RESEARCH ARTICLE OPEN ACCESS

Predicting Student Attrition and Factors Affecting Attrition in Higher Education

Chandandeep[1]
DCS, PU, Patiala, India
ckalra16@gmail.com

Kawaljeet Singh[2]
UCC, PU Patiala, India
singhkawaljeet@pbi.ac.in

Abstract—
India's higher education system is cited as one of the main contributors to the economic rise of India but Student attrition still remains a big challenge to many institutions at large. For every Indian graduate per unit of population, there are 30 non-Indian graduates and the gap is still getting larger. The reason that propulated or stimulated this study was to identify factors affecting student attrition and to predict the student attrition rate in higher education using data mining techniques. Study is based upon the mixed method approach which consists of two phases. The first phase is the quantitative approach. Quantitative data was collected from 135 freshman students using questionnaire based upon factors considered influential based upon analysis part of the study. Results were then further analyzed using SPSS and based upon the results a C program is implemented calculating ―ENTROPY using basics of ID3 and C4.5 algorithm. The results indicate that the information conveyed is 0.845, which is quite reliable; also it calculates the gain ratios for all the influential and dominant factor contributing to the information gain, i.e. affecting attrition WEKA tool applied J48and built a decision tree whose nodes were the verified by the results of C program.

The second phase has been kept as a future work, utilizes a qualitative approach that will verify the results and conclusion drawn by the first phase. Intervention strategies will be made upon these conclusions.

***Keywords*—**Attrition, Entropy, Social integration, Institutional commitment, Academic integration, Goal commitment.

I. INTRODUCTION

With the emergence of the fact that India is getting aware of the importance of higher education and the universities are trying to cash out the situation with the increased quantity and not the quality leading to a serious issue of immense growth in student attrition.

Only 17 percent of Indian students go on to college from high school. And since 50 percent of these students drop before graduation, only 8.5 percent of them enter college. And 82 percent of these college students drop out before they graduate from college i.e. they never earn a degree. Thus, the ratio is of 1 is to 30 for every Indian college graduate to a non-Indian graduate, with the gap getting still larger.

To improve the above figures universities are now trying to encourage students to finish their studies by identifying from huge number of enrolled students, the students who are in need of mentoring. For this, it is important to explore methods that can extract reliable and comprehensive knowledge from the student data that allow prediction of students at risk with a sufficiently high accuracy. Large number of data mining techniques have been applied in past to analyze factors affecting student attrition and provides measures that can lead to better academic planning and management. So, the problem can be stated as: "**Predicting Student Attrition and Factors Affecting Attrition in Higher Education**".

In this study an analytical model has been built by applying classification algorithm on institutional data and the environmental factors, to predict freshmen student attrition and to verify the accuracy decision trees provide, as it has been stated in literature. The study also identifies if educational and financial variables are one of the prominent predictors of attrition in Indian higher educational environment.

This study is beneficial in several aspects. First, this study will contribute to the literature concerned with student attrition in higher education. Secondly it may be beneficial to faculty of educational institutes as it may give them a clear picture of the factors affecting student attrition and thus allow them to develop strategies that aim to prevent students from dropping out.

Finally, the study may be beneficial to future students and their parents since it will provide evidence of the influential factors affecting student attrition.

II. LITERATURE SURVEY

In [8], authors identified factors affecting student retention at King Saud University in Saudi Arabia. It has been estimated that 35% of university students leave higher education before completing their studies. This study was guided by Tinto's (1975) Student Integration Theory. He in his study uses a mixed methods approach which consists of two phases. The first phase was the quantitative approach. Quantitative data were collected using two questionnaires and analyzed using a structural equation modelling (SEM) technique using the AMOS software package. The second phase of this study utilized a qualitative approach. Qualitative data were obtained from three sources: nonpersister students, persister students, and staff members.

In [16], author at his department based on experience claims to be able to distinguish the potentially successful students from amongst the first year influx before the end of the first semester but the selection is only loosely based on assumed student similarities over the years. There is no thorough analysis. Data mining techniques applied using weka improve the accuracy of this prediction. Furthermore, these techniques may point out indicators of success that are missed until now.

A data mining project to generate predictive models for student retention management on campus is conducted [7]. Given new records of incoming students, these predictive models can produce accurate prediction lists identifying students who tend to need the support from the student retention program most. Quality of the predictive models generated has been examined. The results show that some of the machines learning algorithms are able to establish effective predictive models from the existing student retention data. Weka is used to generate predictive models that provide ways to predict whether a new student will continue to enrol or not after one year given the values of the other twenty one attributes.

In [17], author discussed that Classification methods like decision trees, rule mining, Bayesian network etc can be applied on the educational data for predicting the students behavior, performance in examination etc. The C4.5 decision tree algorithm is applied on student's internal assessment data to predict their performance in the final exam. To analyze the accuracy of the algorithm, it is compared with ID3 algorithm and found to be more efficient in terms of the accurately predicting the outcome of the student.

III. CROSS-INDUSTRY STANDARD PROCESS: CRISP–DM [4]

The study has been conducted according following the cross-industry standard which is industry neutral, tool-neutral, and application neutral –CRISP-DM. According to Fig 1, a data mining project always follows a life cycle of six phases. This phase sequence is *adaptive* i.e., the next phase in the sequence often depends on the results of the preceding phase. The most significant dependencies between phases are indicated by the arrows. The iterative nature is symbolized by the outer circle.

.

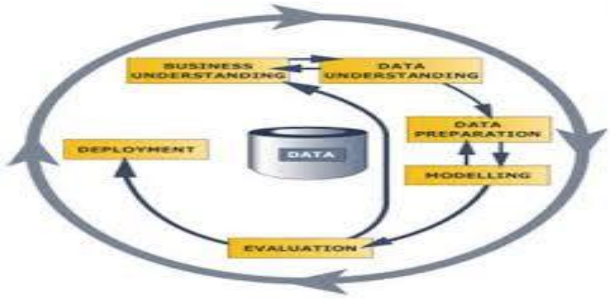


Fig 1: CRISP- DM

Following are the phases in this study, according to the data mining CRISP-DM rule.

3.1 Business Understanding

This was the very first phase that consists of research understanding. Thorough understanding of the causes and factors behind the problem was done. Basic strategy was to use decision trees to develop an analytical model to achieve the objectives. Initial model, Fig 2 was developed at the completion of this phase.

3.2 Data Collection: Questionnaire was designed for collection of quantitative data .The questionnaire measured the variables affecting the 4 constructs of the study.

4 constructs being:

- Initial goal and commitment
- Social integration
- Academic Integration
- Later Goal and Commitment

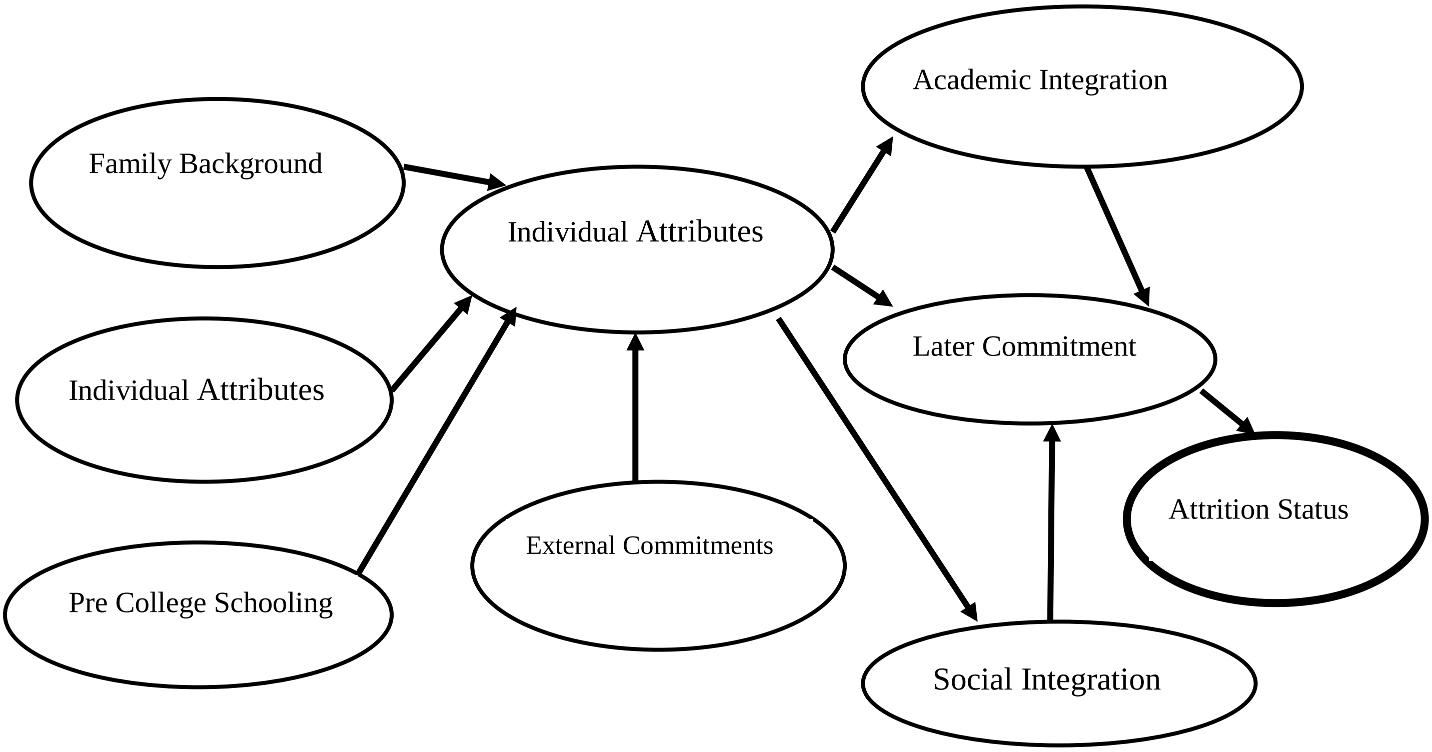


Fig 2: Initial Model

3.3 Data Preparation

The criteria for selection of the participants were that they were first time freshmen students. Selection is based upon the literature review which concludes that most students drop out during their freshmen year. The scales use a five-point Likert scale, ranging from strongly disagree, with a value of one, to strongly agree, with a value of five. The scale items are shown in Table 1.

Table 1: Scale Items

Family Background
1.1. What is your mothers and fathers highest formal education?
2. 2.I am always emotionally and financially supported by my parents regarding tuition's fee and other related expenses
2.4. I am committed to my studies and don't have any kind of family responsibilities on me.
Individual Attributes
1.3 Age Gender Category
Pre-College Schooling
1.4 Pre college specialization and percentage.
Initial Commitments
2.6. It's important for me to graduate else I feel misfit in my peer group (age group).
2.7. It's important for me to graduate because this'll help me to achieve my educational and career goals.
External Commitments
2.5. It is important for me to graduate because my parents (elders) expect me to do it.
Peer Group Interactions
2.11. Since coming to this institute I have developed close personal relationships with other students.
2.13. I often feel misfit in this institute because of differing values/attitudes/customs of the dominant peer group.
2.15. Interpersonal relationships with other students had a positive influence on my personal and intellectual growth.

Interaction with Faculty
2.16. Non-classroom interactions with the faculty have had a positive influence on my personal and intellectual growth.
Faculty Concern for Student Development & Teaching
2.18. Most of the faculty members are willing to spend time to discuss issues of interest to students.
2.22. The course curriculum and attitude of the faculty in this institute burdens and overstresses me.
Academic Development
2.20. I have performed academically well and I anticipate I would. 2.12. I have been attending more of the cultural events now than I was before coming to this institute.
Institutional & Goal Commitment
2.8. I have explored opportunities for transfer, that I feel will prove to be better than the institute I am attending.
2.10. It is likely that I will re-enrol at this college in next semester
2.21. It is likely that I will not re-enrol at this college in next semester.

Before starting the main study, a pilot study was carried out to check the clarity of the questions, to eliminate ambiguities if any, and to estimate time it would take to complete the questionnaires. The questionnaire was administered by the researcher with help from staff. Each member of staff was approached individually to request time in their classes for the administration of the questionnaires. The number of students in each class was between 40 to 50 students and the researcher attended 2 classes. A review of each student's records indicated that 36 of the 135 students had withdrawn voluntarily from the university at the end of their freshman year, while 96 had re-enrolled for their second year. The remaining three students had been required to withdraw for academic reasons. These students were excluded from the analysis because research suggested that voluntary withdrawals are significantly different from forced withdrawals. The final participants for the study consisted therefore of 132 students.

Qualitative approach: As for the Qualitative data a survey was administered which included 10 staff members. The survey included six questions regarding which student do they think are at the risk of dropping out and how can they motivate them to persist. Effective intervention suggestions are made based on the answers to these questions.

3.4 Modelling Phase

- **Pilot Analysis:** Pilot analysis on the data collected by the sample questionnaire prepared on some of the measures mentioned in Table 1, was done using SPSS a powerful program which provides many ways to rapidly examine data and test scientific hunches. The pilot analysis carried out on 17 students resulted in some ambiguous questions which were then modified. However it was seen that Likert scale with undecided option provided a fence to sit at and respondents preferred using it instead of answering the question reliably. SPSS was then used to check for the reliability of the questionnaire by calculating Cronbach's alpha, which resulted in value 0.794 which is reliable.

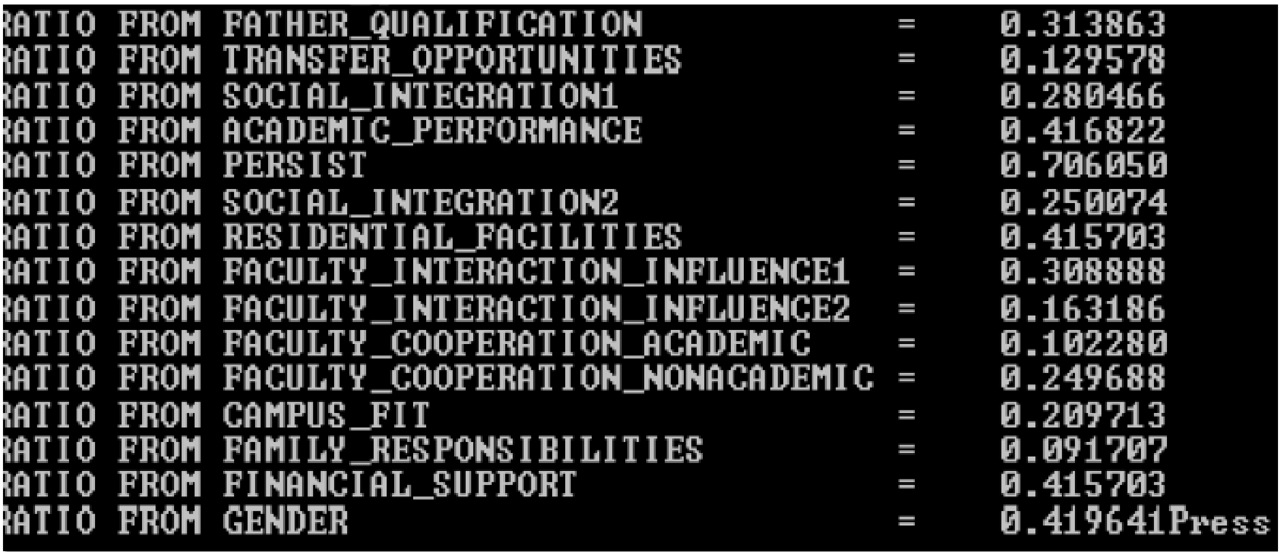
Based upon the pilot survey certain modifications were made and final questionnaire was prepared considering all the items of Table 1.

- **Missing Value Analysis (MVA) [9]:** We need to check the patterns of missing data and its univariate statistics to verify if the data is MCAR or MAR. Little's MCAR test, the null hypothesis is that the data are missing completely at random when *p* value is significant at the 0.05 level. If the value is less than 0.05, the data are not missing completely at random. For our study Little's MCAR test resulted in the significance value .811 which is above 0.05, hence verifies that our data is MCAR. EM method was used to impute the missing values.
- **Factor Analysis:** A common rule of thumb is that a researcher at least needs 10-15 participants per item. Yet, it largely depends on the proportion of variance in a dataset a factor explains how large a sample needs to be. If a factor explains lots of variance in a dataset, variables correlate highly with that factor, i.e. load highly on that factor. A factor having four or more loadings, greater than 0.6 is reliable regardless of sample size. Kaiser-Meyer-Okin (KMO) measure of sampling adequacy can signal in advance whether the sample size is large enough to reliably extract factors. When the KMO is near 0, it is difficult to extract a factor, since the amount of variance just two variables share (partial correlation) is relatively large in comparison with the amount of variance two variables share with other variables

(correlation minus partial correlation). When the KMO is near 1, a factor(s) can probably be extracted.. KMO in our case was .550 and hence verifies that the reliable factors can be extracted from the dataset considered.

3.5 Deployment Phase

The C program implemented uses the C4.5 algorithm [16] to calculate the Entropy or as we say the information that is conveyed by our collected data which resulted in .84 which is quite reliable, also based upon our analysis results of SPSS it calculates the gain and the gain ratios (Figure 3) factors that are determined to be critical to our study. The calculated gain and the gain ratios help us in verifying the sequence of the nodes of decision tree that has been obtained using the WEKA tool.



```
RATIO FROM FATHER_QUALIFICATION                  =     0.313863
RATIO FROM TRANSFER_OPPORTUNITIES                =     0.129578
RATIO FROM SOCIAL_INTEGRATION1                   =     0.280466
RATIO FROM ACADEMIC_PERFORMANCE                  =     0.416822
RATIO FROM PERSIST                               =     0.706050
RATIO FROM SOCIAL_INTEGRATION2                   =     0.250074
RATIO FROM RESIDENTIAL_FACILITIES                =     0.415703
RATIO FROM FACULTY_INTERACTION_INFLUENCE1        =     0.308888
RATIO FROM FACULTY_INTERACTION_INFLUENCE2        =     0.163186
RATIO FROM FACULTY_COOPERATION_ACADEMIC          =     0.102280
RATIO FROM FACULTY_COOPERATION_NONACADEMIC       =     0.249688
RATIO FROM CAMPUS_FIT                            =     0.209713
RATIO FROM FAMILY_RESPONSIBILITIES               =     0.091707
RATIO FROM FINANCIAL_SUPPORT                     =     0.415703
RATIO FROM GENDER                                =     0.419641Press
```

Fig 3: Gain Ratios from Factors Influencing Dropout Decision

The results of Figure 3 verify that educational factors i.e. academic performance and financial support with gain ratio of 0.416 and 0.415 are one of the dominant factors contributing in the prediction of dropouts.

3.5.1 WEKA Tool

J48 is applied to learning dataset using WEKA tool. There were 152 instances considered. Attributes to be worked on were 33. 86.842% of the dataset was training set and rest was the test set i.e. data from 132 students was used to train the analytical model and 20 students made the test set. J48 pruned tree was drawn, which gave number of leaves as 11 and size of the tree build is 21.

Evaluation on Test Split

- Correctly Classified Instances 14 82.3529 %
- Incorrectly Classified Instances 3 17.6471 %

```
            Confusion Matrix:
   a          b <------- classified as
                                    10        1
| a = Yes
                                     2        4
| b = No
```

- 14 instances are classified correctly.
- 3 were incorrectly classified.
- However there were 3 instances those were ignored as class unknown instances.
- 12 persisted, 5 dropped and 3 unknown.
- Dropout rate = 29 % and that of training set was 36/96 = 27 % which is quite a prediction.

3.5.2 Classification Algorithm Comparisons

Table 2 shows the results of different algorithms applied to the same field as analyzed in literature survey and conclude that J48 has the best accuracy and precision when compared to others.

Table 2: Comparison Results

Algorithms	**lazy.IBk (KNN=1)**	**trees.J48**	**rules.JRip**	**functions.Logistic**	**Bayes.NaiveBayes**	**Rules.OneR**
Accuracy	70.58	82.35	64.705	76.47	64.7059	64.70
Precision Yes		.833	.647	.769	.778	.647
No	.667	.8	0	.75	.5	0

3.5.3 Decision tree by J48 and Verification with gain ratios

Persist which is the root of the decision tree has the highest gain contribution of .706 followed by Gender and Academic performance of .419 and .416 respectively which is then followed by financial support of .415 and so on.

The Rule Set

It is easy to *derive a rule set from a decision tree*:
1. Persist>3 and Gender=F => Yes
2. Persist<=3 and Academic Performance >4 => No
3. Persist<=3 and Academic Performance >4 and Financial_Support>4=>Yes
4. Persist>3 and Gender=F and Financial_Support<=2 =>No
5. Persist<=3 and Academic Performance >4 and Financial_Support>4 and Social_Integration>4=> No
6. Persist<=3 and Academic Performance >4 and Financial_Support>4 and Social_Integration>4 and Campus_Fit>3=> Yes
7. Persist>3 and Gender=F and Financial_Support<=2 and Resedential_Facilities<=4 =>Yes

IV. Future Work

Staff Survey can be conducted that'll be including certain questions to verify the factors that analyzes of the student questionnaire resulted in and based upon the certain intervention strategies will be made that help students at risk of dropping out in continuing and completing their degrees.

References

[1] Larose, D.T., Discovering Knowledge in data, *"*An Introduction to Data Mining*", A John Wiley & Sons, Inc.*

[2] The Gartner Group, *www.gartner.com*.

[3] Mannila, H and Smyth, P., "Principles of Data Mining*", MIT Press, Cambridge, MA, 2001*.

[4] Peter Chapman et.al., "CRISP–DM Step-by-Step Data Mining Guide*", 2000, http://www.crisp-dm.org/*.

[5] Ramageri, B.M., "Data Mining Techniques and Applications", *Indian Journal of Computer Science and Engineering, 1,4, 2010, 301-305.*

[6] http://en.wikipedia.org/wiki/Educational_data_mining , *February 20, 2014.*

[7] Shieu-Hong Lin, "Data Mining for Student Retention Management", *http://csci.biola.edu/lin/LinCCSC12.pdf*

[8] Abdullah, S., "A study of the factors affecting student retention at King Saud university, Saudi Arabia: Structural Equation Modelling and Qualitative Methods", *University of Stirling 2008*.

[9] SPSS Base 16.0 User's Guide *http://www.spss.com.*

[10] Frank, E., "WEKA Explorer User Guide for Version 3-4-3", *November 9, 2004*.

[11] http://www.cis.temple.edu/~giorgio/cis587/readings/id3-c45.html

[12] https://students.shu.ac.uk/lits/it/documents/pdf/questionnaire_analysis_using_spss.pdf

[13] SPSS Missing Value Analysis™ 16.0, 2007 by SPSS Inc., *http://www.spss.com.*

[14] http://staff.neu.edu.tr/~ngunsel/files/Lecture%2011.pdf

[15] Bouckaert , R.R., "WEKA Manual for Version 3-6-5", *June 28, 2011.*

[16] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study", *Submitted to the 2nd Int. Conf. on Educational Data Mining (EDM '09), 2009*.

[17] Kumar S.A., and Dr. Vijayalakshmi M.N., "Implication Of Classification Techniques In Predicting Student's Recital*", International Journal of Data Mining & Knowledge Management Process (IJDKP), 1(5), 2011, 41-51.*