analysis, statistical tests, or other dimensional reduction methods. In this case, this study collects data through the decision tree classification method to establish an assessment model; the model can then be used for factor (rule) exploration of admission successes and failures and as a future application for admissions assessment or predictive analysis.

After categorizing the applicants between “enrollment failures” and “enrollment successes,” an evaluation model is established through the decision tree classification method to identify the problems that cause enrollment failures. The decision tree is a data classification method that establishes a tree structure that usually groups cases according to the independent variables or the prediction value of the dependent variable. The established tree structure provides a validation tool for interpretation and confirmation classification analysis. Common algorithms include CART (classification and regression trees), CHAID (chi-square automatic interaction detector), and C5.0 (Combination 5.0).

The CHAID growth method has been used in this decision tree; it is based on the chi-square distribution. The variables assessed must be categorical, and if there is a continuous variable, the data need to be transformed to a categorical variable. Through the automatic interaction detection of the chi-square, independent variables that have the strongest interaction with the dependent variable are selected. When there is no significant difference between categories and related dependent variables, the categories of the independent variables can be automatically merged. The CHAID algorithm is based on chi-square testing to determine the best branching properties, which can be divided into multiple branches. CHAID also has the advantages of fast calculation speed, and does not consider postpruning. In addition, CHAID directly joins the mechanism that stops the growth of the decision tree during the establishment of the decision tree.

**Empirical results and discussion**

**Descriptive statistics and tests.** Table 3 shows the results of the frequency distribution table and the chi-square test of independence for the categorical variables. In Table 3, in the enrolling colleges (X1), School A had the largest number of applicants, and 2810 people accounted for 46.7% of the total sample size. The ratio of applicants to enrollment places is 4.01. A total of 2382 applicants in School D accounted for 39.6% of the total sample

**Table 1 Characteristics of four colleges in the exam-free admissions system in Taiwan.**

Enrollment colleges	Exam-free admissions college	Enrollment department	Enrollment size	Address of admissions school
A School	National university of science and technology	Business, management and nursing	700 spaces	Metropolitan city
B School	Private university of science and technology	Nursing	50 spaces	Metropolitan city
C School	Private university of science and technology	Industry, management	275 spaces	Agricultural county
D School	Private junior college	Nursing, medicine and management	1269 spaces	Agricultural county

**Table 2 The definition and description of the study variables.**

Study variables	Code	Type: scale	Definition and description
The enrollment distribution results	Y1	Categorical: Nominal	Success and fail
The enrolling colleges	X1	Categorical: Nominal	A school, B school, C school and D school
The general location of the enrolling colleges	X2	Categorical: Nominal	T city, D county and M county
Registered schools in a metropolitan area	X3	Categorical: Nominal	Agricultural county and metropolitan city
The distance between the enrollee and the distribution colleges	X4	Categorical: Order	Local county or city to 0; crossing a county or city to give 1; Across two counties or cities to give 2; Across three counties or cities to give 3; Across four counties or cities to give 4; Across five or more counties or cities to 5
The registration threshold	X5	Categorical: Order	No and Yes
Types of junior high-school graduates	U1	Categorical: Nominal	The country, city, county, private and other
From the core urban area	U2	Categorical: Nominal	No and Yes
From remote or outlying areas	U3	Categorical: Nominal	No and Yes
The size of junior high-school graduation	U4	Measure	Number of new graduates
Competition	U5	Measure	0-7 point
Service-learning effectiveness	U6	Measure	0-7 point
Daily life performance	U7	Measure	0-4 point
Physical fitness	U8	Measure	0-6 point
Multiple learning performances	U9	Measure	0-24 point
Performances of technically and artistically	U10	Measure	0-3 point
Disadvantaged status	U11	Measure	0-2 point
Balanced learning performance	U12	Measure	0-6 point
Counseling competency	U13	Measure	0-3 point
Comprehensive assessment program	U14	Measure	0-21 point
Writing test	U15	Measure	0-5 point
Others	U16	Measure	0-5 point (A and B schools with GEPT or TOEIC plus points; C and D schools are not)

size, and the ratio of applicants to enrollment places was 1.8. In School B, there are 544 applicants, accounting for 9.0% of the total sample size, and the ratio of applicants to enrollment places is 10.88. Finally, 277 people in School C accounted for 4.6% of the total sample size, with a ratio of applicants to enrollment places of 1.01. A total of 3354 people, who accounted for 55.8% of the total sample size, chose colleges in the metropolitan area.

Figure 1 shows a bar chart of the exam-free admissions schools in metropolitan areas and agricultural counties in terms of enrollment places and actually enrolled students. It was found that schools in the metropolitan area had full enrollment (when the last students to be admitted have the same conditions, the number of students who can be admitted can be increased so that the number of actual enrolled students is greater than the number of enrollment places), but the schools in agricultural counties were underenrolled (a possible enrollment of 1544 places vs. an actual enrollment of 1286 places).

The distance between the applicants and the registered colleges spread across the county and city, but the largest plurality of applicants were local ( $X4 = 0$ ), with 2140 people accounting for 35.6% of the total sample size. There are 1799 people coming from across the county and city ( $X4 = 1$ ), accounting for 29.9%, and 1202 people coming from across 2 counties and cities ( $X4 = 2$ ), accounting for 20.0%, while the remaining 872 people come from across at least 3 counties and cities ( $X4 \geq 3$ ), accounting for 14.5% of the total sample size. These results are in line with the educational philosophy stressing admission close to the student's home.

The registration threshold set at the enrolling college, which did not meet the entry threshold for registered colleges, accounted for 19.7% of the total sample size with 1182 people. There were 4819 applicants from the noncore urban area, who accounted for 80.1% of the total sample size, and only 122 applicants (2.0%) from remote or outlying areas.

There were 2592 applicants enrolled in City T (metropolitan area) (65.4% of the total enrollment failures and 77.28% of the total applicants were in City T), indicating that the colleges

registered in the metropolitan area were more popular. When the distance between the enrollee and the target college is spread across at least 3 counties and cities ( $X4 \geq 3$ ), the enrollment failure rate is as high as 80%, higher than the nearest county and city enrollees (approximately 61 to 67%). A total of 2783 people (70.2% of all enrollment failures) reached the admissions threshold but failed to enroll. A total of 90.16% of the applicants were from remote or outlying areas; 110 people failed to enroll, indicating that the applicants from remote or outlying areas were in danger of competing with one another.

Table 3 also shows the chi-square test of independence results for the independent variables and the admission distribution results (Y1). At a significance level of 5%, it is found that variables such as enrolling colleges (X1), county or city where the enrolling colleges are situated (X2), registered schools in a metropolitan area (X3), the distance between the enrollee and the target colleges (X4), the admissions threshold (X5), types of junior high-school graduates (U1), and origins from remote or outlying areas (U3) are statistically significant, indicating that there is a dependency between the seven variables and the enrollment distribution results (Y1).

Table 4 shows the descriptive statistics variables, the difference in means *t*-tests (failures–success), Mann-Whitney *U*-tests, and Kolmogorov-Smirnov *Z*-tests. In Table 4, based on the results of the enrollment distribution (Y1), the categorical differences between enrollment failures and enrollment successes are presented in the following manner: the size of junior high-school graduation (U4) (483.820 vs. 467.177), competition (U5) (0.501 vs. 0.424), service-learning effectiveness (U6) (6.462 vs. 6.325), daily life performance (U7) (3.894 vs. 3.795), multiple learning performances (U9) (23.040 vs. 22.785), comprehensive assessment program (U14) (12.850 vs. 12.104), and writing test (U15) (4.082 vs. 3.956). The average performance of enrollment failures is higher than that of enrollment successes and shows higher scores that belong to failed candidates and lower scores that belong to successful candidates, showing the phenomenon of

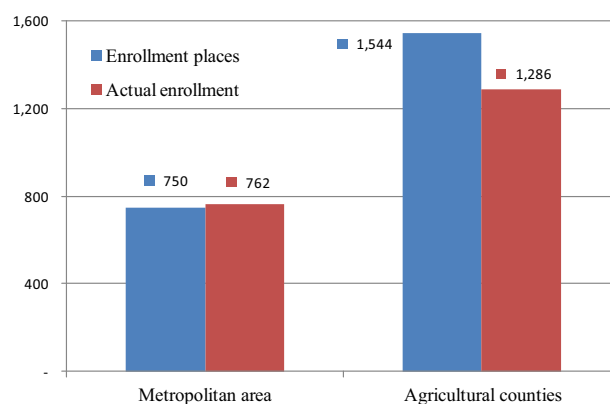
**Table 3 Frequency distribution table and the chi-square test of independence.**

Item	Whole sample		The enrollment distribution results (Y1): enroll failures			The chi-squared test of independence: Pearson's chi-squared test Sig. (2-tailed)
	Frequency	%	Frequency	%	% of item	
Part I. The enrolling colleges (X1)						
A school	2810	46.7%	2098	52.9%	74.7%	$\chi^2_{(3)} = 495.93^{**}$ $p\text{-value} = 0.000 < 0.01$
B school	544	9.0%	494	12.5%	90.8%	
C school	277	4.6%	121	3.1%	43.7%	
D school	2382	39.6%	1252	31.6%	52.6%	
Part II. The enrolling colleges are situated (X2)						
T city	3354	55.8%	2592	65.4%	77.3%	$\chi^2_{(2)} = 443.021^{**}$ $p\text{-value} = 0.000 < 0.01$
S county	277	4.6%	121	3.1%	43.7%	
M county	2382	39.6%	1252	31.6%	52.6%	
Part III. Registered schools in a metropolitan area (X3)						
No	2659	44.2%	1373	34.6%	51.6%	$\chi^2_{(1)} = 434.312^{**}$ $p\text{-value} = 0.000 < 0.01$
Yes	3354	55.8%	2592	65.4%	77.3%	
Part IV. The distance between the enrollee and the distribution colleges (X4)						
0	2140	35.6%	1317	33.2%	61.5%	$\chi^2_{(5)} = 187.911^{**}$ $p\text{-value} = 0.000 < 0.01$
1	1799	29.9%	1099	27.7%	61.1%	
2	1202	20.0%	806	20.3%	67.1%	
3	251	4.2%	219	5.5%	87.3%	
4	273	4.5%	217	5.5%	79.5%	
5	348	5.8%	307	7.7%	88.2%	
Part V. The registration threshold (X5)						
No	1182	19.7%	1182	29.8%	100.0%	$\chi^2_{(1)} = 759.903^{**}$ $p\text{-value} = 0.000 < 0.01$
Yes	4831	80.3%	2783	70.2%	57.6%	
Part VI. Types of junior high-school graduates (U1)						
National	11	0.2%	8	0.2%	72.7%	$\chi^2_{(4)} = 37.070^{**}$ $p\text{-value} = 0.000 < 0.01$
City	2763	46.0%	1730	43.6%	62.6%	
County	2746	45.7%	1855	46.8%	67.6%	
Private	487	8.1%	367	9.3%	75.4%	
Other	6	0.1%	5	0.1%	83.3%	
Part VII. From the core urban area (U2)						
No	4819	80.1%	3180	80.2%	66.0%	$\chi^2_{(1)} = 0.025$ $p\text{-value} = 0.891 > 0.05$
Yes	1194	19.9%	785	19.8%	65.8%	
Part VIII. From remote or outlying areas (U3)						
No	5891	98.0%	3855	97.2%	65.4%	$\chi^2_{(1)} = 32.534^{**}$ $p\text{-value} = 0.000 < 0.01$
Yes	122	2.0%	110	2.8%	90.2%	

\* $p \leq 0.05$ ; \*\* $p < 0.01$ .

a reverse selection of talent. However, in the four performances of technically and artistically gifted (U10) (0.446 vs. 0.713), disadvantaged status (U11) (0.142 vs. 0.164), balanced learning performance (U12) (7.022 vs. 7.866), and other factors (U16) (0.220 vs. 0.710), the performance average of enrollment failures is less than that of enrollment successes, as evident from these results.

Table 4 also shows the results of the two-sample *t*-test for difference in means (failures–success), Mann-Whitney *U*-tests, and Kolmogorov–Smirnov *Z*-tests. At a significance level of 5%, the average difference was significantly positive in categories such as the size of the junior high-school graduation (U4) (16.643;  $p\text{-value} = 0.009$ ), competition (U5) (0.077;  $p\text{-value} = 0.028$ ), service-learning effectiveness (U6) (0.137;  $p\text{-value} = 0.000$ ), daily life performance (U7) (0.098;  $p\text{-value} = 0.000$ ), multiple learning performances (U9) (0.098;  $p\text{-value} = 0.001$ ), comprehensive assessment program (U14) (0.746;  $p\text{-value} = 0.000$ ), and writing test (U15) (0.126;  $p\text{-value} = 0.000$ ), which indicates that the performance average for enrollment failures is significantly higher than that of enrollment successes. In addition, categories such as technically and artistically gifted (U10) (–0.267;  $p\text{-value} = 0.000$ ), balanced learning performance (U12) (–0.844;  $p\text{-value} = 0.000$ ), and other factors (U16) (–0.490;  $p\text{-value} = 0.000$ ) are significantly negative, showing that the performance average for enrollment failures is significantly lower than that of enrollment



**Fig. 1 A bar chart comparing enrollments.** A bar chart of enrollment places and actual enrollment in the exam-free admissions schools of the metropolitan area and agricultural counties.

successes. Finally, the results of the Mann-Whitney *U*-tests and Kolmogorov–Smirnov *Z*-tests are roughly similar to those of the two-sample *t*-test for differences in means.

Table 4 shows that five variables, namely, competition (U5), service-learning effectiveness (U6), daily life performance (U7), multiple learning performances (U9), comprehensive assessment



**Table 4 The descriptive statistics for measured variables, difference of means t-tests, Mann-Whitney U-tests, and Kolmogorov-Smirnov Z-tests.**

Stat. or test	Whole sample (N = 6013)		Enroll failure (N = 3965)		Enroll success (N = 2048)		t-test for equality of means		Mann-Whitney U-test		Kolmogorov-Smirnov Z-test	
	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev	Difference (fail-success)	t-stat.	p-value	z-stat.	p-value	z-stat.
Size of school (U4)	478.146	235.205	483.820	235.419	467.177	234.459	16.643**	2.599	0.009	-2.549*	0.011	1.851**
Competition (U5)	0.475	1.328	0.501	1.365	0.424	1.251	0.077*	2.197	0.028	-1.996*	0.046	0.712
Service-learning effectiveness (U6)	6.415	1.328	6.462	1.262	6.325	1.443	0.137**	3.647	0.000	-3.870**	0.000	1.607*
Daily life performance (U7)	3.860	0.6286	3.894	0.541	3.795	0.767	0.098**	5.168	0.000	-5.261**	0.000	1.218
Physical fitness (U8)	5.664	1.087	5.652	1.095	5.689	1.071	-0.037	-1.243	0.214	-1.831	0.067	0.583
Multiple learning performances (U9)	22.953	2.715	23.040	2.498	22.785	3.086	0.254**	3.225	0.001	-1.430	0.153	0.909
Technically and artistically (U10)	0.537	1.041	0.446	0.963	0.713	1.158	-0.267**	-8.950	0.000	-9.427**	0.000	3.835**
Disadvantaged status (U11)	0.150	0.526	0.142	0.514	0.164	0.549	-0.022	-1.492	0.136	-1.523	0.128	0.401
Balanced learning performance (U12)	7.310	1.519	7.022	1.459	7.866	1.476	-0.844**	-21.095	0.000	-20.664**	0.000	10.271**
Counseling competency (U13)	2.987	0.183	2.988	0.1692	2.984	0.207	0.004	0.809	0.419	-0.390	0.697	0.069
Comprehensive assessment program (U14)	12.596	2.764	12.850	2.4593	12.104	3.218	0.746**	9.191	0.000	-9.302**	0.000	6.864**
Writing test (U15)	4.039	0.712	4.082	0.6517	3.956	0.809	0.126**	6.083	0.000	-5.400**	0.000	1.804**
Other factors (U16)	0.387	1.166	0.220	0.8981	0.710	1.508	-0.490**	-13.525	0.000	-15.162**	0.000	4.699**

\*p ≤ .0.05; \*\*p < 0.01.

**Table 5 Results of classification using the CHAID method by samples of training/test.**

Sample	Observed	Predicted		
		Fail	Success	Ratio of correct predictions (%)
Training	Fail	2588	573	81.9%
	Success	538	1110	67.4%
	Overall ratio (%)	76.9%		
Test	Fail	648	156	80.6%
	Success	147	253	63.2%
	Overall ratio (%)	74.8%		

program (U14), and writing test (U15), are cases where the enroll failures will perform significantly better than the enroll successes. These results were not consistent with general expectations and related to participants' free choice to forgo this admission and then choose another school. Second, the number of students graduating from national secondary schools (size of school (U4)), where enrollment failures are significantly larger than enrollment successes, indicates that the participants' graduating schools do not have a competitive advantage in large student numbers.

**Decision tree analysis.** Table 5 shows the results of classification using the CHAID method by samples of training/test. This is the result of classifying the overall sample (6013 people) into 80% used for training the algorithm (4809 people) and 20% of applicants used to actually test the characteristics of the sample (1204 people) separately. The growth condition of the tree is 6 at the maximum depth of the tree structure, the minimum number of observations in the parent node is 100, and the minimum number of observations in the child node is 50. At a significance level of 0.05, the significant value of the consolidation and segmentation conditions is the result of growth under conditions adjusted by using the Bonferroni method. The training/test model has a risk value of 0.231/0.252, a standard error of 0.006/0.013, a sensitivity rate of 81.9/80.6% (precision rate of 67.4/63.2%) for enrollment failures, 76.9/74.8% for overall accuracy, and 39 nodes. The independent variables of the training/test model are in order: registered schools in a metropolitan area (X3), other factors (U16), registration threshold (X5), technically and artistically gifted (U10), comprehensive assessment program (U14), disadvantaged status (U11), distance between the enrollee and the distribution colleges (X4), types of junior high-school graduates (U1), and writing test (U15).

Figure 2 shows a tree diagram for the training/test sample. This study aims to be able to illustrate in detail the three sub-structure diagrams that divide the decision tree into Up (Figs. 3, 4 is the training/test sample), Left (Figs. 5, 6 is the training/test sample in the metropolitan area colleges), and Right (Figs. 7, 8 is training/test sample in the agricultural county colleges).

Figures 3 and 4 is the first level of the decision tree by training/test samples. In Node 0, 65.7/66.8% of the applicants in the training/test groups failed to enroll. The branches below node 0 are Node 1 and Node 2 (first level). In Node 1 (metropolitan area), 77.0/78.3% of the applicants in the training/test groups failed to enroll, which is a very high percentage. In Node 2 (agricultural county), 51.4/52.6% of the applicants in the training/test groups failed to enroll. This proportion is 25% lower than that of Node 1, indicating that colleges in agricultural counties are less competitive than those in the metropolitan area. The tree structure of the branches below node 1 in Figs. 3 and 4 is the

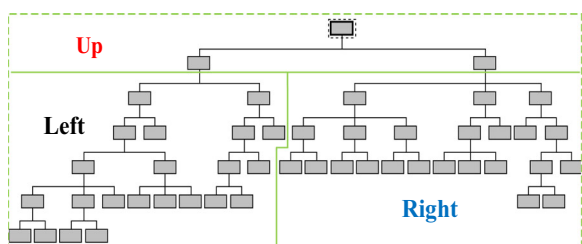
result of the enrollment of schools in the metropolitan area, while the tree structure of the following branches in node 2 is the result of the enrollment of schools in the agricultural county.

*Admission colleges in the metropolitan area.* From Node 1 in Figs. 5 and 6, 77.0/78.3% of the applicants in the training/test groups failed to enroll. The branching structure of the following level is highlighted for Node 1:

- (1) In Node 3, 86.1/85.4% of the applicants in the training/test group failed to enroll, indicating that if the applicant opts for a registered college in the metropolitan area, the bonus score for the English tests (GEPT or TOEIC) (other  $\leq 1.2$ ) becomes the main cause for enrollment failures.
  - I. In Node 8, 75.9/75.1% of the applicants in the training/test groups failed to enroll, and this proportion was high.
  - (I) At Node 19, 82.8/79.5% of the applicants in the training/test groups failed to enroll, which shows that among the applicants who opted for a registered college in the metropolitan area (other  $\leq 1.2$ , who reach the threshold registration, with technically and artistically gifted = 0) in these conditions, there will still be enrollment failures at up to approximately 80%. The fifth level is divided into two nodes by Node 19 below: Node 27 (comprehensive assessment program  $\leq 14$ ; 95.8/95.5% is the ratio of those who failed to enroll) and Node 28 (comprehensive assessment program  $> 14$ ; 64.0/60.0% is the ratio of those who failed to enroll).
    - i. At Node 27, 95.5/95.8% of the applicants in the training/test groups failed to enroll. The sixth level comprises the branch below Node 33 (disadvantaged status  $\leq 0$ ; 98.3/99.1% is the ratio of those who failed to enroll) and Node 34 (disadvantaged status  $> 0$ ; 71.2/72.2% is the ratio of those who failed to enroll). The enrollment failure rate of those

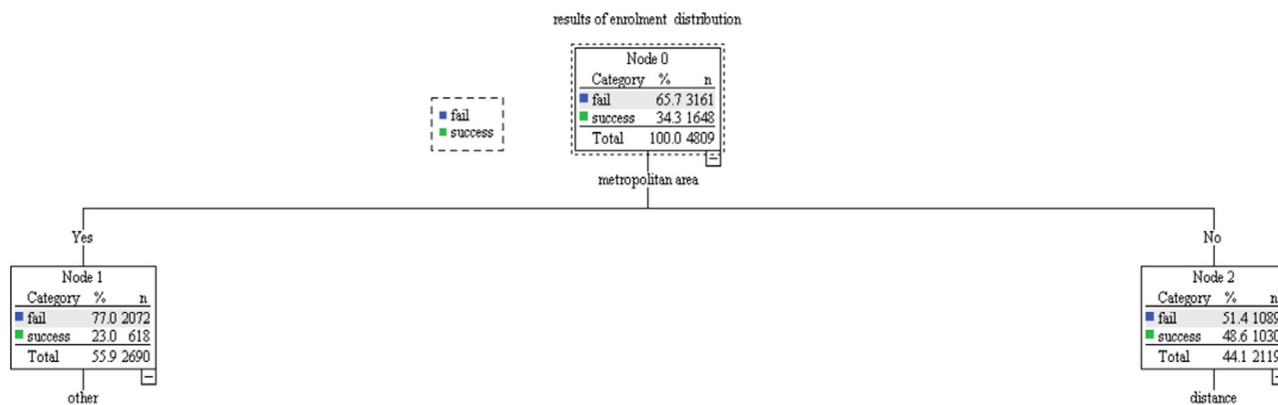
with disadvantaged status is lower than that of nondisadvantaged status, which shows that there is a partial safeguard effect for disadvantaged status.

- ii. At Node 28, 64.0/60.0% of the applicants in the training/test groups failed to enroll. The sixth level comprises the branch below Node 35 (the distance between the enrollee and the distribution colleges  $\leq 0$ ; 52.0/50.9% is the ratio of those who failed to enroll) and Node 36 (the distance between the enrollee and the distribution colleges  $> 0$ ; 75.7/69.1% is the ratio of those who failed to enroll). These results show that the enrollment failure rate will increase significantly for nonlocal applicants.
  - (I) At Node 20 (technically and artistically gifted  $> 0$ ), 53.2/61.7% of the applicants in the training/test groups failed to enroll. The fifth level is divided into two nodes by Node 20 below: Node 29 (the distance between the enrollee and the distribution colleges  $\leq 0$ ; 36.0/34.3% is the ratio of those who failed to enroll) and Node 30 (the distance between the enrollee and the distribution colleges  $> 0$ ; 67.9/82.6% is the ratio of those who failed to enroll). It was found that the enrollment failure rate increased when the distance between the enrollee and the target colleges was greater across the county and city.
    - I. (ii) At Node 9, the registration threshold was not reached, and 100/100% of the applicants in the training/test group failed to enroll.
  - (2) At Node 4 (other  $> 1.2$ ), 64.2/58.3% of the applicants in the training/test group were successful in enrolling. This result shows that when other  $> 1.2$ , the enrollment success rate in metropolitan colleges will be significantly higher than that of Node 3. The third level is composed of the branch below Node 4 to Node 10 (Comprehensive assessment program  $\leq 15.4$ ; 76.1/71.4% is the ratio of those who succeeded in enrolling) and Node 11 (Comprehensive assessment program  $> 15.4$ ; 54.5/56.2% is the ratio of those who succeeded in enrolling). After comparing the enrollment success rates from two horizontal independent nodes 10 and 11, it is found that the lower composite assessment program, instead, has a higher the ratio of those who succeeded in enrolling, showing an abnormal phenomenon.



**Fig. 2 A tree diagram.** A tree diagram for the training/test sample.

*Admissions in agricultural county colleges.* In Figs. 7 and 8, the second level is divided into three nodes by Node 2: Node 5 (the distance between the enrollee and the distribution colleges  $\leq 1.0$ ; 65.7/64.4% is the ratio of those who succeeded in enrolling), Node



**Fig. 3 A sub-tree on up for the training.** A sub-tree structure of the first to the second level for the training sample.

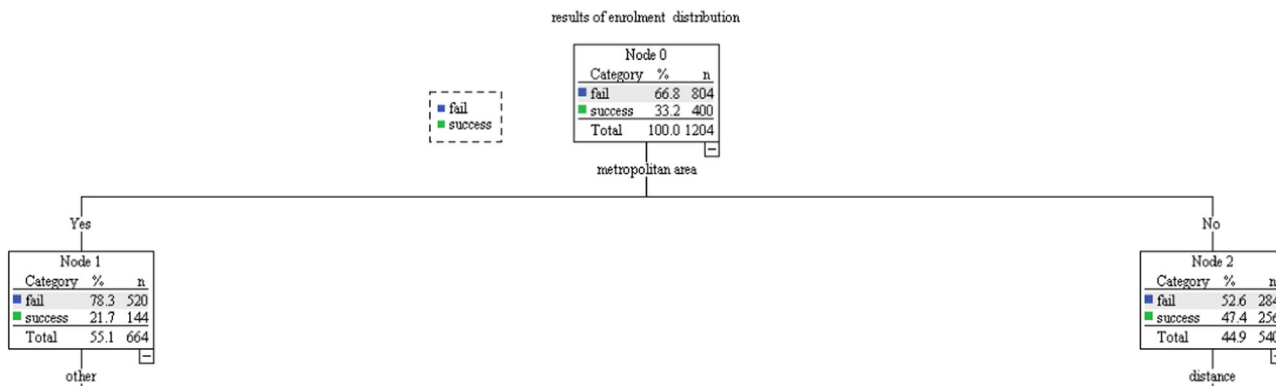


Fig. 4 A sub-tree on up for the test. A sub-tree structure of the first to the second level for the test sample.

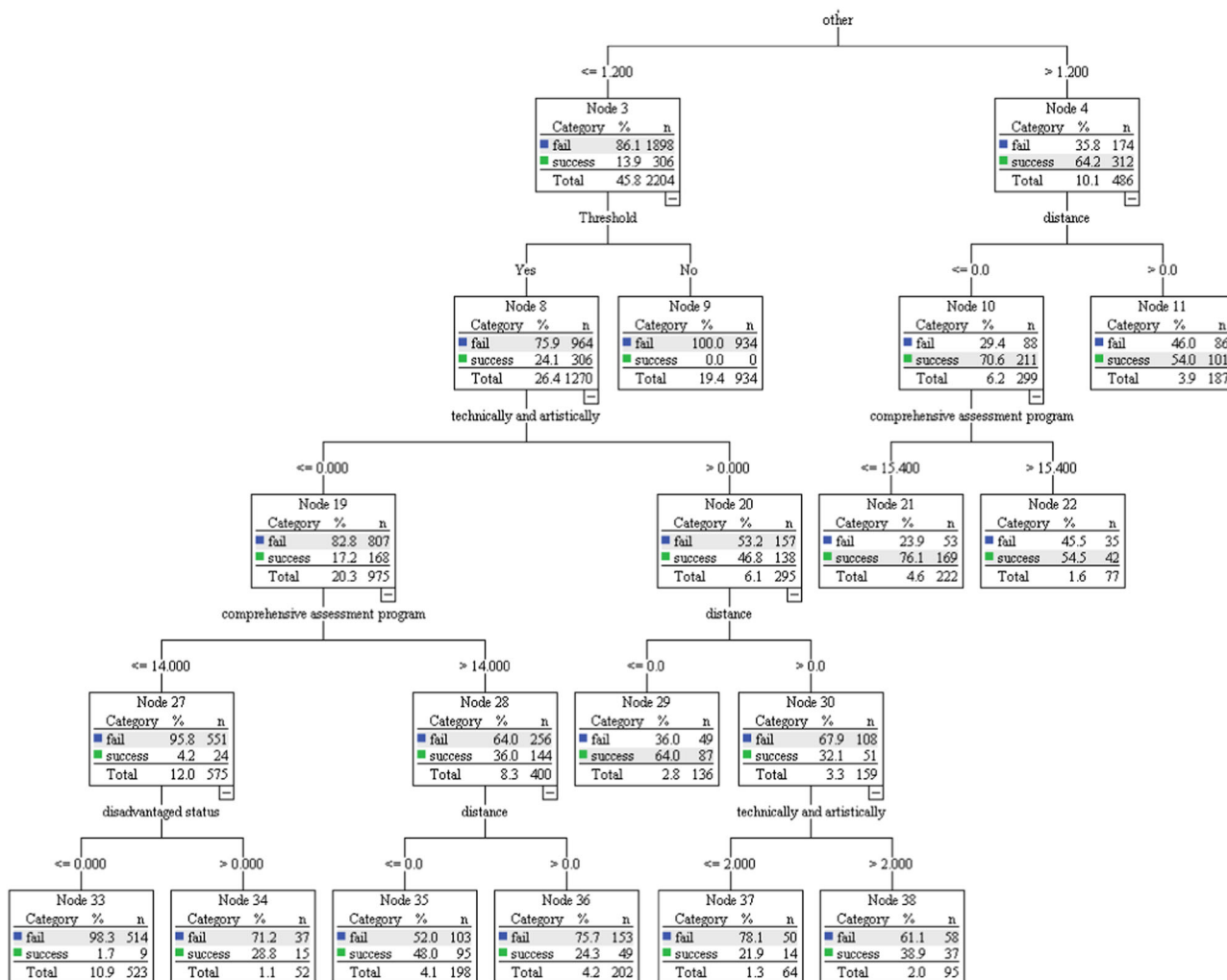


Fig. 5 A sub-tree on left for the training. A sub-tree structure for the training sample in the metropolitan area colleges.

6 ( $1 < \text{the distance between the enrollee and the distribution colleges} \leq 2$ , 60.8/61.7% is the ratio of those who failed to enroll), and Node 7 (the distance between the enrollee and the target colleges  $> 2$ , 80.3/79.8% is the ratio of those who failed to enroll). Comparing the three nodes in this level, it was found that the distance between the enrollee and the admitting college is associated with the ratio of success/failure for enrollments at agricultural county colleges. Applicants will have a higher ratio of successful enrollees when they are in the same county or city as

the school of admission. When applicants and schools span two counties and cities, there is a higher ratio of failed enrollment.

1. At Node 5, the third level is divided into three nodes by Node 5: Node 12 (comprehensive assessment program  $\leq 11.2$ ; 73.0/73.3% is the ratio of those who succeeded in enrolling), Node 13 ( $11.2 < \text{comprehensive assessment program} \leq 12.18$ ; 63.8/54.2% is the ratio of those who succeeded in enrolling), and Node 14 (comprehensive



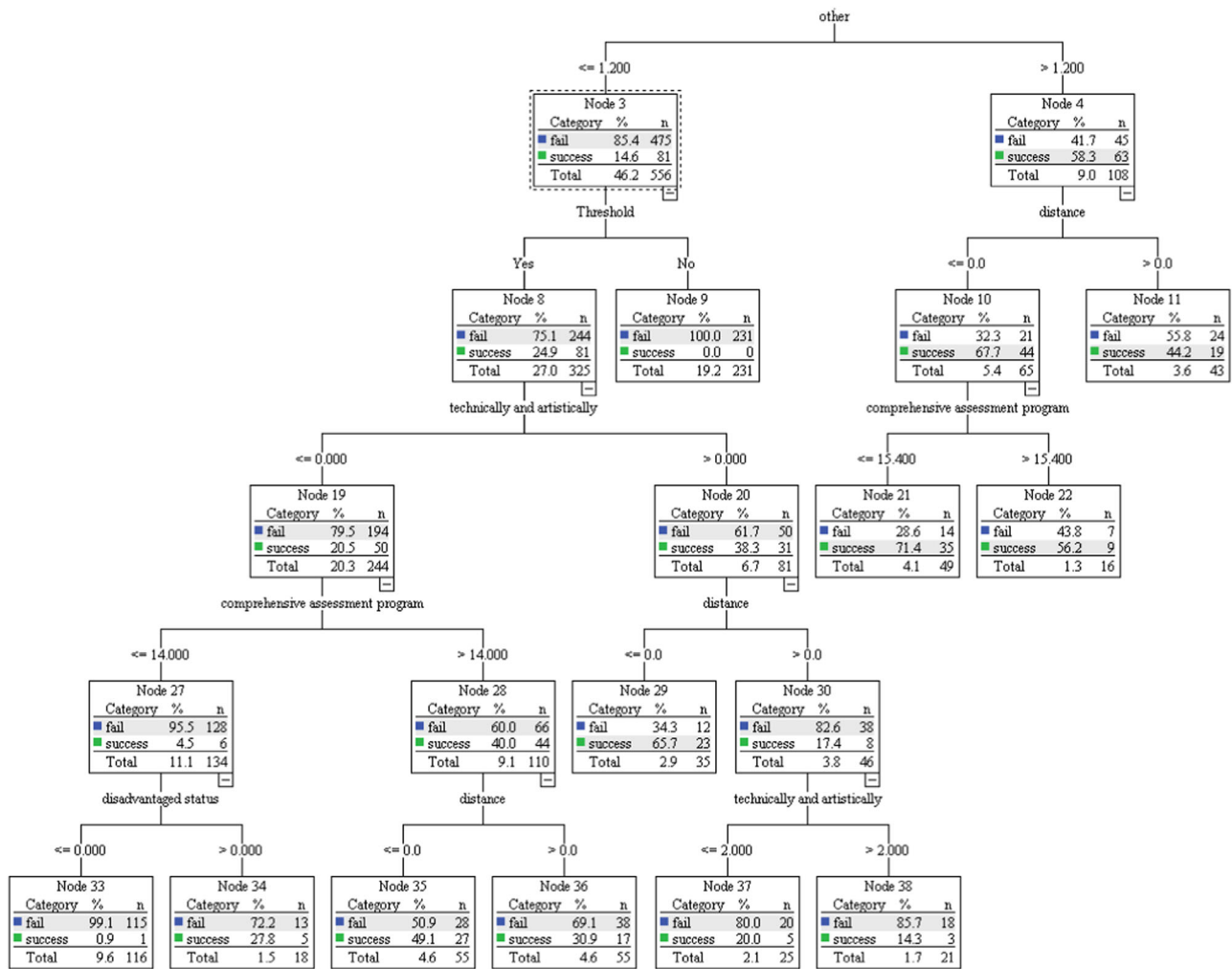


Fig. 6 A sub-tree on left for the test. A sub-tree structure for the test sample in the metropolitan area colleges.

assessment program > 12.18; 51.0/59.4% is the ratio of those who succeeded in enrolling). Comparing the results of the three nodes, it is found that the lowest comprehensive assessment program (Node 12) has the highest ratio of successful enrollees. A higher comprehensive assessment program will instead have a lower ratio of successful enrollees. Therefore, there is a situation of reverse selection of talent.

- I. At Node 12, the fifth level is composed of the branch below Node 12 to Node 23 (writing test ≤ 3; 80.0/73.8% is the ratio of those who succeeded in enrolling) and Node 24 (writing test > 3; 70.4/73.1% is the ratio of those who succeeded in enrolling). It turns out that those with lower scores on a writing test (Node 23) will have a higher ratio of successful enrollees, so there is also a case of talent inverse selection. This phenomenon can occur in agricultural county colleges, which may be the result of an applicant abandoning enrollment and the admissions department creating a strategy to fill up the vacant seats with students.
- II. At Node 14, the fifth level is composed of the branch below Node 14 to Node 25 (Types of junior high-school graduate is the city; 60.0/64.0% is the ratio of those who succeeded in enrolling) and Node 26 (Types of junior high-school graduate include county, private, or national; 73.5/52.6% is

the ratio of those who failed to enroll). Here, it will be found that junior high-school graduates in the city have relatively high success rates at enrollment.

Finally, a possible reason for the emergence of the adverse selection of talent may be that the more qualified applicants gave up this admission opportunity and chose to register at other schools. This represents a potentially difficult problem faced by exam-free admissions colleges in agricultural counties.

- (2) At Node 6, the fourth level is composed of the branch below Node 6 to Node 15 (comprehensive assessment program ≤ 8.4; 46.0/33.3% is the ratio of those who failed to enroll) and Node 16 (comprehensive assessment program > 8.4; 63.5/60.9% is the ratio of those who failed to enroll). A higher comprehensive assessment program has a higher ratio of those who failed to enroll, and there is a phenomenon of inverse talent selection.
- (3) At Node 7, the fourth level is composed of the branch below Node 7 to Node 17 (writing test ≤ 3; 65.1/84.6% is the ratio of those who failed to enroll) and Node 18 (writing test > 3; 83.5/78.2% is the ratio of those who failed to enroll).

Based on the results of Figs. 3 through 8, the seven important decision-making rules are collated in Table 6. This is the result of consolidating some decisions rules from the training group's ratio of predicted failures (or successes) to actual enrolled that reached

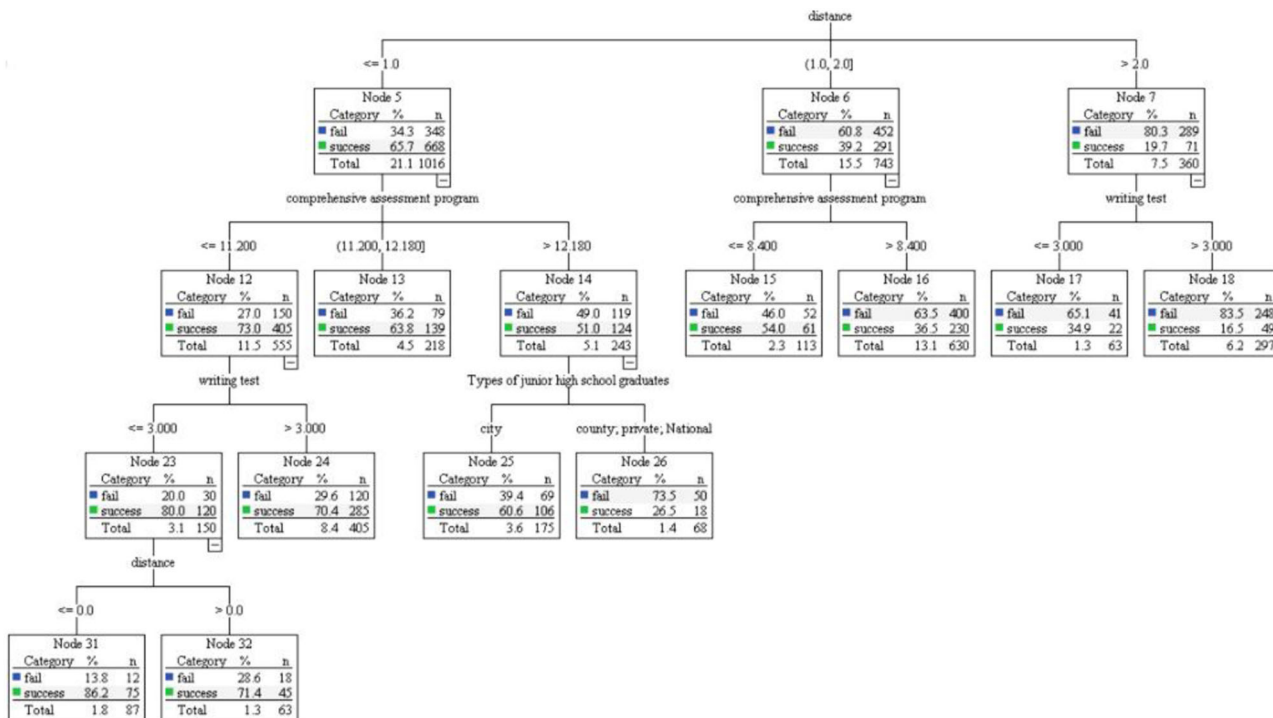


Fig. 7 A sub-tree on right for the training. A sub-tree structure for the training sample in the agricultural county colleges.

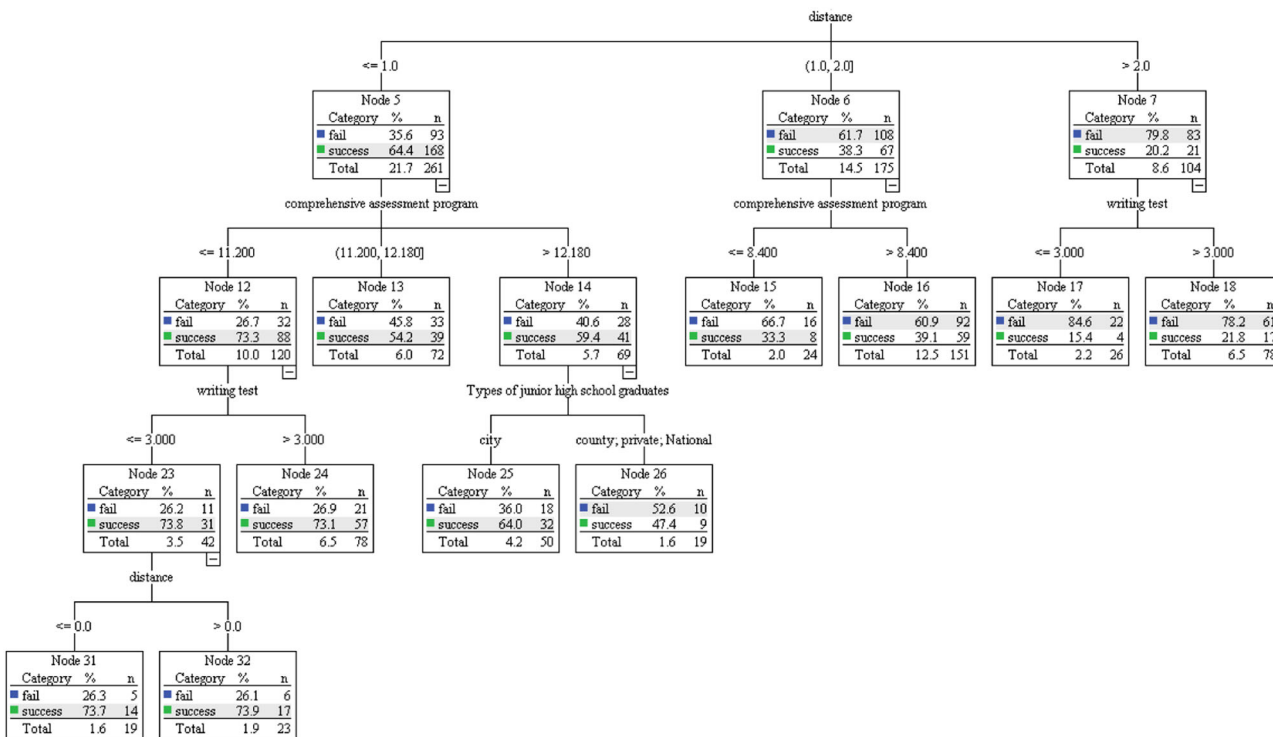


Fig. 8 A sub-tree on right for the training. A sub-tree structure for the test sample in the agricultural county colleges.

an accuracy rate of at least 75% or more. Among them, in Part I Colleges in Metropolitan City, 4 rules determine failure to enroll and 1 rule determines success. Part II Colleges in Agricultural County provides 1 important rule for failed enrollment and 1 important rule for successful enrollment. These rules may be provided to applicants, relatives and friends, as well as to the relevant responsible persons of the exam-free admissions school as a reference.

**Conclusions**

This study used the decision tree analysis method to explore the reasons behind enrollment failures/successes in the joint enrollment and distribution process for Taiwan’s 5-year junior colleges and to observe whether this admissions system fulfills its expected educational objectives. The established training/test model used with the exam-free admissions colleges had a sensitivity rate of 81.9/80.6% for detecting enrollment failures. The location of the

**Table 6 Organization of important rules from the decision tree.**

No.	Observed	Rules	Probability in training/ test
Part I. Colleges in metropolitan city			
1	Fail	WHERE other $\leq 1.2$ AND (threshold = "No")	100/100%
2	Fail	WHERE other $\leq 1.2$ AND (threshold = "Yes") AND ((technically and artistically $\leq 0.0$ ) AND ((comprehensive assessment program $\leq 14$ ) AND ((disadvantaged status $\leq 0$ )))	98.3/99.1%
3	Fail	WHERE other $\leq 1.2$ AND (threshold = "Yes") AND ((technically and artistically $> 0$ ) AND ((distance $> 0$ )) AND ((technically and artistically $\leq 2$ )))	78.1/80.0%
4	Fail	WHERE other $\leq 1.2$ AND (threshold = "Yes") AND ((technically and artistically $\leq 0$ ) AND ((comprehensive assessment program $> 14$ ) AND ((distance $> 0$ )))	75.7/69.1%
5	Success	WHERE other $> 1.2$ AND (distance $\leq 0$ ) AND ((comprehensive assessment program $\leq 15.4$ ))	76.1/71.4%
Part II. Colleges in agricultural county			
6	Fail	WHERE distance $> 2$ AND (writing test $> 3$ )	83.5/78.2%
7	Success	WHERE distance $\leq 1$ AND (comprehensive assessment program $\leq 11.2$ ) AND ((writing test $\leq 3$ ) AND ((distance $\leq 0$ )))	86.2/73.7%

colleges (metropolitan area vs. agricultural county) found in the tree structure is a first-level factor; that is, there is a significant difference between the two categorical colleges for the factors that determine a failure in enrollment.

First, the failed enrollment percentage in metropolitan area colleges is much higher than that of agricultural county colleges. This shows that metropolitan area exam-free admissions colleges still place relatively high competitive pressure on applicants. In the metropolitan area, a college has the advantage of talent selection. The English test results (other) and registration threshold are two important factors in successful enrollment. Then, even if the registration threshold is reached and the other  $\leq 1.2$  then:

- (1) If the technically and artistically gifted (U10) = 0, there was a high probability of approximately 80.0% of failing to enter the college. Attention must be paid to the higher performance of the technically and artistically gifted, which can be accompanied by an improved comprehensive assessment program that can reduce the likelihood of failed enrollment. In addition, there are some safeguards for the conditions of the disadvantaged (see Nodes 33 & 34), and the proportion of enrollment failures due to the distance between the enrollee and the target college in the higher comprehensive assessment program ( $>14.0$ ) (see Nodes 35 & 36) is dependent on the relationship.
- (2) If technically and artistically gifted  $> 0$ , then the proportion of enrollment failures was reduced to 53.2% for the training sample (see Nodes 19 & 20). When the technically and artistically gifted also had a farther distance to their target colleges, the higher the proportion of enrollment failures was (see Nodes 29 & 30). This study found that the enrolling colleges in the metropolitan area have better English test scores (other  $> 1.2$ ), which reduced the proportion of enrollment failures to 35.8% for the training sample (see Nodes 3 & 4); subsequently, the distance (see Nodes 29 & 30) between the enrollee and the target college, as well as the results of the comprehensive assessment program (see Nodes 37 & 38) were secondary factors.

For agricultural county colleges, some applicants retain the flexibility to choose admission, and the colleges may face underenrollment (see Fig. 1; actual enrollment ratios at schools C and D were 56.73 and 89.05%). The distance between the enrollee and the target college is a major factor in enrollment failures (see Nodes 5, 6, & 7 in Figs. 2 and 3); it is found that in the local or neighboring counties and cities of the applicant, the results of the

comprehensive assessment program are not ideal, but the enrollment failure rate is lowered (see Nodes 12, 13, & 14 in Figs. 2 and 3). Moreover, some of the results of the comprehensive assessment program or writing test (see Nodes 23 & 24 in Figs. 2 and 3) show better results among applicants, but the enrollment failure rate is higher, showing that the colleges experience a reverse selection phenomenon for talent. The reason for this phenomenon may be that better-performing applicants prefer to choose a different college or that it is formed by the lax conditions for college admission.

The results of this study show that although the admissions systems for Taiwan's national secondary schools are based on the goal of a multi-intelligence balanced education, under the influence of enrollees' college preferences, the admission competition between metropolitan area colleges and agricultural county colleges is worsened. The enrolling colleges in the metropolitan areas have a greater ability to make selections, and the rural enrolling colleges in the agricultural county have been marginalized by the reverse selection of talent. This phenomenon of reverse selection of talents will not be conducive to agricultural enrollment schools to select excellent talents for education, cannot effectively use educational resources, it is not conducive to the local cultivation of talents. To address the under-enrollment of colleges in agricultural counties, it is proposed that the Wu (2020) approach should be used to recommend enrollment policies in agricultural county college districts that give students guaranteed or priority admission to improve student enrollment at schools of their choice.

**Data availability**

Data sharing is not applicable to this research as no data were generated or analyzed.

Received: 10 May 2022; Accepted: 12 October 2022; Published online: 25 October 2022

**References**

Amburgey WOD, Yi J (2011) Using business intelligence in college admissions: a strategic approach. *Int J Bus Intell Res* 2(1):1–15  
 Asif R, Merceron A, Ali SA, Haider NG (2017) Analyzing undergraduate students' performance using educational data mining. *Comput Educ* 113:177–194  
 Aulck L, Nambi D, Velagapudi N, Blumenstock J, West J (2019) Mining university registrar records to predict first-year undergraduate attrition. In: *Proceedings*

- of the 12th international conference on educational data mining (EDM 2019). International Educational Data Mining Society, Montreal, Canada, pp. 9–18
- Berry MA, Linoff G (1997) Data mining techniques for marketing, sales, and customer support. Wiley and Sons, New York
- Chou CP (2009) Toward a twelve-year basic education program in Taiwan. *Bull Educ Resour Res (Taiwan)* 42:25–42
- Delibasic B, Vukicevic M, Jovanovic M, Suknovic M (2013) White-box or black-box decision tree algorithms: which to use in education? *IEEE Trans Educ* 56(3):287–291
- Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. *AI Mag* 17(3):37–54
- Gardner H (2011) Intelligence, creativity, ethics: reflections on my evolving research interests. *Gift Child Q* 55(4):302–304
- Han S (2022) Identifying the roots of inequality of opportunity in South Korea by application of algorithmic approaches. *Humanit Soc Sci Commun* 9(1):18
- Howard E, Meehan M, Parnell A (2018) Contrasting prediction methods for early warning systems at undergraduate level. *Internet High Educ* 37:66–75
- Kingsford C, Salzberg SL (2008) What are decision trees? *Nat Biotechnol* 26(9):1011–1013
- Kirby NF, Dempster ER (2014) Using decision tree analysis to understand foundation science student performance. Insight gained at one South African university. *Int J Sci Educ* 36(17):2825–2847
- Kiss B, Nagy M, Molontay R, Csabay B (2019) Predicting dropout using high school and first-semester academic achievement measures. In: 2019 17th international conference on emerging elearning technologies and applications (ICETA). IEEE, Starý Smokovec, Slovakia, pp. 383–389
- Križanić S (2020) Educational data mining using cluster analysis and decision tree technique: a case study. *Int J Eng Bus Manag* 12:184797902090867
- Lee L, Liu YS (2021) Use of decision trees to evaluate the impact of a holistic music educational approach on children with special needs. *Sustainability* 13(3):1410
- Lin CF, Yeh YC, Hung YH, Chang RI (2013) Data mining for providing a personalized learning path in creativity: an application of decision trees. *Comput Educ* 68:199–210
- Lynch CF (2017) Who prophets from big data in education? New insights and new challenges. *Theory Res Educ* 15(3):249–271
- Maltz EN, Murphy KE, Hand ML (2007) Decision support for university enrollment management: implementation and experience. *Decis Support Syst* 44(1):106–123
- Ministry of Education (2009) The implementation plan for expansion of application admission on the high school and five-year junior college. Ministry of Education in Taiwan, Taipei City
- Nagy M, Molontay R (2018) Predicting dropout in higher education based on secondary school performance. In: 2018 IEEE 22nd international conference on intelligent engineering systems (INES). IEEE, Las Palmas de Gran Canaria, Spain, pp. 389–394
- Oranye NO (2016) The validity of standardized interviews used for university admission into health professional programs: a Rasch analysis. *SAGE Open* 6(3):215824401665911
- Park E, Dooris J (2020) Predicting student evaluations of teaching using decision tree analysis. *Assess Eval High Educ* 45(5):776–793
- PhridviRaj MSB, GuruRao CV (2014) Data mining—past, present and future—a typical survey on data streams. *Procedia Technol* 12:255–263
- Ragab AHM, Mashat AFS, Khedra AM (2012) HRSPCA: hybrid recommender system for predicting college admission. In: 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA). IEEE, Kochi, India, pp. 107–113
- Rastrollo-Guerrero JL, Gómez-Pulido JA, Durán-Domínguez A (2020) Analyzing and predicting students' performance by means of machine learning: a review. *Appl Sci* 10(3):1042
- Singer G, Golan M, Rabin N, Kleper D (2020) Evaluation of the effect of learning disabilities and accommodations on the prediction of the stability of academic behaviour of undergraduate engineering students using decision trees. *Eur J Eng Educ* 45(4):614–630
- Tanna M (2012) Decision support system for admission in engineering colleges based on entrance exam marks. *Int J Comput Appl* 52(11):38–41
- Vialardi C, Chue J, Peche JP, Alvarado G, Vinatea B, Estrella J, Ortigosa Á (2011) A data mining approach to guide students through the enrollment process based on academic performance. *User Model User Adapt Interact* 21(1-2):217–248
- Waterhouse L (2006) Multiple intelligences, the Mozart effect, and emotional intelligence: a critical review. *Educ Psychol* 41(4):207–225
- Wu MJ (2020) Predicting outcomes of school-choice policies using district characteristics: empirical evidence from Hong Kong. *J School Choice* 14(4):633–654
- Yao G, Wang J, Cui B, Ma Y (2022) Quantifying effects of tasks on group performance in social learning. *Humanit Soc Sci Commun* 9(1):282
- Zeng X, Yuan S, Li Y, Zou Q (2014) Decision tree classification model for popularity forecast of Chinese colleges. *J Appl Math* 2014:675806

### Acknowledgements

The authors would like to thank Mr. Kuo-En Wang (Director of the Admissions Center) for providing anonymous participants registration data set in this study.

### Author contributions

Y-SL designed the study, analyzed the data, and wrote the manuscript. LL contributed to the study supervision and project administration. The authors read and approved the final manuscript.

### Competing interests

The authors declare no competing interests.

### Ethical approval

The study follows the principles of the Declaration of Helsinki. This article does not contain any studies with human participants or animals performed by the authors. This study is an educational issue and does not involve human experiments.

### Informed consent


This article does not contain any studies with human participants performed by any of the authors

### Additional information

Correspondence and requests for materials should be addressed to Ying-Sing Liu.

Reprints and permission information is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022