An Illustration of Using Adaptive Data Mining to Develop Strategic Knowledge Bases for Student Retention

Pi-Sheng Deng and Abhijit Chaudhury

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 with students to effectively provide additional value and service. In this study we developed a two-module system: a decision tree for predicting a student's decision to stay until graduation and an affinity analysis algorithm for identifying the relationship between student attributes and student decisions. We followed a three-phase-six-stage adaptive data mining cycle in developing a knowledge base for student retention strategies. The affinity analysis initially identified more than 400 association relationships with student retention. By applying inductive inference, the association rule set was refined iteratively down to less than 30 rules, and useful strategic implications were developed regarding how the selected factors were associated with a student's decision. This set of implications and factors was then integrated into the development of strategies for student retention.**

Index Terms—**Affinity analysis; Association rules; Adaptive DM Development Cycle; Decision trees; Inductive inference; Student retention**

I. Introduction

The term data mining refers to the use of algorithms to extract patterns of information embedded in large databases, can refer to the techniques or the entire systemic development process, and can also refer to the discipline of study. Data mining techniques have been applied to different areas of applications and for different types of tasks, including classification, regression, prediction, clustering, and association among others [1]. Recently, educational researchers have applied data mining techniques to educational databases for the development of knowledge bases for supporting administrative decision making at colleges and universities.

With the outbreak of the COVID-19 pandemic and the growth of Internet/Web technology, educational institutions worldwide have also experienced an unprecedented impact caused by the increasing amount of various kinds of information available to students on the Web. Students can now compare almost every aspect of different institutions in making their educational decisions. Such technological development has allowed students to become more selective consumers in the educational marketplace. As one consequence, the concept of geography-based service regions of a college 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 [2][3]. Global competition and the rise of digital technology have forced educational institutions to think strategically about their processes for managing their relationships with students and other stakeholders. To be competitive, educational institutions must provide more value and service to stakeholders and must improve institutional processes and programs for interacting with their stakeholders, especially students [2].

Students are long-term assets of an educational institution, and the relationship with them should be nurtured through 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 [4]," and has been a significant measure of the effectiveness of institutional processes and programs [2][5]. Student retention focuses on managing 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 [6][7]. It is also one of the most challenging issues for higher education nationwide, and even worldwide [3][8][9].

This research is motivated by the resource-intensive efforts devoted to student outreach and retention, yet with marginal return. We employed data mining techniques to analyze student demographic data and student profiles for discovering hidden trends or patterns of the antecedent ➔ consequent relationships among 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 processes so that the institution can enhance its relationships with its stakeholders, especially students, and allocate student retention resources and efforts more effectively.

Manuscript received November 29, 2022; revised June 10, 2023

P.S. Deng is a full professor of the MIS Department, College of Business Administration, California State University at Stanislaus, Turlock, CA 95382 USA (corresponding author; phone: +1 209 667-3018; e-mail: pdeng@csustan.edu)

A. Chaudhury is a full professor of the Department of Information System and Analytics at Bryant University, Smithfield, RI 02917 USA (e-mail: achaudhu@bryant.edu)

II. 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 [10][11]][12]. Data mining draws heavily on statistical techniques, especially linear regression, logistic regression, discriminant analysis, and principal components analysis [13]. In addition, data mining also includes techniques from artificial intelligence (AI), such as decision trees, production rules, neural networks, fuzzy logic, and genetic algorithms [14].

Rather than simply assuming that one technique is analytically superior to others, studies were conducted on how logistic regression, probit analysis, and linear regression compared in predicting college student retention [15]. Results indicate that though the former two offered theoretical advantages, they showed little practical advantages over traditional linear regression. Also, a data mining model was proposed in [16] to predict at-risk students and to explain the reasons behind student attrition so that colleges can intervene to retain them. This study showed that educational and financial variables were among the most important predictors. Further, studies in [17] employed survival analysis using demographic, institution, student GPA and workload variables was employed to identify students at high risk of dropping out [17]. Logistic regression analysis has been often applied to investigate the relationship between first-year student retention and factors, such as full-time status, ethnicity, and GPA, and found no significant relationship between the retention rate and ethnicity [18]. Recently, decision trees, support vector machines (SVM), and neural networks have been applied to investigate the main factors that influence undergraduate student retention in 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 [19].

Another study identified transfer status, residency, and ethnicity as crucial factors to retention [20]. The studies presented in [21] indicated that academic and social integration factors were the most significant factors in persistence. while financial considerations were of less importance in decisions.

In addition to identifying the factors related to a student's decision to drop out, studies in [22] applied marketing concepts help institutions of higher education align educational and service processes more closely to students needs for alleviating student retention issues. According to [23] 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 studies used 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 machine learning techniques, including data mining, fall short of providing easy-to-understand and transparent results to end users [24]. In this study we applied the decision tree, which is a non-parametric and easy to interpret technique, and the rule-based affinity analysis, which has been relatively less employed for student retention strategy development. We iteratively refined the 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 develop strategies for student retention.

It is not our intention to compare the prediction performance of different data mining techniques. Our intention is to show the use of affinity analysis for developing a knowledge base system for student retention. The rest of this manuscript is organized as follows: in the following section we describe an adaptive data mining development cycle. We, then, describe how we collected data from a medium sized college in California to illustrate the execution of the development cycle. We especially emphasize the importance of the adaptive refinement process to knowledge bases generated by affinity analysis. We, then, discussed the strategic implications derived from the current models. Finally, a conclusion section discusses the important findings of this study and some possible future development directions.

III. Description of Adaptive Data Mining Development Cycle

In this section, we follow a three-phase six-stage adaptive DM development cycle in [25] for developing knowledge bases for supporting student retention. The first phase, the "pre-data mining" 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, 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.

Next is the data mining phase, characterized by two continually interacting processes that result in a refined knowledge base [26]. 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 then *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 its content will continually be refined with new insights gained through the interactive modeling-refining process.

The last phase is the post-data mining phase which consists of model interpretation/implication and deployment. One of the major limitations of today's machine learning techniques is the lack of transparency behind their behaviors [24]. Thus, if the model developed 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.

Though the data mining development cycle we followed may look very similar to the Systems Development Life Cycle (SDLC) used in structured systems analysis and design, there are some important distinctions. SDLC assumes that end users have clear understanding of their own information needs, and that they are able to articulate their own information needs. These two assumptions encourage systems analysts to invest much of their time and effort in the early stages of SDLC for extracting as much end user information needs as possible. The spiral development process of SDLC is more of a (mental) model-driven iterative approach than an adaptive process. The nature of SDLC approach is also a stage-by-stage iterative development process. From any stage in the development process, we can return to any certain earlier stage for correcting or revising glitches or mistakes.

On the other hand, our data mining development cycle is a data-driven approach to system development. This approach requires the availability of a huge amount of data and the availability of data mining algorithms. Data mining algorithms render knowledge extracted from the data accumulated over the years. Though it is also capable of iterative development, it is specifically characterized by the knowledge base refining process which is an adaptive man-machine collaborative process. This continual adaptive tuning process results in a refined knowledge base.

IV. An Illustration of Applying Affinity Analysis to Data Mining Development Cycle

The case used in this section is a small-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 development cycle. The data mining task in this study belongs to the association rules or affinity analysis problem. The association rules technique has been applied to discover general associations patterns among items in large databases [10][27].

Due to the Human Subjects 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.

A. 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, Majors have 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 has three colleges, and each college provides one major degree program, with several concentrations under each major. In this case, only two of the variables are 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 ALS: 1; BA: 2; EDU: 3

In this study, we adopt the second alternative. Students were classified into thirteen different groups As a result of converting all input variables into categorical or binary variables. The coding system is shown in Table I. Though more detailed categories could be designed [28], consideration of the number of samples in each category made more detailed categorization inappropriate for a medium sized university.

Table I. Converting input variables into binary or categorical variables.

Input	Domain	Coding	Definition	Type
Gender	{F, M}	F: 1 M: 2	gender of students	Categorical
Major	{ALS, BA, EDU}	ALS: 1 BA: 2 EDU: 3	one of the three university colleges (College of Arts, Letters, and Sciences, College of Business Administration, and College of Education) students belong to	Categorical
County	{Calaveras, Foreign Countries, Mariposa, Merced, other states, Others, San Joaquin, Stanislaus, Tuolumne}	Calaveras: 1 Foreign Countries: 2 Mariposa: 3 Merced: 4 other states: 5 Others: 6 San Joaquin: 7 Stanislaus: 8 Tuolumne: 9	the service counties students coming from	Categorical
1st_G_College	{N, Y}	N: 1 Y: 2	is the student a first-generation college student in the family	Binary
Transfer	{N, Y}	N: 1 Y: 2	is the student a transferred student	Binary
Original_College	{CA Independent, CSUS, Foreign, JC, Other CSU, Out of State, UC}	CA Independent: 1 CSUS: 2 Foreign: 3 JC: 4 Other CSU: 5 Out of State: 6 UC: 7	the original institution a student came from, such as regional community colleges or other four-year colleges or universities	Categorical
Ethnicity	{American Indian, Asian Pacific, Black, Hispanic, Intl, Other, White}	American Indian: 1 Asian Pacific: 2 Black: 3 Hispanic: 4 Intl: 5 Other: 6 White: 7	the ethnicity of a student	Categorical
Classification	{Freshman, Junior, Post Bacc, Senior, Sophomore}	Freshman: 1 Junior: 2 Post Bacc: 3 Senior: 4 Sophomore: 5	the class status of a student, including freshman, sophomore, junior, senior, and post-baccalaureate	Categorical
Status	{F, P}	F: 1 P: 2	full time or part time students	Categorical
Age<=24	{N, Y}	N: 1 Y: 2	if a student is 24-year old or younger	Binary
Married	{N, Y}	N: 1 Y: 2	the marital status of a student	Binary

Financial_Aide	{N, Y}	N: 1 Y: 2	if a student needs a financial aide	Binary
Transfer_or_Dropout	{Dropout, Stay, Other College}	Dropout: 1 Stay: 2 Other college: 3	transferring to other four-year colleges or universities, completing the entire undergraduate education at CSUS, or dropping out of school	Categorical

After data collection, we ensured the quality of the dataset by removing, i.e., "scrubbing" [29], 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 were removed.

B. *Data Mining Phase*

Our system consists of two modules: one is a decision tree which is responsible for the prediction and classification of a student's decision to persist at the college until graduation; the other module conducts affinity analysis by discovering the relationship between student attributes and a student's decision. This can be regarded as the meta-data generated by the affinity analysis and will be deposited into the knowledge base of the institution. Through the adaptive interaction between administrative end users and the knowledge base, strategies for student retention can be developed gradually. See Fig 1.

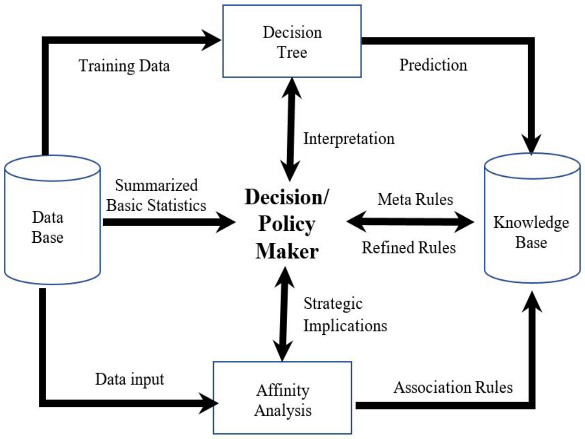

Fig 1. A two module-driven knowledge discovery process for forming student retention knowledge base.

Conducting Decision Tree Analysis for Prediction

Due to the consideration of the interpretability issue of machine learning techniques, we predicted a student's decision by developing a decision tree. The easy-to-understand tree structure has made the decision tree a most popular tool for tasks of classification and regression.

There are two processes employed to construct classification trees: One is recursive partitioning of the space of the independent variables (i.e., predictors); the other one is pruning using a validation dataset. In classification, the dependent variable Y is a categorical variable. Recursive partitioning divides up the p-dimensional space of the independent variables (i.e., predictors) X_1, X_2,…, X_p into non-overlapping multi-dimensional hyper-cubes. The division is performed recursively by dividing up the results of previous divisions. First, one of the predictors, X_i, which has the smallest entropy, is selected, and a value c_i is chosen to split the p-dimensional space into two parts: one part consists of points with $X_i \leq c_i$, and the other consists of points with $X_i > c_i$. In this study, the splitting value, c_i, for a predictor is simply the midpoints between pairs of consecutive values of this predictor. Entropy is a measure of impurity (or heterogeneity) of a dataset, and is defined as: $\sum_{k=1}^{m} p_k \, log_2(1/p_k)$, where $\sum_{k=1}^{m} p_k = 1$, and k =1, 2,…, m.

Then, one of these two parts is split in a similar manner by choosing a predictor and a splitting value for the predictor again. This results in three rectangular regions. This process continues. The decision tree approach to classification is through recursive partitioning to divide the entire search space into rectangular regions such that, eventually, each rectangle is as homogeneous (i.e., belonging to the same class) as possible.

The beginning state of our training dataset is: 105/600 of class 1, 422/600 of class 2, and 73/600 of class 3. Thus, the initial entropy of our training set is:

$$(105/600)\log_2(600/105) + (422/600)\log_2(600/422) \\ + (73/600)\log_2(600/73) = 1.1669$$

Next, since we were constructing a binary decision tree based on categorical variables, we needed to convert each predictor variable into a binary variable, for instance, the predictor ***County*** with nine categories. We converted the variable into a binary variable by treating {1}as one category and {2, 3, 4, 5, 6, 7, 8, 9} as the (aggregated) second category. We adopted the weighted entropy value to measure the heterogeneity (or randomness) of each splitting predictor. The splitting predictors are ranked based on how much they reduce heterogeneity. The weighted entropy for each splitting predictor is shown in Table II.

Table II. Weighted entropy associated with each splitting predictor.

Predictors	Weighted Entropy	Info Gain (the training set initial entropy – weighted entropy)
Gender	1.1619	0.00495
County	1.1620	0.00489
Major	1.1471	0.01982
1st_G_ College	1.1624	0.00453
Transfer_In	0.9676	0.19926
Original_College	1.1629	0.00403
Ethnicity	1.1601	0.00682
Classification	0.9951	0.17175
Status	1.1645	0.00238
Age<=24	1.1434	0.02346
Married	1.0366	0.13025
Financial_Aide	1.1196	0.04725

From Table II, the predictor Transfer_In has the smallest weighted entropy (i.e., the largest information gain). Thus, it was used as the first splitting predictor. We computed the entropy for both branches. The combined heterogeneity of the two rectangles resulting from the split is a weighted average of the two entropies, weighted by the number of records in each class. By comparing the reduction in heterogeneity across all predictors, the next split is chosen.

Thus, as this process proceeds, this set of constituent rectangles will become more homogeneous. Finally, a full-grown tree will develop.

Usually, the full tree is too widespread to be easily displayed on a page. In addition, the full tree might over fit the training data and thus will not perform well when encountering a new case. We can improve the full tree by fitting it to training dataset and pruning it using the validation dataset. We pruned the full tree by one decision node at a time by applying the full tree to the validation dataset. We noticed that a truncated tree with 5 decision nodes had the lowest validation error rate and the fewest number of nodes as shown in Table III. We showed the truncated tree which made the minimum error rate in Fig 2.

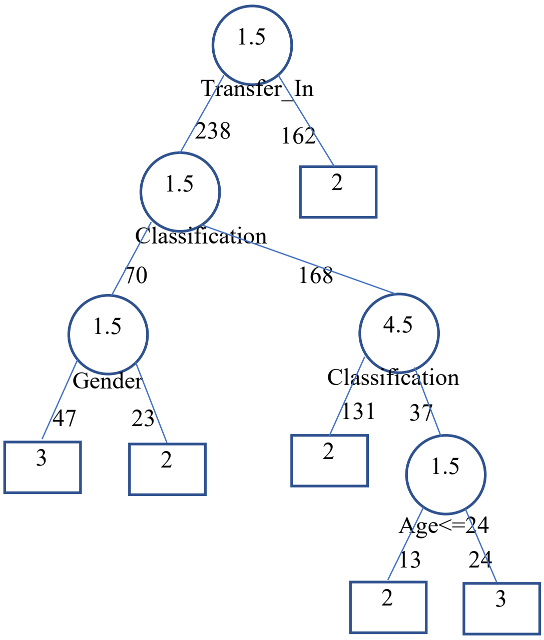

Fig 2. The minimum error rate truncated tree with five decision nodes based on the validation dataset.

The above classification tree reads as follows: When Transfer_In is larger than 1.5 (i.e., "Y"), 162 students are classified as "Stay". In other words, transferred students chose to stay with us. When Transfer_In is less than or equal to 1.5 (i.e., "N"), there are 238 students who are still yet to be determined by the next decision node—Classification. There are 70 students with Classification ≤ 1.5 (i.e., freshman), who need to be determined by one more decision node—Gender. Among those, 47 students are female (Gender ≤ 1.5) and transferred to other institutions, while the rest of 23 students are male and stayed with us. Going back to the Classification node, there are 168 students with Classification > 1.5 (i.e., other than freshmen). Of these 168 students, 131 students are juniors, post baccalaureate or seniors, and who chose to stay with us, while the other 37 students, who are classified as sophomores, still need to be determined by one more factor—Age. Among these 37 sophomore students, 13 were older than 24 and chose to stay with us, while the other students who were younger than 24 chose to transfer to other institutions. The error rate made by this truncated tree when applied to the validation tree is only 22.25%.

Table III. Validation error rate made by trees with different number of decision nodes.

# Decision Nodes	% Validation Error	
20	23.5	
19	23.5	
18	23.5	
17	23.5	
16	23.5	
15	23.5	
14	23.5	
13	23.5	
12	22.25	
11	22.25	
10	22.25	
9	22.25	
8	22.25	
7	22.25	
6	22.25	
5	**22.25**	**Min. Err. Tree**
4	22.75	
3	23.25	
2	23.25	
1	26	
0	26	

Conducting Affinity Analysis for Association Rule Identification

We also applied the association rule analysis to identify which group of student attributes tends to be associated with a student's decision on staying with us until graduation. Result of this analysis can help us develop more effective strategies for retaining students. The discovered relationships among attributes are represented as an *antecedent* ➔ *consequent* type of rule and can help the administrators compare and understand the behavioral patterns of students in different groups. 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 are probabilistic [30].

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 sometimes expressed as a percentage of the total number of records in the dataset is 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 and measures the strength of association between $\boldsymbol{A}$ and $\boldsymbol{C}$, i.e., the degree of uncertainty about the rule. Confidence is the ratio of the number of records that include all items in $\boldsymbol{C}$ and $\boldsymbol{A}$ (i.e., the support) to the number of records that include all items in $\boldsymbol{A}$. In other words, the confidence is an estimated conditional probability that a randomly selected record will include all the items in $\boldsymbol{C}$ given that the record includes all the items in $\boldsymbol{A}$. 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 $\boldsymbol{A}$ and $\boldsymbol{C}$ 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:

$$\frac{P(\boldsymbol{A}) \times P(\boldsymbol{C})}{P(\boldsymbol{A})} = P(\boldsymbol{C})$$

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

The larger the lift ratio, the greater the strength of the association.

We conducted association analysis for our student dataset by using XLMiner®. In this study, 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 a group of students fits 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. This original knowledge base consisted of 424 association rules. 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* AND *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 the proportion of transactions 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 as desirable as a less efficient rule with strong support. The confidence shows the rate at which consequents (*C*) will be found among the transactions involving the antecedent (*A*). 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 review Support(*A*) and Support(*A* AND *C*). If Support(*A*) is already low, then even though the rule has high confidence, the rule is still not valuable.

C. Knowledge Base Refinement by Inductive Inference

The original output included 424 rules. We applied rules of inductive inference to refine the knowledge base, such as rule generalization, rule elimination, rule specialization, rule combination, and rule subsumption. After reviewing these rules, we found that some rules involved the same set of factors, with different antecedents and consequents. Those rules could be combined. Since we were interested in learning about the characteristics of different groups of students who chose to stay with us, transfer to other institutions, or drop out of schools, we deleted those association rules not containing the factor Transfer or Dropout. After refinement, the reduced rule set (i.e., the knowledge base) consisted of 137 rules. One parameter, the lift ratios, quantifies the usefulness of a rule. Rules with lift ratios further from 1 are more useful than rules with lift ratios closer to 1, which implies $\boldsymbol{A}$ and $\boldsymbol{C}$ are independent events. Thus, knowledge of the occurrence of $\boldsymbol{A}$ does not alter the probability of the occurrence of $\boldsymbol{C}$. The higher the lift value, the more useful the rules are.

To refine the model further, some rules are trivial and can be removed from the rule set, and some can be combined. For example, the following two rules can be merged:

IF 1st G College = "N" and Gender = "F" and Major = "ALS" THEN Transfer or Dropout = "Stay", and

IF 1st G College = "N" and Gender = "F" THEN Transfer or Dropout = "Stay"

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

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

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

We eliminated the trivial rules and combined rules into a smaller set of rules. In this study 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 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 IV. This set of rules populates the student retention knowledge base.

V. 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

Table IV. Further reduced association rules output for the knowledge base.

Rule Conf. (%)	Ant (*A*)	Cons (*C*)	Sup (*A*)	Sup (*C*)	Sup (*A* U *C*)	Lift
95.33	Original College_JCs=>	Transfer or Dropout_Stay	364	406	347	2.3480
94.66	Gender_F, Transfer_Y=>	Transfer or Dropout_Stay	281	718	266	1.3184
97.75	Classification_S, Transfer_Y=>	Transfer or Dropout_Stay	222	718	217	1.3614
97.23	1st G College_N, Transfer_Y=>	Transfer or Dropout_Stay	253	718	246	1.3542
95.92	Classification_S=>	Transfer or Dropout_Stay	245	718	235	1.3359
95.31	Transfer_Y=>	Transfer or Dropout_Stay	426	718	406	1.3274
95.16	Financial Aide_Y, Transfer_Y=>	Transfer or Dropout_Stay	289	718	275	1.3253
94.98	Status_F, Transfer_Y=>	Transfer or Dropout_Stay	279	718	265	1.3229
94.85	Married_N, Transfer_Y=>	Transfer or Dropout_Stay	233	718	221	1.3210
94.51	Major_ALS, Transfer_Y=>	Transfer or Dropout_Stay	328	718	310	1.3163
93.86	Major_ALS, Status_F, Transfer_Y=>	Transfer or Dropout_Stay	228	718	214	1.3072
93.78	Financial Aide_Y, Major_ALS, Transfer_Y=>	Transfer or Dropout_Stay	225	718	211	1.3061
77.69	Financial Aide_Y, Major_ALS=>	Transfer or Dropout_Stay	363	718	282	1.0820
93.53	Gender_F, Major_ALS, Transfer_Y=>	Transfer or Dropout_Stay	232	718	217	1.3027
86.89	Classification_J=>	Transfer or Dropout_Stay	267	718	232	1.2102
80.56	Major_ALS, Married_Y=>	Transfer or Dropout_Stay	252	718	203	1.1219
79.46	Age <= 24_N, Financial Aide_Y=>	Transfer or Dropout_Stay	297	718	236	1.1067
79.45	Age <= 24_N, Major_ALS=>	Transfer or Dropout_Stay	253	718	201	1.1065
78.41	Age <= 24_N, Status_F=>	Transfer or Dropout_Stay	301	718	236	1.0920
93.52	Age <= 24_Y, Transfer_Y=>	Transfer or Dropout_Stay	247	718	231	1.3025
76.58	Age <= 24_N=>	Transfer or Dropout_Stay	444	718	340	1.0665
78.29	Married_Y, Status_F=>	Transfer or Dropout_Stay	281	718	220	1.0904
77.96	Financial Aide_Y, Married_Y=>	Transfer or Dropout_Stay	363	718	283	1.0858
77.03	Married_Y=>	Transfer or Dropout_Stay	444	718	342	1.0728
76.88	Financial Aide_Y, Status_F=>	Transfer or Dropout_Stay	359	718	276	1.0708
76.41	Financial Aide_Y=>	Transfer or Dropout_Stay	568	718	434	1.0642
76.43	Gender_M=>	Transfer or Dropout_Stay	314	718	240	1.0645

the combination of multiple rules further implications regarding the factors important to a student's decision to stay with the university until graduation, to transfer to another institution, or to drop out of college. These implications can also be adapted by the student retention office as guidelines for developing student retention strategies.

Table V. Interpretation of the rules in Table 4.

Rule	Interpretation	Conf %	Lift Ratio
1	For junior-college transfers, about 95% of them transfers, around 95% of them are about 2.35 times as likely to stay until graduation as non-junior-college transfers.	95.33	2.35
2	For female transfers, around 95% of them are about 1.32 times as likely to stay until graduation as the remaining students.	94.66	1.32
3	For the senior transfer students, around 98% of them are about 1.36 times as likely to stay until graduation as the remaining students.	97.75	1.36
4	For the non-1st generation transfer students, around 97% of them are about 1.35 times as likely to stay until graduation as the remaining students.	97.23	1.35
5	For senior students, 96% of them are about 1.34 times as likely to stay until graduation as non-senior students.	95.92	1.34
6	For transfer students, 95% of them are about 1.33 times as likely to stay until graduation as non-transfer student.	95.31	1.33
7	Among those transfer students with financial needs, about 95% of them are about 1.32 times as likely to stay until graduation as the remaining students.	95.16	1.33
8	Among transfer full-time students, 95% of them are expected to be 1.32 times as likely to stay until graduation as the remaining students.	94.98	1.32
9	Among transfer unmarried students, almost 95% of them are expected to be 1.32 times as likely to stay until graduation as the remaining students.	94.85	1.32
10	Among transfer ALS students, about 94.5% of them are 1.32 times as likely to stay until graduation as the rest of the students.	94.51	1.32
11	Among transfer full-time ALS majors, about 94% of them are 1.31 times as likely to stay until graduation as the rest of the students.	93.86	1.31
12	Among transfer ALS majors with financial needs, about 94% of them are 1.31 times as likely to stay until graduation as the remaining students.	93.78	1.31
13	For the ALS majors with financial needs, about 78% of them are 1.08 times as likely to stay until graduation as the remaining students.	77.69	1.08
14	Among transfer ALS female students, about 93.5% of them are 1.30 times as likely to stay until graduation as the remaining students.	93.53	1.30
15	Among junior students, about 87% of them are 1.21 times as likely to stay until graduation as the non-junior students.	86.89	1.21
16	Among married ALS students, almost 81% of them are 1.12 times as likely to stay until graduation as the remaining students.	80.56	1.12
17	For students older than 24 and with financial needs, about 79% of them are 1.11 times as likely to stay until graduation as the remaining students.	79.46	1.11
18	For those ALS students older than 24, about 79% of them are 1.11 times as likely to stay until graduation as the remaining students.	79.45	1.11
19	For those full-time students older than 24, almost 78% of them are 1.10 times as likely to stay until graduation as the remaining students.	78.41	1.09
20	For those transfer students younger than 24, almost 94% of them are 1.30 times as likely to stay until graduation as the remaining students.	93.52	1.30
21	For students older than 24, about 76% of them are 1.07 times as likely to stay until graduation as students younger than 24.	76.58	1.07

22	For those married full-time students, about 78% of them are 1.09 times as likely to stay until graduation as the remaining students.	78.29	1.09
23	For those married students with financial needs, almost 78% of them are 1.09 times as likely to stay until graduation as the remaining students.	77.96	1.09
24	For the married students, about 77% of them are 1.07 times as likely to stay until graduation as the un-married students.	77.03	1.07
25	For those full-time students with financial needs, almost 77% of them are 1.07 times as likely to stay until graduation as the remaining students.	76.88	1.07
26	For students with financial needs, about 76% of them are 1.06 times as likely to stay until graduation as students without financial needs.	76.41	1.06
27	For male students, about 76% of them are 1.06 times as likely to stay until graduation as the remaining students.	76.43	1.06

A. Model Interpretation

The corresponding IF-THEN rules and interpretations of Table IV are provided in Table V. This set of rules has confidence above 76.4%, higher than many confidence settings for most of other applications. Each rule in this rule set has lift ratio above 1.0.

Table VI. Identifying strategically important factors by single-condition, single-factor rules.

Rule 1	Rule 5	Rule 6	Rule 21	Rule 24	Rule 26	Rule 27
Original college	Classification	Transfer_in	Age	Married	Financial aide	Gender
0.95	0.96	0.95	0.77	0.77	0.76	0.76
2.35	1.34	1.33	1.07	1.07	1.06	1.06

B. Model Strategic Implications

From Table V we can derive from each rule or from the combination of multiple rules further implications regarding the factors important to a student's decision. At first, we use Table V in identifying the rules with a single condition in the antecedent. We call the corresponding factors strategically important factors (see Table VI). The implication of each strategically important factor is shown in Table VII.

Table VII. Implications of strategically important factors.

Rules	Strategical Implications
1	*Majority of JC transfers are much more likely to stay until graduation than those transferred from other types of institutions. In other words, JC transfers generally have much higher level of loyalty than transfers from other types of colleges. So, JC students are a most important group of students to retention.*
5	*Majority of senior students are much more likely to stay until graduation than non-senior students. Senior students are another most important group of students to retention.*
6	*Majority of transfer students are much more likely to stay until graduation than non-transfer students. Transfer students are another most important group of students to retention.*
21	*For almost every 7.6 out of 10 students who are older than 24 are more likely to stay until graduation than students who are 24 or younger. So, student age is another important factor to retention.*
24	*For almost every 7.7 out of 10 married students are more likely to stay until graduation than unmarried students. So, student marital status is another important factor to retention.*
26	*For almost every 7.6 out of 10 students who have financial needs are more likely to stay until graduation as the remaining students. So, student financial needs is another important factor for us to monitor for retention.*
27	*For almost every 7.6 out of 10 male students are more likely to stay until graduation than the remaining students. So, student gender is another important factor to retention.*

Some rules might exhibit reinforcing effects through interacting with each other. We apply inductive inference to eliminate duplicate rules and rules subsumed by other rules, and we combine multiple rules for generating new rules. For each of the important factors identified in Table VI, we search for rules whose antecedents include this important factor and another factor. The process generates a new set of two-condition rules. We compare every two-rule set to rank the influence of those factors included in these two rules. We repeat the processes for more interacting factors until there are no more for comparison. For example, according to Table VI, Rule 6 has a single-factor antecedent consisting of Transfer_In. According to the above algorithm, a two-factor (including Transfer_In) rule set is generated as shown in Table VIII.

Table VIII. Implications of rules with reinforcing interactions with each other.

Interacting Rules	**Strategical Implications**
Rule 2 × Rule 6	*From these two rules, we see there is a reinforcing interaction between a student's transfer status and the gender of a student in affecting a student's decision. It seems that a student's transfer status is more important than a student's gender in affecting a student's decision.*
Rule 3 × Rule 6	*From these two rules, we see there is a reinforcing interaction between a student's transfer status and the classification of a college student in affecting a student's decision to stay until graduation. Among the transfer students, senior students tend to be more likely to stay with us than those as non-senior students.*
Rule 4 × Rule 6	*From these 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 × Rule 15	*From these two rules, we understand that senior students are more likely to stay until graduation than students of other standings.*
Rule 6 × Rule 8	*From these 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. 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 × Rule 6	*From these 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 × Rule 6	*From these 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. 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 12 × Rule 13	*From these 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 × Rule 14	*From the above two rules, it seems that a student's transfer status and major are more important than Gender 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 × Rule 24	*From these 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 × Rule 21	*From these two rules, it seems that a student's age is more important than the financial need in affecting a student's decision. 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 × Rule 21	*From these 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 × Rule 6	*From these 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. 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 × Rule 21	*From these 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 × Rule 24	*From these two rules, it seems that the full-time status of a student does not affect a student's decision 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 × Rule 23	*From these 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 × Rule 26	*From these 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.*
Rule 3 × Rule 5 × Rule 6	*From these 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 11 × Rule 8 × Rule 6	*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.*

The antecedent of Rule 2 consists of Gender and Transfer_In with confidence 0.9466. Comparing with Rule 6 whose confidence was 0.9531, we can infer, at the macro-level, that there is a reinforcing interaction between Rule 2 and Rule 6. Further investigation is needed to understand how Gender and Transfer_In interact with each other, at the micro-level to reinforce the interaction between Rule 2 and Rule 6 in affecting a student's decision. We denote the interaction between Rule 2 and Rule 6 as: Rule 2 × Rule 6. Next, Rule 3 is Classification and Transfer_In. Different confidence ratios show Gender and Classification having different levels of reinforcing interactions with Transfer_In. There is a reinforcing interaction between Classification and Transfer_In. We thus infer the influence order as: (Classification × Transfer_In) > (Gender × Transfer_In).

Similarly, Rule 4 has a stronger reinforcing interaction between 1st_G_College and Transfer_In than that of (Gender × Transfer_In) in Rule 2. Thus, the influence order is: (Classification × Transfer_In) ≈ (1st G_College × Transfer_In) > (Gender × Transfer_In). Similarly, we can infer the importance of reinforcing interactions for the remaining factors, and thus for the entire set of paired factors as:
(Classification × Transfer_In) ≈ (1st G_College × Transfer_In) > (Gender × Transfer_In) > (Financial Aide × Transfer_In) > (Status × Transfer_In) > (Married × Transfer_In) > (Major × Transfer_In) > (Age × Transfer_In). (Note that "≈" means "almost equal to.")

VI. Discussion and Conclusions

In this study, we followed a three-phase-six-stage adaptive data mining cycle in developing a decision tree and association rules hybrid system for investigating strategically important factors which can be used to provide administrators suggestions regarding retention strategy development. A student data file of 1,000 records was employed for this purpose 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. According to [30] college institutions fail to translate what they learn 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, our study showed that Original_college was the strongest factor, and Classification and Transfer_in were the next strongest factors influencing a student's decision. We also found out that 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 studies, a recent study [16] proposed that the financial factor was one of the most important factors. It is worthwhile to investigate what might have caused such a difference.

VII. Future Research

In their studies, [31] developed a logit model to predict the retention of graduate students. Their predicted retention rate for graduate students was independent of age and gender, but was a function of ethnicity, registration status, and the interactions between academic division and registration status and between academic division and ethnicity. Still, according to [8], 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 applicability limitation. With the availability of a broader and more recent student data file, we might obtain 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 large data set. Still, COVID-19 has dramatically changed, not only the teaching modality, but also administrative and strategic processes, including reduced student retention activities which may contribute to fallen student transfer rates [32]. Since the transfer student status was a most important factor affecting a student's decision, the effects of such a paradigm shifting will be a worthy topic for future research when more relevant data become available. Studies showed chatbot technology holds the promise of revolutionizing education [33]. Using multi-modes of communications, chatbots may interact with administrators individually or collaboratively. We plan to investigate the role of a chatbot as a strategic decision-making partner for creating cost-effective processes for combating the COVID challenges. We believe this will be an exciting and valuable direction to pursue.

Since CSUS has historically served under-represented students, we plan to include this factor as a predictor for future studies. Especially, loss of STEM majors continues to be an ongoing national concern. According to [34] there were millions of job openings, yet only a small fraction of today's workforce has the skills to keep up with innovations. To address our share of this multifaceted problem, we plan to build support structures—including peer and alumni mentorship—that significantly affect retention, offer career visibility for marginalized student population, engage alumni seeking to encourage student persistence, and fill the talent gap with a population that represents modern society.

Most research on educational data mining focuses on developing the best model for identifying important factors. However, deployment of the learned model as an integrated component of daily administrative processes is less emphasized [35]. One of our plans is to study how to best adapt student retention practices by leveraging the model-generated predictive warning signs that a student is in academic peril.

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 to help a college develop better understanding of 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; cultivate a familial environment; connect students with financial resources; 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.

Acknowledgment

We would like to express our appreciation to Dr. Mark Thompson of California State University, Stanislaus, for improving the technical writing of this manuscript.

References

[1] M.H. Dunham, *Data Mining*. Upper Saddle River, NJ: Pearson, 2003.

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

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

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

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

[6] L. Ang and F. Buttle, "Customer retention management processes: A quantitative study," *European Journal of Marketing*, vol..40 no.½, pp. 83-99, 2006.

[7] O. Asil, "A hybrid data analytic approach to predict college graduation status and its determinative factors," *Industrial Management & Data Systems,* vol.116, no. 8, 2016.

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

[9] J. Roberts, "Professional staff contributions to student retention and success in higher education," *Journal of Higher Education Policy and Management*, vol. 40, no. 2, pp. 140-153, 2018.

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

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

[12] J. Kleinberg, C. Papadimitriou, P. Raghavan, "A microeconomic view of data mining," *Data Mining & Knowledge Discovery*, vol. 2, no. 4, pp. 311-324, 1998.

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

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

[15] E.L. Dey, A.W.Astin, “Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression,” *Research in Higher Education*, vol. 34, no.5, pp. 569–581, 1993.

[16] D. Delen, “A comparative analysis of machine learning techniques for student retention management,” *Decision Support Systems*, vol. 49, no. 4, pp. 498-506, 2010.

[17] R. Villano, S. Harrison, G. Lynch, & G. Chen, “Linking early alert systems and student retention: a survival analysis approach”, *Higher Education*, vol. 76, no. 5, pp. 903–920, 2018.

[18] J.P. Grayson, “Racial origin and student retention in a Canadian University,” *Higher Education*, vol. 36, no. 3, pp. 323–352, 1998.

[19] J.W. Jia and M. Mareboyana, “Undergraduate Student Retention Prediction Using Wavelet Decomposition,” in G.C. Yang, S.I. Ao, L. Gelman (eds), *Transactions on Engineering Technologies*. Springer, Dordrecht, pp. 643-655, 2015.

[20] M.A.Kuhail, N. Alturki, S. Alramlaw, and K. Alhjon, “Interacting with educational chatbots: A systematic review,”Educatiob and InformationTechnologies, vol.28, pp.973-1018, 2023.

[21] J.N. Wetzel, D. O’Toole, & S. Peterson, “Factors affecting student retention probabilities: A case study,” *Journal of Economics and Finance*, vol. 23, no.1, pp. 45–55, 1999.

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

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

[24] M. Du, N. Liu, & X. Hu, “Techniques for interpretable machine learning,” *Communications of the ACM*, vol. 63, no.1, pp. 68-77, 2020.

[25] P.S. Deng, “Using affinity analysis-driven adaptive data mining life cycle for the development of a student retention DSS,” *WSEAS Transactions on Advances in Engineering Education*, vol. 18, pp. 135-147, 2021.

[26] P.S. Deng, “Applying a market-based approach to the development of a sharing-enabled KM model for knowledge-intensive small firms,” *Information Systems Management*, vol. 25, no. 2, pp. 174-187, 2008.

[27] G. Shmueli, G., P.C. Bruce, N.R. Patel,, *Data Mining for Business Analytics, 3rded.* New York: Wiley, 2016.

[28] M. Heagney, “Student success and student diversity,” in *Improving student retention in higher education: The Role of teaching and learning*, G. Crosling, L. Thomas, & M. Heagney (eds.), London: Routledge. 2008.

[29] N. Mercurius, “Scrubbing data for D3M,” *T.H.E. Journal*, Oct., pp. 15-18, 2005.

[30] V. Tinto, “From theory to action: Exploring the institutional conditions for student retention,” in J. Smart (ed), *Higher Education: Handbook of Theory and Research,* vol. 25, Springer, Dordrecht, 2010.

[31] M.D. Ott, T.S. Markewich, & N.L. Ochsner, ”Logit analysis of graduate student retention,” *Research in Higher Education*, vol. 21, no. 4, pp. 439–460, 1980.

[32] S. Weissman, “Report: Ripple effects of the pandemic on student retention” [Blog post]. Retrieved from https://www.insidehighered.com/quicktakes/2022/01/28/report-ripple-effects-pandemic-student-retention , January 28, 2022.

[33] M.A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, “Interacting with educational chatbots: A systematic review,” *Education and Information Technologies*, vol. 28, pp. 973-1018, 2023.

[34] B. Scarpelli, N. Miller, & R. Stephens, “State of the app economy, 5th ed. Available at https://actonline.org/wp-content/uploads/App_Economy_Report_2017_Digital.pdf.

[35] M.T. Nietzel. (2022, October 01). How colleges are using artificial intelligence to improve enrollment and retention [Blog post]. Retrieved from https:// www.forbes.com/sites/michaeltnietzel/2022/10/01/colleges-are-turning-to-artificial-intelligence-to-improve-enrollment-and-retention/