International Journal of Computer Applications (0975 - 8887)
Volume 67 - No. 21, April 2013

Mining Students' Characteristics and Effects on University Preference Choice: A Case Study of Applied Marketing in Higher Education

Muhammed Basheer Jasser*
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Fatimah Sidi*
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Aida Mustapha
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Abdulelah Khaled T Binhamid
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

University servers and databases store a huge amount of data including personal details, registration details, evaluation assessment, performance profiles, and many more for students and lecturers alike. Mining such data offers a huge potential in advancing the educational field in the country because data mining is able to extract important models and hidden patterns beneath the data, which will help in decision-making to improve the outcome of educational establishments. This work concerns with data related to students. Understanding the characteristics of students enrolled in the university is important as it helps the university or institution to strategize on marketing their education programmes. This paper analyzes the student characteristics of Universiti Putra Malaysia based on their preference choice during registration at the university. The experiments are carried out using the Oracle Data Miner software and the results are analyzed and discussed.

General Terms:

Educational Data Mining

Keywords:

Classification, Decision Tree, Oracle Data Miner

1. INTRODUCTION

As the higher educational system at university level becomes more automated and computerized, higher amount of data are being generated and are made available. For example, a student course registration system stores course selection list for every student and a learning management system helps students to evaluate a specific course or provides feedback about specific lecturer. From both system databases, historical records for a specific lecturer teaching a specific course could be extracted for the purpose of performance analysis and performance evaluation.

One of the most important issues to be addressed in any educational establishment is the student behaviors such as their academic performance, course selections, added and dropped courses or preference level. Another example is the university registration system, which stores data concerning student characteristics such as personal details and qualifications of the new applicants. This data contains numerous hidden patterns that could be used to strategize the management effort to produce better graduates.

Among the previous works include a student retention model to analyze the factors that affect students decision to further their higher education as mentioned by Tinto [6].Thomas and Galambos[5] proposes a student satisfaction model that relates how students situation and characteristics affect their satisfaction towards some university or faculty. This study also analyzes how the university facilities relates to student achievement. Both works demonstrate the potential use of such model in targeting the management effort towards specific issues at hand. While decision making in managing the students and resources are important, an equally important task is planning and managing the marketing cost of educational programmes. As the cost of marketing in recent years is increasing, universities could capitalize on data mining tasks in such cost.

This paper attempts to model relationship between the student profiles with their university preference choice at the application level using the student application database. The university preference choice will show how the applied university is ranked among other universities in the country as preferred by the students. Studying the applicant characteristics will help targeting the marketing efforts and thus saving costs concentrating in a smaller group of potential students. The remaining of this paper is structured as follows. Section 2 introduces previous works about data mining applications in higher education. Section 3 provides an overview on the classification task in data mining using the decision tree algorithm. Section 4 details out the experiments and the Oracle Data Miner tool, Section 5 discusses the results and and finally Section 6 concludes the paper.

2. RELATED WORK

Data mining has been widely applied in the higher education field as universities provide huge masses of data. Some of the application is to study factors that affect student retention through monitoring the academic behavior and providing powerful strategies to intervene as proposed by C. H. Yu et al. [8]. The study

showed that some demographic factors like transferred hours, residency, and ethnicity are very important to predict retention. The classification decision tree of their study is shown in Figure1

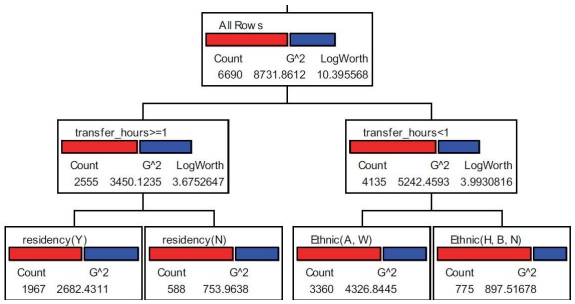

Fig. 1. Classification Decision tree of Retention.

Yadav and Pal [7] use the ID3 decision tree to generate the important rules that can help to predict student enrollment into an academic programme called the Master of Computer Application. The generated tree yields that bachelor of science students in mathematics and computer applications will enroll and will likely to perform better as compared to bachelor of science students without any background in mathematics.

Aher and Lobo[1] uses data from undergraduate student exam grades to cluster the students and predict their performance in future tests using decision tree algorithm. Research has also shown the benefit of data mining applications in a course management system for online instructors as in the study proposed by C. Romero et. al [4] whereby the instructors are able to view statistical graphs and to apply clustering task to find similar characteristics among a group of students using data generated from a system. Table1 shows the details of each work including the objective, datamining task applied, data source and the tool used.
In this paper, the objective is to predict the student's university preference choice and focus on the characteristics of student with low preference value. Oracle data miner ODM will be used on the data obtained from university students registration profiles.

3. DECISION TREE CLASSIFICATION

Classification is one of data mining task using statistical and machine learning algorithms to predict group membership for data instances. One common example is to predict weather for a day, whether it will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks [2].

The classification algorithm used in this case study will be the decision tree, which is represented as a graph of nodes and branches whereby in every branch a decision is made and the consequence is presented in the resulting node connected to it. Meanwhile, the leaf nodes represent the classes [8]. Decision tree is chosen because the algorithm is able to generate rules that explain every condition and step that has been taken in performing the classification task.

The objectives of our proposed classification task using the decision tree algorithm is two-fold. The first is to know the preference level of new enrolling student applicants. The second is to use the generated rules from the decision tree and focus on characteristic of applicants that has low preference level to the particular university with data on hand. Since this study is interested in all the details about specific classes, decision tree is considered the ideal solution to achieve the objectives.

4. EXPERIMENTS

The classification experiment proposed in this paper attempts to inevstigate the characteristics of students who did not choose Universiti Putra Malaysia (UPM) as their best preferred choice in effort to strategize for targeted marketing for similar group of potential students in the future.

4.1 Dataset

The dataset is sourced from student application database into UPM, which consists of rich information ranging from applicants personal information, qualifications, English tests, and applicants course choices. The main targeted attribute to be studied will be the preference choice in the applicant course choice table. Figure 2 shows the original table with all applicant attributes.

Attribute	Attribute
ID	APPLIED_YEAR_ID
SESSION_ID	LEVEL_OF_STUDY_ID
SEMESTER_ID	METHOD_OF_STUDY_ID
BUMISTATUS_ID	APPLICANT_ID
CITIZEN_ID	DOMAIN_ID
CITIZENSHIP_STATUS_ID	CHOICE_ID
GENDER_ID	PROG_APPLIED_ID
MARITALSTATUS_ID	PROG_QUALIFIED_ID
PARLIAMENT_ID	APPLIED_STATUS1_ID
RACE_ID	CGPA_RANGE_ID
DISABILITIES_ID	AGE_APPLIED
HANDICAP_ID	CGPA_ENTRY
BIRTHDATE_ID	MERIT_SCORE
BIRTH_STATE_ID	PROGRAM_OFFER_ID
BIRTH_COUNTRY_ID	INSTITUTION_ID
PASSPORT_COUNTRY_ID	MAJOR_ID
ORIGIN_COUNTRY_ID	MINOR_ID
PERM_STATE_ID	FIELD_OF_STUDY_ID
PERM_COUNTRY_ID	APPLIED_STATUS2_ID
MAILING_STATE_ID	APPLIED_STATUS3_ID
MAILING_COUNTRY_ID	APPLIED_STATUS4_ID
FAMILY_INCOME_ID	APPLIED_STATUS5_ID
NEXT_OF_KIN_ID	APPLIED_STATUS6_ID
APPLIED_DATE_ID	APPLIED_STATUS_ID

Fig. 2. Attributes from Faculty Applicant Course Choice.

However, while this table includes a lot of attributes, only a handful is useful for the proposed classification task. Figure 3 shows the final attributes after the attribute removal process, carried was carried out manually.

Attribute
ID
AGE_APPLIED
LEVEL_OF_STUDY_ID
BUMISTATUS_ID
GENDER_ID
MARITALSTATUS_ID
PROG_APPLIED_ID
CGPA_RANGE_ID
CHOICE_ID

Fig. 3. Final attributes after attribute removal process.

Based on Figure 2, the following are details of the attributes.

—`LEVEL_OF_STUDY_ID` contains twelve possible distinct numerical values each represent the study level owned by the applicant, some of these values are certificate, bachelor and master.

Table 1. Related Works Details.

Work	Objective	Applied Data Mining Task	Data Source	Tool
C. H. Yu et al. [8]	Study the factors affecting university student retention	Classification of decision tree, multivariate adaptive regression splines (MARS), neural networks	Track of continuous enrollment or withdrawal of students enrolled at university	JMP
Yadav and Pal [7]	Predict good student enrollment	Classification of decision tree	Data of 432 students of the Department of MCA	WEKA
Aher and Lobo [1]	Improve students performance	Classification and clustering	Database of final year students for Information Technology UG course	WEKA
C. Romero et. al [4]	Show the benefits of applying data mining in course management systems	Different tasks: classification, clustering	Course management learning system (Moodle data)	WEKA/ Keel

—`PROG_APPLIED_ID` contains also identifiers to program type offered by the university including all faculties and levels.

—`CGPA_RANGE_ID` includes eight distinct numerical identifiers each represents a CGPA range, for example the identifier value eight represents the range 3.5-4, knowing that this university follows the scale of CGPA 1 to 4 in evaluating students.

—`BUMISTATUS` represents the locality of the applicant, in other words the state of living is far from the university location of not. This attribute helps to study the effect of living location in predicting the preference choice of UPM.

—`CHOICE_ID` represents the applicant preference choice in a set of eight distinct values 1 to 8, the lower value means higher priority for example the value 1 means that the applicant prefers UPM as the best university.

`AGE_APPLIED`, `GENDER_ID`, and `MARITALSTATUS_ID` attributes include distinct numerical values and are self-explanatory.

4.2 Data Preprocessing

Looking at the final attributes in Figure 3, note that some of the chosen attributes contain missing values and need to be handled in order to produce the best result. More specifically, `AGE_APPLIED` is handled by replacing the missing values with the attribute mean. Missing values for `GENDER_ID` are replaced with the attribute mode. This is similar to other attributes such as `MARITALSTATUS_ID`, `PROG_APPLIED_ID`, `CGPA_RANGE_ID`, `BUMISTATUS`, and `LEVEL_OF_STUDY_ID`. Table 2 details out the process for each attribute.

Table 2. Treatment for missing values.

Attribute	Type	Treatment
AGE_APPLIED	Ratio	Null values replaced by mean
GENDER_ID	Nominal	Null values replaced by mode
MARITALSTATUS_ID	Nominal	Null values replaced by mode
PROG_APPLIED_ID	Nominal	Null values replaced by mode
CGPA_RANGE_ID	Ordinal	Null values replaced by mode
BUMISTATUS	Nominal	Null values replaced by mode
LEVEL_OF_STUDY_ID	Nominal	Null values replaced by mode

Based on Table 2, mean represents the average from total sum of attributes while the mode represents the most frequent value. Note that missing values for ratio attribute are handled using the attribute mean while the nominal and ordinal types are handled by the attribute mode.

4.3 Oracle Data Miner

Based on the selected attributes, the classification experiment is carried out using Oracle Data Miner (ODM) Tool [3]. ODM provides powerful data mining functionality as native SQL functions within the Oracle Database. It is a component of the Oracle Advanced Analytics option that helps companies to do conduct better analysis on business or organizational data. The Oracle Data Work Flow is Graphical User Interface (GUI)-based, hence providing easy navigation to building and evaluating models, apply the data mining models to new data, and save and share their analytical results.

Because of this, the ODM tool is able to build and apply predictive models directly from the star schema data from the Oracle database. Therefore, the models and results remain in the Oracle database, thus eliminating the need for data movement, minimizing the information latency, while maintaining the security. Potential use of the ODM is building predictive models that help to target best customers, develop detailed customer profiles or detect and prevent fraudulent actions during financial transactions.

The choice of ODM for this research is also influenced by the nature of the UPM student data, which resides in an Oracle database. Figure 4 shows the main user interface of oracle data miner including some built model and workflow processes.

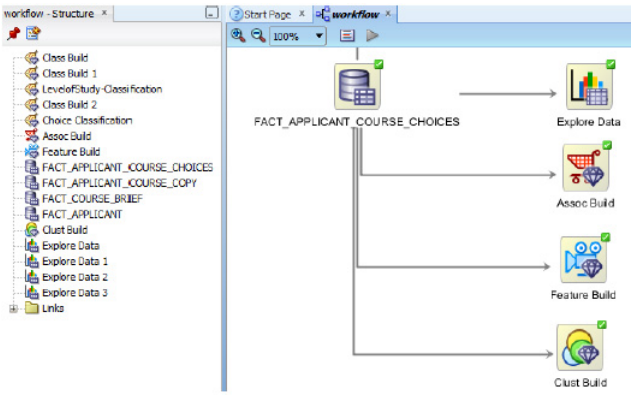

Fig. 4. Interface of Oracle Data Miner (ODM).

5. RESULTS AND DISCUSSIONS

A decision tree algorithm was applied for classification task using the Oracle Data Miner tool with the model parameters as shown in Figure 5. The homogeneity metric represents the way of selecting the split node, which has two possible values Gini and Entropy. In this experiment, the Entropy is chosen because it allows the decision tree to have multiway splits while Gini will enforce it to be binary [4].

The maximum depth represents the longest path depth from the root to the leaves. In this experiment, it is set to 5 because assigning less value will make the decision tree useless since the rules generated will not provide much detail about a specific class or leaf node. Having higher depth value makes the generated rules with less support since the database instances are more scattered over the tree nodes. Other model attributes like minimum records in a node and minimum records in a split represents the support

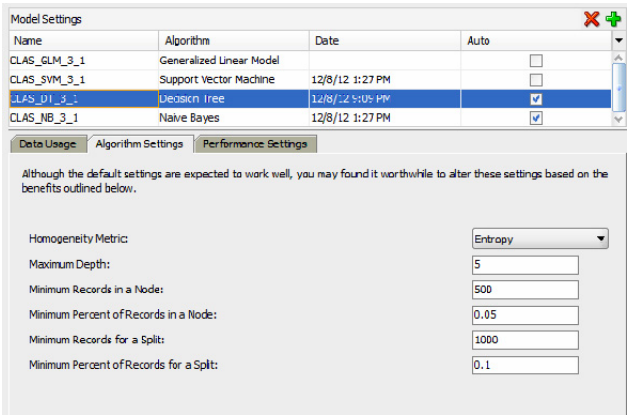

Fig. 5. Model parameters in ODM.

for specific attribute value. We notice that the model parameters must be chosen carefully to get the most accurate result. The resulted decision tree is shown in Figure 6.

Based on Figure 6, every node in the decision tree has its own classification rule. This rule consists of a number of conditions that helps in tracing one data instance to the right tree node. The condition sequence length is proportional with the node level. For example `node10` has the following rule: if `BUMISTATUS_ID` isIn (2) And `LEVEL_OF_STUDY_ID` isIn (3) Then 1.

For more accurate results we consider the rules of leaf nodes since the main idea is to consider the tree depth of 5 in order to consider more node splits with more conditions otherwise less depth value could be considered. Table 3 shows the rules of every leaf node. Note that the numbers appearing in the rules are only representative of the actual values except the applicants attribute.

The main goal of this study is to address the characteristics of student who does not choose Universiti Putra Malaysia (UPM) as their first preference. From the results, this means the interesting values reside in the range 4 to 8 so only rules that produces a value within the range is considered. It is noted in Table 2 that `node11` gives the value 6, which is the only interesting one representing the important characteristic to be studied. All other rules belongs to the applicants that have UPM as their choice within the accepted range 1 to 4.

The rule for `node11` is: if `LEVEL_OF_STUDY_ID` isIn (5, 6) And `PROG_APPLIED_ID` ¡= 50.5 Then 6. The values 5,6 of `LEVEL_OF_STUDY` and 50.6 for `PROG_APPLIED_ID` represent foreign keys to other tables that contains the actual values. This is shown in Figure 7. After replacing the corresponding values from the reference tables, the modified rule will be: if `LEVEL_OF_STUDY` isIn (Bachelor, Diploma) And `PROG_APPLIED` isIn (1, 2 until 50) Then 6.

LEVEL_OF_STUDY_ID	LEVEL_OF_STUDY_CODE	LEVEL_OF_STUDY
1	02	MATRICULATION
2	17	STPM
3	MASTER	MASTER
4	PHD	PHD
5	04	BACHELOR
6	03	DIPLOMA
7	16	SPM
8	07	ASASI
9	CERTIFICATE	CERTIFICATE
10	12	PRA DIPLOMA
11	NONGRADUATING	NONGRADUATING

Fig. 7. Actual values from foreign keys.

6. CONCLUSION

This paper presented an application of Oracle Data Miner (ODM) tool to perform decision tree classification using student application database of Universiti Putra Malaysia (UPM). The aim of this work is to study the characteristic of students who does not choose UPM as their preferred choice of educational provider. After data preprocessing and cleaning, a decision tree was constructed to investigate the student characteristics. The classification experiments showed that student choice is mostly affected by level of study and the program type they applied for. Based on this finding, an improvement could be done to organize specific programs so as to increase student interest and reputation of the university. UPM student database is indeed a rich repository of hidden models and relationships that could be capitalized to build different models for other types of problems or interests.

7. ACKNOWLEDGEMENT

This work was partially funded by eNCoral Digital Solution Sdn. Bhd. Corresponding authors: Fatimah Sidi(fatimah@upm.edu.my), Muhammed Basheer Jasser(mbjasser@gmail.com)

8. REFERENCES

[1] S. B. Aher and L. M. R. J. Lobo. Data mining in educational system using weka. In *IJCA Proceedings on International Conference on Emerging Technology Trends*, pages 20–25, New York, 2011. Foundation of Computer Science.

[2] J. Han, M. Kamber, and J. Pei. *Data Mining Concepts and Techniques*. Morgan Kaufman, San Francisco, 3rd edition, 2006.

[3] Oracle. Odm: Oracle data miner. http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html.

[4] C. Romero, S. Ventura, and E. Garcia. Data mining in course management systems: Moodle case study and tutorial. 2008.

[5] E. H. Thomas and N. Galambos. Article: What satisfies students? mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, 45(3), May 2004.

[6] V. Tinto. Article: Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45:89–125, 1975.

[7] S. K. Yadav and S. Pal. Data mining application in enrollment management: A case study. *International Journal of Computer Applications*, 41(5), March 2012.

[8] C. H. Yu, S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8:307–325, 2010.

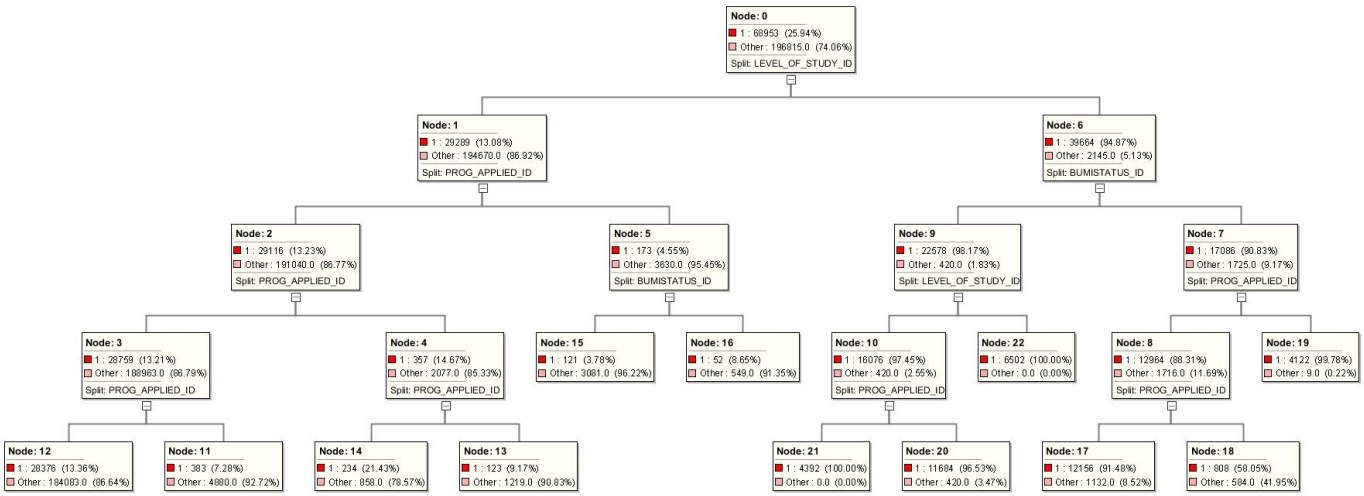

Fig. 6. Resulting decision tree.

Table 3. The rules of the decision tree leaves at depth 4.

Node	Rules
node12	If LEVEL_OF_STUDY_ID isIn (5, 6) And $50.5 <$ PROG_APPLIED_ID $<= 1464$ Then 2
node13	If LEVEL_OF_STUDY_ID isIn (5, 6) And $1464 <$ PROG_APPLIED_ID $<= 1469.5$ Then 3
node20	If BUMISTATUS_ID isIn (2) And LEVEL_OF_STUDY_ID isIn (3) And PROG_APPLIED_ID $<= 240.5$ Then 1
node18	If LEVEL_OF_STUDY_ID isIn (3, 4) And BUMISTATUS_ID isIn (1) And $239.5 <$ PROG_APPLIED_ID $<= 240.5$ Then 1
node11	If LEVEL_OF_STUDY_ID isIn (5, 6) And PROG_APPLIED_ID $<= 50.5$ Then 6
node14	If LEVEL_OF_STUDY_ID isIn (5, 6) And $1469.5 <$ PROG_APPLIED_ID $<= 1492$ Then 1
node21	If BUMISTATUS_ID isIn (2) And LEVEL_OF_STUDY_ID isIn (3) And PROG_APPLIED_ID > 240.5 Then 1
node19	If LEVEL_OF_STUDY_ID isIn (3, 4) And BUMISTATUS_ID isIn (1) And PROG_APPLIED_ID $<= 239.5$ Then 1