# Using Affinity Analysis-Driven Adaptive Data Mining Life Cycle for the Development of a Student Retention DSS

PI-SHENG DENG
Management Information Systems
California State University, Stanislaus
Turlock, CA 95382
USA

*Abstract:* Technological development has engaged educational institutions in fierce global competition. To be competitive in meeting the changing needs of today's student population, educational institutions find it imperative to prioritize student retention efforts and to develop strategies that interact and serve students more effectively in providing them more value and service. In this research we proposed a three-phase-six-stage adaptive data mining development life cycle, and we applied the affinity analysis to this methodology in identifying more than 400 association relationships with student retention, refining iteratively the association rule set down to less than 30 rules, and developing useful strategic implications regarding how the important factors were associated with a student's decision. This set of implications and factors could then be integrated into the development of strategies for student retention.

*Key-Words*: Affinity analysis; Association rules; Adaptive Data Mining Development Cycle; Student retention

## 1 Introduction

In recent years, the Internet/Web technology has enabled organizations to provide their services or information to customers or clients whom they could not reach easily before. This has elevated the competition among organizations to the global level. Without exception, educational institutions worldwide have also experienced the unprecedented impact caused by the Interne array of information available to them on the Web, students can now compare easily almost every aspect of different institutions in making their education decisions. Such technological development has engaged educational institutions in fierce global competition. As a consequence, concept of the service region of an institution has become less clear. To meet the changing needs of today's student population, educational institutions find it imperative to prioritize student retention efforts and to develop strategies that serve students more effectively.[1][2] Global competition and the rise of digital technology have made educational institutions think strategically about their institutional processes for managing their relationships with students and other stakeholders. To be competitive, educational institutions need to provide more value and service to its stakeholders, and improve institutional processes and programs for interacting with their stakeholders, especially students.[1]

Students are long-term assets of an educational institution, and the relationship with them should be nurtured through institutional processes or programs, such as institutional discourse, student services, outreach, and educational programs. Student retention is "the process of helping students meet their own needs so they will persist in their education toward the achievement of the educational aims they value,"[3] and has been a significant measure of the effectiveness of institutional processes and programs.[1][4] Student retention focuses on managing all of the ways that an educational institution interacts with its existing and potential new students, and is often regarded as one of the most important indicators for assessment of institutional performance and commitment to student success in undergraduate education.[5][6] It is also one of the most challenging issues for higher

education institutions nationwide, and even worldwide.[2][7][8]

This research is motivated by the resource-demanding efforts being devoted to student outreach and retention, yet with marginal return. We employed data mining techniques in this research to analyze student demographic data and student profile for discovering hidden trends or patterns of the *antecedent* ➔ *consequent* relationships between college-related characteristics, activities and a student's decisions on college selection, transfer, or continuance. The discovered relationships can be incorporated into the development of strategies for coordinating institutional processes so that the institution can enhance its relationships with its stakeholders, especially students, and allocate student retention resources and efforts more effectively.

## 2. Background on Data Mining and Student Retention

In recent years, an increasing number of researchers started to apply data mining and machine learning techniques to the study of student retention issues. Data mining employs a set of statistical and machine learning techniques for exploring and extracting useful and meaningful patterns or relationships from a large dataset.[9][10][11] Data mining draws heavily on statistics techniques, especially linear regression, logistic regression, discriminant analysis, and principal components analysis.[12] In addition, data mining also includes techniques from artificial intelligence (AI), such as decision trees, production rules, neural networks, fuzzy logic, and genetic algorithms.[13]

Rather than simply assuming that one technique is analytically superior to others, Dey and Astin[14] studied how logistic regression, probit analysis, and linear regression compared in predicting college student retention. Results indicate that though the former two offered theoretical advantages, they showed little practical advantages over traditional linear regression. Delen[15] developed a data mining model to predict at-risk students and to explain the reasons behind student attrition so that college can intervene to retain them. This study showed the educational and financial variables were among the most important predictors. Similarly, Villano et al.[16] employed survival analysis to develop a model for identifying students of high risk of dropping out by using demographic, institution, student GPA and workload variables. Grayson[17] employed logistic regression analysis to investigate the relationship between first-year student retention and factors, such as full-time status, ethnicity, and GPA, and it was found that there was no significant relationship between the retention rate and ethnicity. Jia and Mareboyana[18] applied decision trees, support vector machines (SVM), and neural networks, to investigate the main factors that influence undergraduate student retention in the historically black colleges and universities. The investigation revealed that cumulative grade point average (GPA) and total credit hours were two main factors affecting a student's decision.

In their studies, Yu et al.[19] identified transfer status, residency, and ethnicity as crucial factors to retention. Wetzel et al.[20] proposed a model of retention for studying which factors would affect a student's decision to stay with a college until graduation. Their findings indicated that academic and social integration factors were found to be the most significant factors in persistence in these years. Financial considerations were of less importance in the persistence decision.

In addition to identifying the factors related to a student's dropout decision, Jung et al.[21] applied marketing concepts to help institutions of higher education to align educational and service processes more closely to their students for alleviating student retention issues. Rahman[22] claimed that the selectivity of a college was not the sole factor affecting student retention. The contributory factors include six institutional initiatives, such as academic advising and new student orientation program.

Most of the research studies have been using parametric techniques for predicting retention decisions. The parametric approach is not adaptive, in terms of its inability to allow the revision of parameters without re-running the parametric model. In addition, most of machine learning techniques, including data mining, fall short of providing end users easy-to-understand transparent results,[23] like in the rule-based format. In this research we applied the rule-based affinity analysis, which has been relatively less employed for student retention strategy development, to iteratively refining data mining development cycle in providing actionable information in the form of *antecedent* ➔ *consequent* rules, which can be combined for generalization, specialization or reduction, to help academic institutions in developing strategies for student retention.

It is not the intention of this study to compare the prediction performance of various different data mining techniques. Our intention is to show the feasibility of using the affinity analysis for developing a decision support system for student

retention strategies. The rest of this manuscript is organized as follows: in the next section, we describe a three-phase-six-stage iteratively refining data mining development cycle. We, then, collected data from a small~medium-sized college in California to illustrate the execution of the development cycle in the next following section. We especially highlighted the importance of the iterative refinement process for the generated rule-based model by affinity analysis. We, then, discussed the strategic implications derived from the current models. Finally, a conclusion section was used to discuss the important findings of this study and some possible future research directions.

## 3. A Description of Adaptive Data Mining Development Life Cycle

In this section, we proposed a three-phase six-stage adaptive data mining development cycle, as shown in Fig. 1, in guiding our application of data mining techniques to investigate the relationship between student attributes and the student's decision outcome. The discovered relationship can be incorporated into the administrative processes a college for developing strategies for student retention.

In Fig. 1, the "pre-data mining" phase includes the work that has to be done before the application of data mining techniques and models. This phase consists of two stages: the *data investigation stage* and the *data pre-processing stage*. The data investigation stage allows us to develop an understanding of the nature and purpose of the data mining project, and the project-related data needs and data sources in an organization. The second stage is the *data pre-processing stage*. In this stage we perform data cleaning and organizing, distribution fitting, and descriptive statistics (such as average, standard deviation, median, mode, minimum, maximum). This allows us to see how variation in scale across variables, the skewness on each variable, outliers, etc.

Next is the data mining phase. This phase is characterized by two continually interacting processes that result in a refining knowledge base, as shown in Fig. 2. These two processes are *data mining modeling* (functionally represented by the DM system developer) which allows us to develop models based on given datasets, and the next process is *model refinement* (functionally represented by the administrator) which fine-tunes the developing models. Through the continual fine-tuning interaction between these two mutually complementary processes, a rule-based knowledge base is formed. This knowledge base is the accumulated result of the modeling process and whose content will continually be refined with new insights gained through the interactive modeling-refining process.

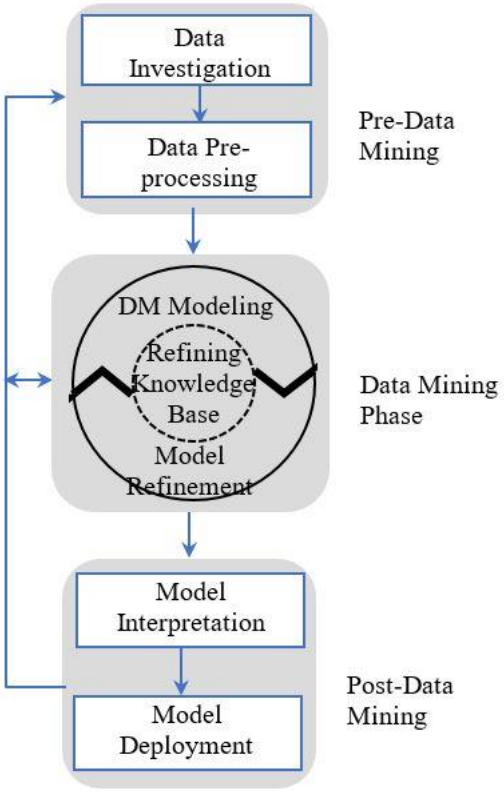


**Fig. 1.** Data mining development cycle as an iteratively refining three-phase six-stage process.

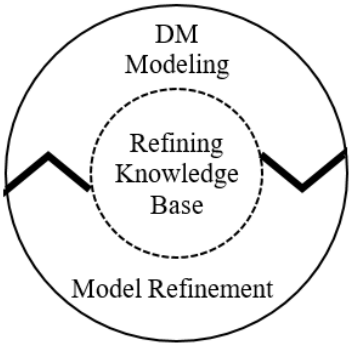


**Figure 2.** A refining knowledge base as a result of two continually interactive processes. Adapted from the sharing-enabled continuous-learning model (Deng, 2008).

The last phase is the post-data mining phase which consists of model interpretation and deployment. As pointed out by Du et al. [24], one of the major limitations of today's machine learning techniques is the lack of transparency behind their behaviours. Thus, the model developed in the previous phase must be understandable to users and the model performance is satisfactory, then the model will be incorporated into the business processes for daily operational decision support. Otherwise, we need to select another model or even need to go back to the first phase to choose another pre-processing technique.

# 4 An Illustration of Applying Affinity Analysis to Data Mining

The case used in this section is a small to medium-sized four-year public university in California. In this section we illustrated the application of affinity analysis to the collected student data set by following the data mining life cycle. The data mining task in this study belongs to the association rules or affinity analysis problem and is a type of classification problem. The association rules technique has been applied to discover general associations patterns among items in large databases[25][26].

Due to the Human Subject research policy, the original demographic data of individual students were not accessible to those without the approval of the UIRB Committee. The data set we obtained consisted of 1,000 student records and thirteen attributes. All of the attributes were categorical variables, including binary variables.

## *4.1. Pre-Data Mining Phase*

Most of the current data mining techniques require categorical variables be pre-processed before the application of the techniques. One alternative is to transform a categorical variable into a series of dummy binary variables. For example, ***Major*** has values of "ALS" (Arts, Letters, and Science), "BA" (Business Administration), or "EDU" (Education), and can be split into three separate variables:

Major_ALS: Yes/No
Major_BA: Yes/No
Major_EDU: Yes/No

Assume this university had three colleges, and each college provided one major degree program, with several concentrations under each major. In this case, only two of the variables will be needed, for if the values of two are known, then the third is also known. For instance, if a student is neither an ALS major nor a BA major, then that student must be an EDU major. Another alternative is to convert the values of a categorical variable into a series of scores. For example, the three values of ***Major*** can be assigned ordinal or nominal values as follows:

ALS: 1; BA: 2; EDU: 3

In this study, we adopt the second alternative. Students were classified into thirteen different groups and the coding system is as follows:

- ***Sex***: sex of students; a binary variable
  F: 1; M: 2
- ***County***: the service counties students coming from; a categorical variable
  Calaveras: 1; Foreign Countries: 2; Mariposa: 3; Merced: 4; other states: 5; Others: 6; San Joaquin: 7; Stanislaus: 8; Tuolumne: 9
- ***Major***: one of the three university colleges (College of Arts, Letters, and Sciences, College of Business Administration, and College of Education) students belong to; a categorical variable
  ALS: 1; BA: 2; EDU: 3
- ***1st_G***: is the student a first-generation college student in the family; a binary variable
  N: 1; Y: 2
- ***Transfer_In***: is the student a transferred student; a binary variable
  N: 1; Y: 2
- ***Original_College***: the original institution a student came from, such as regional community colleges or other four-year colleges or universities; a categorical variable
  CA independent College or University: 1; CSUS: 2; Foreign: 3; JCs: 4; Other CSUs: 5, Out of State Institutions: 6; UCs: 7
- ***Ethnicity***: the ethnicity of a student; a categorical variable
  American Indian: 1; Asian/Pacific: 2; Black: 3; Hispanic: 4; Intl: 5; Other: 6; White: 7
- ***Classification***: the class status of a student, including freshman, sophomore, junior, senior, and post-baccalaureate; a categorical variable
  Freshman: 1; Junior: 2; Post Baccalaureate: 3; Senior: 4; Sophomore: 5
- ***Status***: full time or part time students; a binary variable
  Full time: 1; Part time: 2
- ***Age <= 24***: if a student is under age 24; a binary variable

N: 1; Y: 2

- ***Married***: the marital status of a student; a binary variable
  N: 1; Y: 2
- ***Financial_aide***: if a student needs a financial aid; a binary variable
  N: 1; Y: 2
- ***Transfer_or_Dropout***: transferring to other four-year colleges or universities, completing the entire undergraduate education at CSUS, or dropping out of school; a categorical variable
  Dropout: 1; Stay: 2; to other institutions: 3

Though more detailed categories could be designed[27], consideration of the number of samples in each category made more detailed categorization inappropriate for a small-medium sized university. This current research focused on the analysis of undergraduate students only due to the time consideration; however, this research can be extended to include the analysis for graduate students in future studies.

After data collection, we need to ensure the quality of the dataset by removing, i.e., "scrubbing"[28], erroneous pieces of data from the dataset, such as inaccurate inputting, incomplete information, improperly formatted structures, and duplication of data. Also, unnecessary data contained in the student records, such as a student's name, identification number, street address, and phone number, are removed.

### *4.2. Data Mining Phase*

Due to the consideration of the interpretability issue of machine learning techniques, we conducted the rule-based affinity analysis to extract interesting associations and correlation relationships between a student's decision on staying with us until graduation and the attributes of the student. The discovered relationships among attributes are represented as an *antecedent* ➔ *consequent* type of rule, and can help the college compare and understand the behavioral patterns of students in different groups. Through this analysis, the college can leverage the analysis result in designing unique outreach and retention programs or activities for different segments of students.

Association rules show attribute value conditions that occur frequently together in a given dataset, and provide information of this type in the form of "IF-THEN" statements. In association analysis the "IF" part is called the antecedent (***A***), and the "THEN" part is called the consequent (***C***). Both are sets of items that are disjoint (i.e., do not have any item in common). These association rules are computed from the data, and probabilistic in nature.[25]

In addition to the antecedent and the consequent, an association rule has two numbers that express the degree of uncertainty about the rule. The first number is called the support for the association rule. The support is simply the number of records that include all items in the antecedent (***A***) and consequent (***C***) parts of the rule. The support is sometimes expressed as a percentage of the total number of records in the dataset. This is equivalent to an estimated probability that a record selected randomly from the entire dataset will contain all items in the antecedent (***A***) and consequent (***C***):

$$\text{Support} = P(\boldsymbol{A} \text{ AND } \boldsymbol{C}).$$

The other number is called the confidence of the rule for measuring the strength of association between ***A*** and ***C***, i.e., the degree of uncertainty about the rule. Confidence is the ratio of the number of records that include all items in ***C*** and ***A*** (i.e., the support) to the number of records that include all items in ***A***. In other words, the confidence is an estimated conditional probability that a randomly selected record will include all the items in ***C*** given that the record includes all the items in ***A***. Confidence can be defined as follows:

$$\text{Confidence} = \frac{\text{Number of records containing all items in } \boldsymbol{A} \text{ AND } \boldsymbol{C}}{\text{Number of records containing all items in } \boldsymbol{A}}$$

$$= \frac{P(\boldsymbol{A} \text{ AND } \boldsymbol{C})}{P(\boldsymbol{A})} = P(\boldsymbol{C}|\boldsymbol{A})$$

Another important parameter in association analysis is the lift ratio. Lift is the ratio of Confidence to Benchmark Confidence as

$$\text{Lift Ratio} = \frac{\text{Confidence}}{\text{Benchmark Confidence}}$$

with the assumption that ***A*** and ***C*** are independent. Under independence, the support is computed as:

$$P(\boldsymbol{A} \text{ AND } \boldsymbol{C}) = P(\boldsymbol{A}) \times P(\boldsymbol{C}),$$

and the benchmark confidence is defined as:

$$\frac{P(\boldsymbol{A}) \times P(\boldsymbol{C})}{P(\boldsymbol{A})} = P(\boldsymbol{C}) = \frac{\text{Number of records containing all items in } \boldsymbol{C}}{\text{Total number of records in the entire dataset}}$$

For example, if a supermarket database has 100,000 transaction records, out of which 2,000 include all items included in ***A***, and 800 of these also include all items in ***C***. Then, the association rule "If all the items in ***A*** are purchased, then each item in ***C*** is also purchased on the same shopping trip" has a support of 800 transaction records (alternatively, 0.8% = 800/100,000) and a confidence of 40% (= 800/2,000). Assume the total number of transaction records which include all items in ***C*** is 5,000, then the Benchmark Confidence is 5% = 5,000/100,000. Thus, the lift ratio is 40%/5% = 8. Hence, lift is a value that gives us information about the increase in probability of ***C*** given ***A***. A lift ratio greater than 1.0 suggests that the level of association between ***A*** and ***C*** is higher than would be expected if they were independent. The larger the lift ratio, the greater the strength of the association.

*4.2.1. Conducting Affinity Analysis for DM Modeling*

We then conducted association analysis for our student dataset by using XLMiner®. In this analysis, we set the following criteria for any interesting association rule to be displayed: the minimum support equal to 200, and confidence as 60%. This means that for each association rule, among the 1,000 students in our data file there is a group of students fit the attributes in the antecedent and at least 200 of them also fit the attributes in the consequent. Also, for the group of records containing attributes in the antecedent of the rule, at least 60% of them also contain the attributes in the consequent of the rule. The association rules generated are shown in Table 1. The output includes information on Support(***A***)—the support of the antecedent, Support(***C***)—the support of the consequent, and the support of the combined set—Support(***A*** ∪ ***C***). The output also includes the confidence of each rule and the lift ratio.

In interpreting results, it is useful to look at the various measures. The support for each rule indicates how many transactions (or the proportion of transactions) are represented by this rule. If only a small number of transactions are represented, then this rule may be not that useful. The lift ratio indicates how efficient the rule is in finding consequents, compared to random selection. Though a very efficient rule is desirable, a very efficient rule with low support is not desirable as a less efficient rule with strong support. The confidence shows the rate at which consequents will be found among the transactions involving the antecedent. A rule with low confidence may find consequents at too low a rate to be worth the cost of promoting the consequent in all the transactions involving the antecedent. However, when a rule has high confidence, we also need to look into Support(***A***) and Support(***A*** ∪ ***C***). If Support(***A***) is already low, then even though the rule has high confidence, the rule is still not valuable to us.

*4.2.2. Model Refinement*

The original output included 424 rules. In reviewing these rules, we found out some of the rules involved the same set of attributes, with different antecedents and consequents. Those rules could be combined together. Since we were interested in learning about the characteristics of different groups of students who chose to stay with us, or chose to transfer to other institutions, or to drop out of schools, we deleted those association rules did not contain the attribute Transfer or Dropout. After refinement, the reduced rule set (or the knowledge base) consisted of 137 rules. Due to length consideration, we showed only the first two rules in Table 1.

**Table 1.** Association Rules Output for Student Data File.

| Rule # | Conf. % | Antecedent (*A*) | Conseq (*C*) | Support (*A*) | Support (*C*) | Support (*A* ∪ *C*) | Lift Ratio |
|---|---|---|---|---|---|---|---|
| 1 | 84.19 | 1st G College_N, Transfer_Y=> | Original College_JCs, Transfer or Dropout_Stay | 253 | 347 | 213 | 2.42 |
| 2 | 97.26 | 1st G College_N, Original College_JCs=> | Transfer or Dropout_Stay, Transfer_Y | 219 | 406 | 213 | 2.40 |

In Table 1, the first rule, for example, has ***A*** = {1st G College_N, Transfer_Y} and ***C*** = {Original College_JCs, Transfer or Dropout_Stay}. Number of rules containing ***A*** is 253, number of rules containing ***C*** is 347, and number of rules containing both ***A*** and ***C*** is 213. Thus, we can compute

$$\text{Confidence} = \frac{P(\boldsymbol{A} \text{ AND } \boldsymbol{C})}{P(\boldsymbol{A})} = \frac{213/1000}{253/1000} = 0.8419.$$

Benchmark Confidence = 347/1000. Thus, the Lift Ratio = Confidence / Benchmark Confidence = 0.8419 / 0.347 = 2.4262.

However, Table 1 still contains 137 rules, and many of them are trivial. For instance, a rule: IF Original College = "JCs" and Transfer or Dropout = "Stay" THEN Transfer Student = "Y", is trivial, for if a student's original college is junior colleges, then of course that student is a transfer student. Or, if a student's original college is CSUS, then of course that student is not a transfer student. Or, if a student is not a transfer student, then of course that student's original college must be CSUS.

To refine the model further, some rules can be combined together. For example, the following two rules can be merged together:

IF 1st Generation College Student = "N" and Sex = "F" and Major = "ALS" THEN Transfer or Dropout = "Stay", and

IF 1st Generation College Student = "N" and Sex = "F" THEN Transfer or Dropout = "Stay"

Since the first one is just a subset of the second one, the first rule can be eliminated. Still, some rules are spurious. For example, the existence of the following two rules implies the status attribute is irrelevant:

IF Status = "Full-time" THEN Transfer or Dropout = "Stay", and

IF Status = "Part-time" THEN Transfer or Dropout = "Stay"

We thus further eliminate the trivial rules and combine rules into a smaller set of rules. In this research we are interested in investigating the factors which are related to student retention. In other words, we are interested in learning about those factors which are important to a student's decision to stay with our university until graduation. Thus, we retain only those rules in Table 1 with the consequent including: Transfer_or_Dropout_Stay (i.e., staying with our university until graduation). The result consisting of 27 rules is shown in Table 2 at the end of this paper, due to space and readability considerations.

Based on Table 2, we derived a set of IF-THEN rules as follows:

**Rule 1** For those junior college transfers, about 95% of them stayed until graduation.

**Rule 2** For female transfers, about 95% of them stayed until graduation.

**Rule 3** For the transfer students, if they stayed until senior year, then almost all of them would stay until graduation.

**Rule 4** For those transfer, non-first-generation college students, almost all of them would stay until graduation.

**Rule 5** For senior students, 96% of them would stay until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

**Rule 7** Among those transfer students with financial needs, about 95% of them would stay until graduation.

**Rule 8** Among transfer full-time students, almost 95% of them would stay until graduation.

**Rule 9** Among transfer unmarried students, almost 95% of them would stay until graduation.

**Rule 10** Among transfer ALS students, about 94.5% of them would stay until graduation.

**Rule 11** Among transfer full-time ALS majors, about 94% of them would stay until graduation.

**Rule 12** Among transfer ALS majors with financial needs, about 94% of them would stay until graduation.

**Rule 13** For the ALS majors with financial needs, only 78% of them would stay until graduation.

**Rule 14** Among transfer ALS female students, about 93.5% of them would stay until graduation.

**Rule 15** Among junior students, about 87% of them would stay until graduation.

**Rule 16** Among married ALS students, about 80.5% of them would stay until graduation.

**Rule 17** Among those students older than 24 and with financial needs, about 79% of them would stay until graduation.

**Rule 18** For those ALS students older than 24, about 79% of them would stay until graduation.

**Rule 19** For those full-time students older than 24, about 78% of them would stay until graduation.

**Rule 20** For those transfer students younger than 24, about 93.5% of them would stay until graduation.

**Rule 21** For students older than 24, about 76% of them would stay until graduation.

**Rule 22** For those married full-time students, about 78% of them would stay until graduation.

**Rule 23** For those married students with financial needs, about 78% of them would stay until graduation.

**Rule 24** For the married students, about 77% of them would stay until graduation.

**Rule 25** For those full-time students with financial needs, about 77% of them would stay until graduation.

**Rule 26** For students with financial needs, about 76% of them would stay until graduation.

**Rule 27** For male students, about 76% of them would stay until graduation.

### *4.3. Post Data Mining Phase--Model Interpretation & Strategic Implications*

The post data mining phase mainly consists of model interpretation and deployment of strategic importance. From the above set of rules, we can derive from each rule or from the combination of multiple rules further implications regarding the factors important to a student's decision on staying

with our university until graduation or transferring to other institutions or simply dropping out of school. These implications can also be adapted by the student retention and outreach office as guidelines for developing student retention and outreach strategies. There are 19 derived implications as follows: (Note that each percentage represents the Confidence level.)

**Rule 1** For those junior college transfers, about 95% of them stayed until graduation.

***Implication 1****: JC transfers are more likely to stay until graduation than those transfer from other types of institutions.*

**Rule 2** For female transfers, about 95% of them stayed until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

***Implication 2****: From the above two rules, it seems that a student's transfer status is more important than a student's sex in affecting a student's decision to stay with our university.*

**Rule 4** For those transfer, non-first-generation college students, almost all of them would stay until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

***Implication 3****: From the above two rules, we see there is a reinforcing interaction between a student's transfer status and the status of non-first-generation college student in affecting a student's decision to stay with us until graduation. Among the transfer students, students of non-first-generation college students tend to be more likely to stay with us than those as first-generation college students.*

**Rule 5** For senior students, 96% of them would stay until graduation.

**Rule 15** Among junior students, about 87% of them would stay until graduation.

***Implication 4****: From the above two rules, we understand that senior students are more likely to stay until graduation than students of other standings.*

**Rule 3** For the transfer students, if they stayed until senior year, then almost all of them would stay until graduation.

**Rule 5** For senior students, 96% of them would stay until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

***Implication 5****: From the above three rules, we see there is a reinforcing interaction between a student's transfer status and standing as a senior student in affecting a student's decision to stay with us until graduation.*

**Rule 6** For transfer students, 95% of them would stay until graduation.

**Rule 8** Among transfer full-time students, almost 95% of them would stay until graduation.

***Implication 6****: From the above two rules, it seems that a student's transfer status is more important than a student's full-time status in affecting a student's decision to stay with our university. These two rules also seem to imply that among transfer students, full-time students seem to be more likely to stay with us until graduation than part-time students.*

**Rule 7** Among those transfer students with financial needs, about 95% of them would stay until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

***Implication 7****: From the above two rules, it seems that a student's transfer status is more important than a student's financial needs in affecting a student's decision to stay with our university. These two rules also seem to imply that among transfer students, students with financial needs seem to be more likely to stay with us until graduation than those without financial needs.*

**Rule 9** Among transfer unmarried students, almost 95% of them would stay until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

***Implication 8****: From the above two rules, it seems that a student's transfer status is more important than a student's marital status in affecting a student's decision to stay with our university. These two rules also seem to imply that among transfer students, unmarried students are more likely to stay with us until graduation than married students.*

**Rule 11** Among transfer full-time ALS majors, about 94% of them would stay until graduation.

**Rule 8** Among transfer full-time students, almost 95% of them would stay until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

***Implication 9****: From the first two rules, it seems that a student's transfer status and full-time status together is more important than a student's major in affecting a student's decision to stay until graduation. From the third rule, it seems that a student's transfer status is more important than a student's full-time status in such a decision. However, among transfer full-time students, ALS majors seem to be more likely to stay until graduation than other majors.*

**Rule 12** Among transfer ALS majors with financial needs, about 94% of them would stay until graduation.

**Rule 13** For the ALS majors with financial needs, only 78% of them would stay until graduation.

***Implication 10****: From the above two rules, a student's transfer status is more important than a student's major or financial need. These two rules also seem to imply that among the ALS majors with financial needs, the transfer students are more likely stay until graduation than non-transfer students.*

**Rule 10** Among transfer ALS students, about 94.5% of them would stay until graduation.

**Rule 14** Among transfer ALS female students, about 93.5% of them would stay until graduation.

***Implication 11****: From the above two rules, it seems that a student's transfer status and major are more important than Sex in affecting a student's decision to stay with our university. These two rules also seem to imply that among the ALS transfers, female students tend to be more loyal to our university than male students.*

**Rule 16** Among married ALS students, about 80.5% of them would stay until graduation.

**Rule 24** For the married students, about 77% of them would stay until graduation.

***Implication 12****: From the above two rules, it seems that a student's marital status is more important than a student's major in affecting a student's decision to stay with our university. These two rules also imply that among married students, ALS majors are more likely to stay with us until graduation than other majors.*

**Rule 17** Among those students older than 24 and with financial needs, about 79% of them would stay until graduation.

**Rule 21** For students older than 24, about 76% of them would stay until graduation.

***Implication 13****: From the above two rules, it seems that a student's age is more important than the financial need in affecting a student's decision to stay with our university. These two rules also imply that among students older than 24, those with financial needs are more likely to stay with us until graduation than those without financial needs.*

**Rule 18** For those ALS students older than 24, about 79% of them would stay until graduation.

**Rule 21** For students older than 24, about 76% of them would stay until graduation.

***Implication 14****: From the above two rules, it seems that a student's age is more important than a student's major in affecting a student's decision to stay with our university. These two rules also imply that among students older than 24, ALS majors are more likely to stay with us until graduation than other majors.*

**Rule 20** For those transfer students younger than 24, about 93.5% of them would stay until graduation.

**Rule 6** For transfer students, 95% of them would stay until graduation.

***Implication 15****: From the above two rules, it seems that a student's transfer status is more important than a student's age in affecting a student's decision to stay until graduation. These two rules also imply that among the transfer students, students younger than 24 are more likely to stay with us until graduation than students older than 24.*

**Rule 19** For those full-time students older than 24, about 78% of them would stay until graduation.

**Rule 21** For students older than 24, about 76% of them would stay until graduation.

***Implication 16****: From the above two rules, it seems that a student's full-time status is not as important as the age in determining a student's decision to stay with our university until graduation. These two rules also imply that among students older than 24, full-time students are more likely to stay with us until graduation than part-time students.*

**Rule 22** For those married full-time students, about 78% of them would stay until graduation.

**Rule 24** For the married students, about 77% of them would stay until graduation.

***Implication 17****: From the above two rules, it seems that the full-time status of a student does not affect a student's decision to stay with our university as much as the marital status. These two rules also imply that among married students, full-time students are more likely to stay with us until graduation than part-time students.*

**Rule 24** For the married students, about 77% of them would stay until graduation.

**Rule 23** For those married students with financial needs, about 78% of them would stay until graduation.

***Implication 18****: From the above two rules, it seems that a student's financial needs would not affect a student's decision to stay with our university as much as the marital status. These two rules also imply that among married students, those with financial needs are more likely to stay with us until graduation than those without financial needs.*

**Rule 25** For those full-time students with financial needs, about 77% of them would stay until graduation.

**Rule 26** For students with financial needs, about 76% of them would stay until graduation.

***Implication 19****: From the above two rules, it seems that the full-time status of a student does not affect a student's decision to stay with our university as much as the financial needs. These two rules also imply that among students with financial needs, full-time students are more likely to stay with us until graduation than part-time students.*

## 5. Conclusions & Future Research

In this paper, we followed a three-phase-six-stage ADMDC in applying data mining techniques to a student data file of 1,000 records based on the basic information obtained from Campus Data Portfolio. The data file consists of thirteen attributes. The first twelve attributes are student-related attributes, and the last one is the decision made by a student to stay with the college until graduation. We applied the Association Rules, including the affinity analysis, to identify the relationships between student-associated attributes and the student decision on staying until graduation, transferring to other institutions, or dropping out of school.

Tinto[29] argued that college institutions failed to translate what they had learned on student retention into a set of guidelines for actions and policies to increase rates of college completion. This has been evidenced by the increased accessibility to college education over the past several decades, especially for students of low-income and underserved backgrounds, without seeing similar increases in college completion. For generating a set of guidelines, we conducted the affinity analysis by using the association rule technique. We set the support level to be at least 200, and the confidence level to be at least 60%. The original model consisted of a rule set, i.e., knowledge base, of more than 400 rules. In the stage of model refinement, we eliminated trivial rules and redundant rules, and combined subsumed rules with their containing rules. Since our study was about student retention, we also eliminated those rules from the developing model whose consequents did not contain the outcome variable. The refined model or knowledge base was composed of less than 30 rules.

From the refined model, we derived a set of interesting and useful implications regarding the important factors affecting a student's decision to stay with us until graduation were discussed in the previous section. Important factors associated with a student's decision included: Transfer, Age, Marital Status, Financial Needs, Major, Full-time Status, Sex, First-Generation college student or not, and the standing classification. Among the important factors, we found out the following interesting relationship: whether the student was a transfer or not was more important than the age and the marital status which in turn were more important than the financial needs and the student's major, and which in turn were more important than a student's status as a full-time or part-time student. Different from our findings, Delen[15] proposed that the financial factor was one of the two most important factors. It is worthwhile to investigate what might have caused such a difference.

Since CSUS has long been serving the under-represented students, we plan to include this factor as a predictor for future studies. Ott, Markewich, & Ochsner[30] developed a logit model to predict the retention of graduate students. In their study, predicted retention rates for graduate students were independent of age and sex, but were a function of ethnicity, registration status, and the interactions between academic division and registration status and between academic division and ethnicity. Still, according to Bilquise et al.[7], ensemble predictors outperformed traditional classification techniques in predicting student retention. We also plan to investigate the effect of interactions among predictors on student retention, and compare it with that of ensemble predictors.

Since our analysis was based on a sample of 1,000 students, the findings here have their applicability limitation. With the availability of a broader and more recent student data file, we might get more and deeper insights about our analysis. Our knowledge base will be refined continually. Through this research, we have demonstrated the usefulness and application of data mining techniques to the discovery of useful and interesting relationships among data from a huge set of data. Still, COVID-19 has changed, not only the teaching modality, but also administrative and strategic processes, including student retention activities. The effects of such a paradigm shifting will be a worthy topic for future research when more data become available.

The result from our study will have both short-term and long-term strategic implications. We expect the direct and immediate effects of this study are that it can help a college to develop better understanding on the factors affecting a student's continuance, transfer or drop-out decisions; maintain or increase its students' loyalty to the college; tailor a college's various student retention and outreach programs and activities more

effectively to the characteristics and needs of both the potential and current students; and modify a college's development strategies. From the long-term point of view, this study will help an educational institution identify its own competition niche, and thus enable the institution to reposition itself in this highly competitive global educational market.

## References


[1] Crosling, G. (2017). Student retention in higher educator: A shared issue. In *Encyclopedia of International Higher Education Systems and Institutions*. Shin, J.C. & Teixeira, P. (eds.) 1-6.

[2] Soussa, T. (2015). Student retention is more important than ever. *Higher Ed Live*, September 9.

[3] Moxley, D., Najor-Durack, A., & Dumbrigue, C. (2001). *Keeping students in Higher Education*. London: Routledge.

[4] Wignall, A. (2019). Are freshmen retention rates a good indicator of a college 's quality? *College Raptor*, January 10.

[5] Asil, O. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems*, *116*(8).

[6] Ang, L. & Buttle, F. (2006). Customer retention management processes: A quantitative study. *European Journal of Marketing*, *40*(½), 83-99. https://doi.org/10.1108/03090560610637329

[7] Bilquise, G., Abdallah, S., & Kobbaey, T. (2020). Predicting student retention among a homogeneous population using data mining. In: Hassanien, A., Shaalan, K., & Tolba, M. (eds.) *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019*, 35-46.

[8] Roberts, J. (2018). Professional staff contributions to student retention and success in higher education. *Journal of Higher Education Policy and Management*, *40*(2), 140-153.

[9] Hand, D., Mannila, H., & Smyth, P. (2001). *Data Mining*. Cambridge, MA: MIT Press.

[10] Berry, M.J.A. and Linoff, G.S. (2000). *Mastering Data Mining*. New York: Wiley.

[11] Kleinberg, J., Papadimitriou, C., Raghavan, P. (1998). A microeconomic view of data mining. *Data Mining & Knowledge Discovery*, *2*(4), 311-324.

[12] Hastie, T., Tibshirani, R., Friedman, J. (2013). *The Elements of Statistical Learning*, 2nd ed. Springer.

[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

[14] Dey, E.L., Astin, A.W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, *34*(5), 569–581. https://doi.org/10.1007/BF00991920

[15] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

[16] Villano, R., Harrison, S., Lynch, G., Chen, G. (2018). Linking early alert systems and student retention: a survival analysis approach. *Higher Education*, *76*(5), 903–920. https://doi.org/10.1007/s10734-018-0249-y

[17] Grayson, J.P. (1998). Racial origin and student retention in a Canadian University. *Higher Education*, *36*(3), 323–352. https://doi.org/10.1023/A:1003229631240

[18] Jia, J.W. & Mareboyana, M. (2015). Undergraduate Student Retention Prediction Using Wavelet Decomposition. In: Yang, G.C., Ao, S.I., Gelman, L. (eds) *Transactions on Engineering Technologies*. Springer, Dordrecht, 643-655.

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

[20] Wetzel, J.N., O'Toole, D., & Peterson, S. (1999). Factors affecting student retention probabilities: A case study. *Journal of Economics and Finance*, *23*(1), 45–55. https://doi.org/10.1007/BF02752686

[21] Jung, R., Kochbeck, J., Nagel, A. (2008). Student retention through customized service processes. In: Oya, M., Uda, R., Yasunobu, C. (eds) *Towards Sustainable Society on Ubiquitous Networks*. IFIP – The International Federation for Information Processing, 286. Springer, Boston, MA.

[22] Rahman, S.I. (2014). Spelman College: A case study of student retention strategies. In: Gasman, M. & Commodore, F. (eds.) *Opportunities and Challenges at Historically Black Colleges and Universities*. Palgrave Macmillan, New York.

[23] Deng, P.S. (2008). Applying a market-based approach to the development of a sharing-enabled KM model for knowledge-intensive small firms. *Information Systems Management*, *25*(2), 174-187.
[24] Du, M., Liu, N., & Hu, X. (2020). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68-77.
[25] Shmueli, G., Bruce, P.C., Patel, N.R. (2016). *Data Mining for Business Analytics*. New York: Wiley.
[26] Berry, M.J.A. and Linoff, G.S. (1997). *Data Mining Techniques*. New York: Wiley.
[27] Heagney, M. (2008). Student success and student diversity. In *Improving student retention in higher education: The Role of teaching and learning*. In: Crosling, G., Thomas, L., & Heagney, M. (eds.) London: Routledge.
[28] Mercurius, N. (2005). Scrubbing data for D3M. *T.H.E. Journal*, Oct., 15-18
[29] Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention. In: Smart, J. (ed) *Higher Education: Handbook of Theory and Research*. Volume 25. Springer, Dordrecht.
[30] Ott, M.D., Markewich, T.S. & Ochsner, N.L. (1984). Logit analysis of graduate student retention. *Research in Higher Education*, *21*(4), 439–460.

**Table 2.** Further Reduced Association Rules Output for Student Data File.

| Rule # | Conf. % | Antecedent (*A*) | Consequent (*C*) | Support(*A*) | Support(*C*) | Support(*A* ∪ *C*) | Lift Ratio |
|---|---|---|---|---|---|---|---|
| 1 | 95.33 | Original College_JCs=> | Transfer or Dropout_Stay | 364 | 406 | 347 | 2.348021 |
| 2 | 94.66 | Sex_F, Transfer_Y=> | Transfer or Dropout_Stay | 281 | 718 | 266 | 1.318411 |
| 3 | 97.75 | Classification_S, Transfer_Y=> | Transfer or Dropout_Stay | 222 | 718 | 217 | 1.361389 |
| 4 | 97.23 | 1st G College_N, Transfer_Y=> | Transfer or Dropout_Stay | 253 | 718 | 246 | 1.354223 |
| 5 | 95.92 | Classification_S=> | Transfer or Dropout_Stay | 245 | 718 | 235 | 1.33591 |
| 6 | 95.31 | Transfer_Y=> | Transfer or Dropout_Stay | 426 | 718 | 406 | 1.32737 |
| 7 | 95.16 | Financial Aide_Y, Transfer_Y=> | Transfer or Dropout_Stay | 289 | 718 | 275 | 1.325288 |
| 8 | 94.98 | Status_F, Transfer_Y=> | Transfer or Dropout_Stay | 279 | 718 | 265 | 1.32287 |
| 9 | 94.85 | Married_N, Transfer_Y=> | Transfer or Dropout_Stay | 233 | 718 | 221 | 1.321028 |
| 10 | 94.51 | Major_ALS, Transfer_Y=> | Transfer or Dropout_Stay | 328 | 718 | 310 | 1.316326 |
| 11 | 93.86 | Major_ALS, Status_F, Transfer_Y=> | Transfer or Dropout_Stay | 228 | 718 | 214 | 1.307237 |
| 12 | 93.78 | Financial Aide_Y, Major_ALS, Transfer_Y=> | Transfer or Dropout_Stay | 225 | 718 | 211 | 1.306097 |
| 13 | 77.69 | Financial Aide_Y, Major_ALS=> | Transfer or Dropout_Stay | 363 | 718 | 282 | 1.081977 |
| 14 | 93.53 | Sex_F, Major_ALS, Transfer_Y=> | Transfer or Dropout_Stay | 232 | 718 | 217 | 1.302709 |
| 15 | 86.89 | Classification_J=> | Transfer or Dropout_Stay | 267 | 718 | 232 | 1.210186 |
| 16 | 80.56 | Major_ALS, Married_Y=> | Transfer or Dropout_Stay | 252 | 718 | 203 | 1.121944 |
| 17 | 79.46 | Age <= 24_N, Financial Aide_Y=> | Transfer or Dropout_Stay | 297 | 718 | 236 | 1.106703 |
| 18 | 79.45 | Age <= 24_N, Major_ALS=> | Transfer or Dropout_Stay | 253 | 718 | 201 | 1.106499 |
| 19 | 78.41 | Age <= 24_N, Status_F=> | Transfer or Dropout_Stay | 301 | 718 | 236 | 1.091996 |
| 20 | 93.52 | Age <= 24_Y, Transfer_Y=> | Transfer or Dropout_Stay | 247 | 718 | 231 | 1.302539 |
| 21 | 76.58 | Age <= 24_N=> | Transfer or Dropout_Stay | 444 | 718 | 340 | 1.066526 |
| 22 | 78.29 | Married_Y, Status_F=> | Transfer or Dropout_Stay | 281 | 718 | 220 | 1.090415 |
| 23 | 77.96 | Financial Aide_Y, Married_Y=> | Transfer or Dropout_Stay | 363 | 718 | 283 | 1.085814 |
| 24 | 77.03 | Married_Y=> | Transfer or Dropout_Stay | 444 | 718 | 342 | 1.0728 |
| 25 | 76.88 | Financial Aide_Y, Status_F=> | Transfer or Dropout_Stay | 359 | 718 | 276 | 1.070755 |
| 26 | 76.41 | Financial Aide_Y=> | Transfer or Dropout_Stay | 568 | 718 | 434 | 1.064185 |
| 27 | 76.43 | Sex_M=> | Transfer or Dropout_Stay | 314 | 718 | 240 | 1.064528 |