PREDICTING STUDENT ATTRITION USING DATA MINING PREDICTIVE MODELS

[1]MARIACRYSTAL E. OROZCO, [2]JASMINDE CASTRO – NIGUIDULA

[1,2]Technological Institute of the Philippines
E-mail: [1]Orozco.MariaCrystal@gmail.com, [2]JasminNiguidula@yahoo.com

Abstract - This paper demonstrates how educational data mining can help institution in decision-making specifically to reduce student attrition. Phases of CRISP-DM (Cross Industry Standard Process for Data Mining) methodologyare followed in order to determine students at-risk of dropping out after the first semester in their freshmen year. Predictive models namely, decision tree, naïve bayes, and rule induction were built and applied to process the data set. Subsequently, these models were tested for accuracy using 10-fold cross validation. Results show that, given sufficient data and appropriate variables, these models are capable of predicting freshmen attrition with roughly 80% accuracy. Moreover, the average grades of the students can be used as predictor in determining student attrition unlike the gender attribute that yielded no significant result.

Keywords - Student attrition, Cross industry standard process-data mining, Decision tree, Rule induction, Naïve bayes,

I. INTRODUCTION

Student attrition is a major concern in the education and policy-making communities worldwide.[1]In the Philippines, very few of the children who enter school when they reach the eligible age fail to complete at least 14 years of education, from elementary to college. [2]

Based on the Commission on Higher Education's (CHED) 2008 data, out of 100 Grade 1 pupils, only 66 finish Grade 6 and only 58 of them enroll in first year high school. Of the 58, only 43 finish high school. Of the 43 only 23 finished high school enrolled in college, and of the 23, only 14 eventually graduate from college. The dropout rate among college studentshas reached an alarming 83.7 percent. This means that the country is producing 2.13 million college dropouts annually while graduates stand at close to 500,000 only. [2]

Eventually, the number of additional students who can be identified as at risk will grow commensurately because institutions are likely to experience continued strong enrollment growth over the next ten years. Improvement in the degree completion time by better identification of at-risk students translates not only to higher graduation rates but also to substantial cost savings for students. [3]

With attention to year level of students, the extreme high attrition rates happen during the freshman year because of the difficulties students face in making the adjustment to college life. [4]

Meanwhile, a study revealed that institutional database variables (database-driven) out-perform the institutional integration survey (survey-based)scales in conducting institution-specific retention research. [5]In particular, more researches used data mining, a database-driven process used to discover patterns for a large data set. It is an expert system that uses its historical experience (stored in relational databases or cubes) to predict the future. [6]

Application of data mining in education sector, which is referred as educational data mining (EDM) is an emerging trend in the global competitive business. Understanding the data mining terms, tasks, techniques and application are foundation of developing data mining in education sector. [7] It is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process. [8]

This paper closely investigates freshmen attrition under five programs in a selected university in the Philippines to improve retention rates. It aims to develop analytical models to identify freshmen students who are most likely to drop after their first semester in college, to present the attrition of students according to gender and average grade, and to compare the accuracy of the three predictive models used.

II. RELATED WORKS

There are a number of studies related to predicting student attrition that have utilized different data mining methods.

One study focused on the experience of monitoring the first year student attrition in a large metropolitan multi-campus university during 2004-2010. It discussed the trends in student attrition which have been found and identified key issues which have been and need to be addressed by the university in order to increase retention.[9]

Another study examined the prediction of dropouts through data mining approaches in an online program using k-Nearest Neighbour (k-NN), Decision Tree (DT), Naïve Bayes (NB) and Neural Network (NN). These methods were trained and tested using 10-fold cross validation. [10]

A simple methodology based on decision tree was applied in a study to determine the set of students

that are likely to leave institution after confirming the admission. [11]
Rule induction was applied in predicting student attrition in nursing courses that resulted in 94% prediction accuracy, however, large amounts of high quality data was involved.[12]
Another research work proposed a prototype machine learning tool which can automatically recognize whether the student will continue their study or drop their study using classification technique based on decision tree and extract hidden information from large data about what factors are responsible for dropout student. Traditional metrics such as accuracy, precision, recall, and AUC were used to evaluate the performance of the models. [13]
Lastly, using eight years of student data obtained from the university's existing databases along with three popular data mining techniques (artificial neural networks, decision trees, and logistic regression), a study has developed analytical models to predict freshmen attrition. K-fold cross validation was used to split data and to train and test the models.[14]

III. METHODOLOGY

In this research, a popular data mining methodology called CRISP-DM (Cross Industry Standard Process for Data Mining was followed. Figure 1 shows its six-step process.

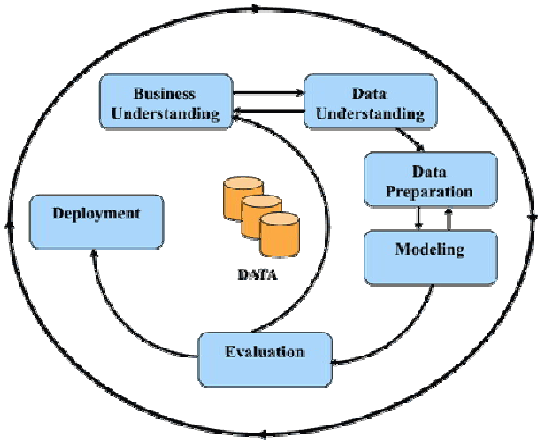


Figure 1 Process of CRISP-DM

CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These phases help organizations understand the data mining process and provide a road map to follow while planning and carrying out a data mining project. [15]
Most of the attention was focused on the first three steps of the CRISP-DM process because it sets the stage for successful data mining.

Phase One: Business Understanding
This phase has involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan. [15]
In determining business objectives, the institution would like to know which factors greatly affect student attrition. Upon assessing the situation, it is found that 20% of freshmen students leave the institution after first semester. This led towards a data mining goal, which is to identify freshmen students who are most likely to drop after their first semester. To achieve this goal, this data mining project plan was proposed.

Phase Two: Data Understanding
The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality. [15]
The data for this study came from a single institution with an average enrolment of 8000 per semester. Table 1 presents the freshmen student data under study. It is depicted that the average attrition rate in this institution is 12.85%.

Year	Total Freshmen Students in the First Semester	Returned for Second Semester	Freshmen attrition (%)
2012	760	669	11.97 %
2013	1000	933	6.70 %
2014	1201	1000	16.74 %
2015	1132	951	15.99 %
Total	**4093**	**3553**	**12.85** %

Table 1 Freshmen Student Data Used in this Study

The data consists of a four-year student enrolment data (academic years from 2012 to 2015) of five selected programs namely BSCS, BSIT, BSHRM, BSBA, and BST students. Highest population share belongs to BSHRM program and lowest to BSCS as shown in Table 2. Although BSHRM has the highest population share, it is the BSBA and BST which has the highest attrition rate.

Program	Population Share (%)		Attrition Rate
	1st Sem	2nd Sem	
BSBA	26.12	25.84	14.13
BSCS	4.23	4.33	10.98
BSHRM	27.46	27.58	12.81
BSIT	21.11	21.59	11.23
BST	21.08	20.66	14.95
Total	**100.00**	**100.00**	**12.82**

Table 2 Population Share and Attrition Rate per Program

Table 3presents the student enrolment data attributes and corresponding descriptions.

Attribute	Data Type	Possible Values
ID_Number	Polynomial	
School Year	Polynomial	2012, 2013, 2014.

		2015
Semester	Binomial	1, 2
Program Code	Polynomial	BSCS, BSIT, BSHRM, BSBA, BST
Gender	Binomial	M, F
Average Grade	Number	0 - 100
Attrition Rate	Binomial	Yes, No

Table 3. Enrolment Data Attributes with Description and Possible Values

Phase Three: Data Preparation

The data preparation phase covers all activities to construct the final data set or the data that will be fed into the modeling tools from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data.[15]

The data was collected from the institution's database. Also, it is consolidated with the use of RapidMiner, which is a code free data science platform that unites data preparation, machine learning, and model deployment.

In Figure 2, two raw data sets (first semester enrollees and second semester enrollees) were imported into RapidMiner's Local Repository. Set Role function is used to identify the id special attribute from the two data sets. Set Minus operator was applied to determine those records of the first data set whose IDs are not contained within the second set. The result will be tagged as "Yes" in the newly-generated attribute "Attrited" using the Generate Attribute operator. Intersect operator, on the other hand, returns those records of the first data set whose IDs are contained within the second data set. The result is the retained students so this will be tagged as "No" in the Attrited attribute.

The two data sets with the Attrited column was joined together using the Union operator. The outcome is a single flat file that was eventualy cleansed using Handle Missing Values and Filter Examples operators as shown in Figure 3.Set Role operator is used again to define the label attribute which is the Attrited column. Using Discretize by Entropy, the average grade was discretized (converted to nominal type) into 4 bins (0-26, 26-49, 49-77, and 77-100). The final data set to be fed into the modelling process is 3557 out of 4093 records.

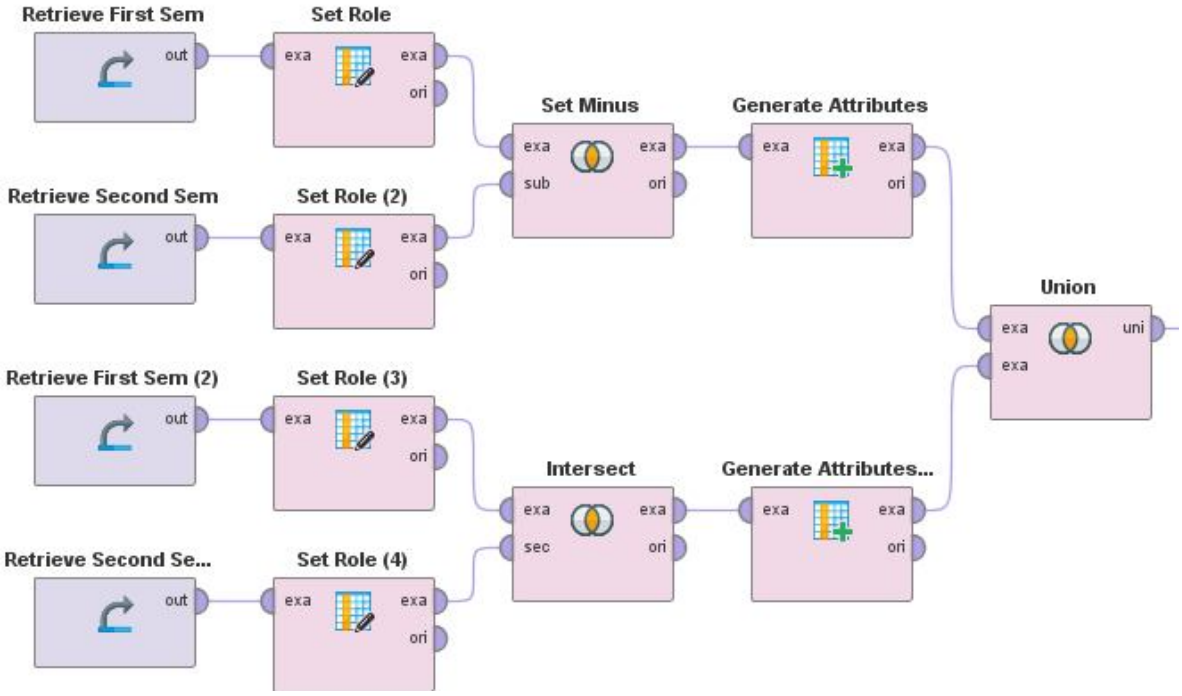


Figure 2 Data Preparation Process using RapidMiner

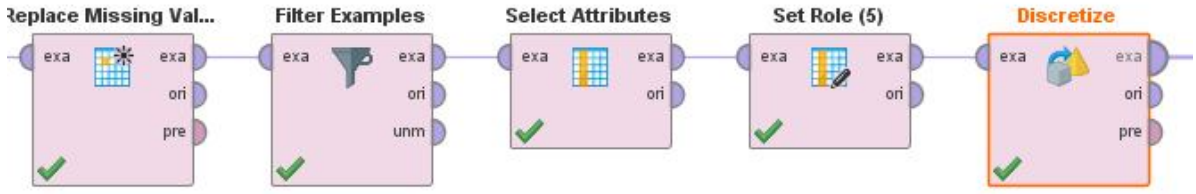


Figure 3RapidMiner Operators used for Data Preparation

Phase Four: Modeling

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models. [15]

Three modeling techniques were selected for this study namely: Decision Tree, Naïve Bayes, and Rule Induction as shown in Figure 4. These models were selected because of their popularity and superiority based on many studies.

Decision tree is commonly used for gaining information for the purpose of decision -making. It starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. [8]

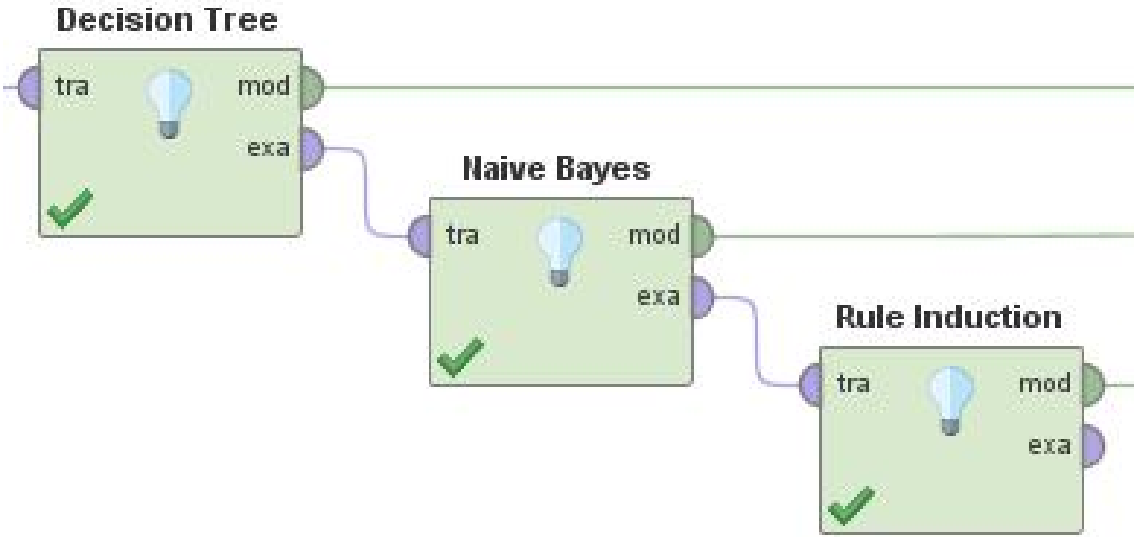


Figure 4 Predictive Models used in this Study

Naive Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence, as observed in the data. In simple terms, it assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other features.

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data. [16]

Phase Five: Evaluation

Inthis step, the three models were compared for their predictive accuracy using the Cross Validation, Apply Model and Performance operators as shown in Figure 5 and 6.

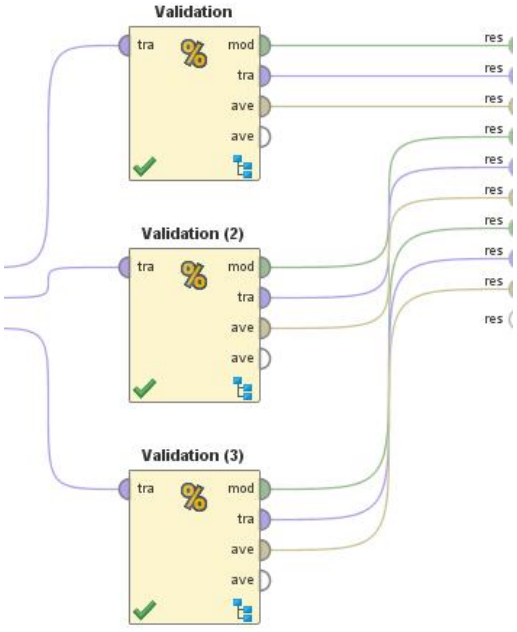


Figure 5 Validation Operators for the three models

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. [17]

The Apply Model operator applies an already learnt or trained model on the testing data set. The Performance operator is used for performance evaluation. It delivers a list of performance criteria values like accuracy.

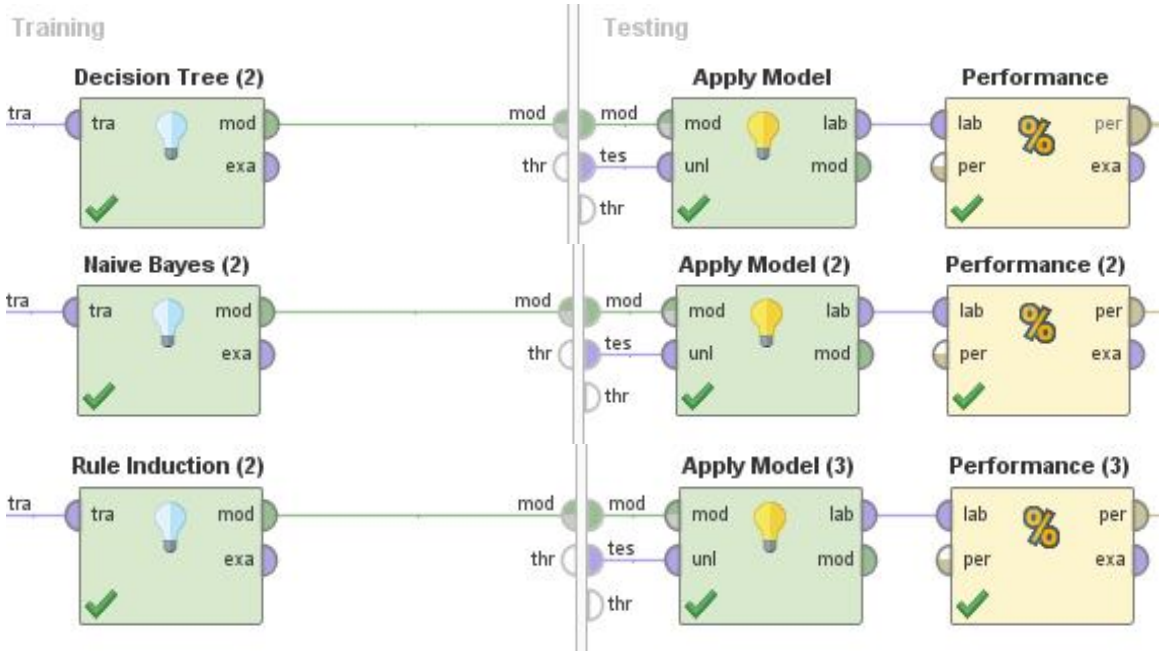


Figure 6 Apply Model and performance Operators

RESULTS AND DISCUSSION

Results show that, given sufficient data and appropriate variables, data mining methods are capable of predicting freshmen attrition with roughly 80% accuracy as shown in Table 4. Among the three models, Naïve Bayes performed the best with 83.41% accuracy, followed by Rule Induction with 82.94%, and Decision Tree with 80.79%.

Table 4 Prediction Results

Predictive Model	Accuracy
Naïve Bayes	83.50%
Decision Tree	82.96%
Rule Induction	82.65%

Decision Tree Results

Based on the decision graph and tree on Figure 6 and 7 respectively, this study confirms that academic performance during the first semester is an important predictor for student attrition.[14] Those students whose average grade is between 0% and 25% are the most at-risk freshmen to leave. Meanwhile, gender hardly contributes to student attrition. However, female students are more likely to attrite than male students if the grade average belongs to range 1 and 2.

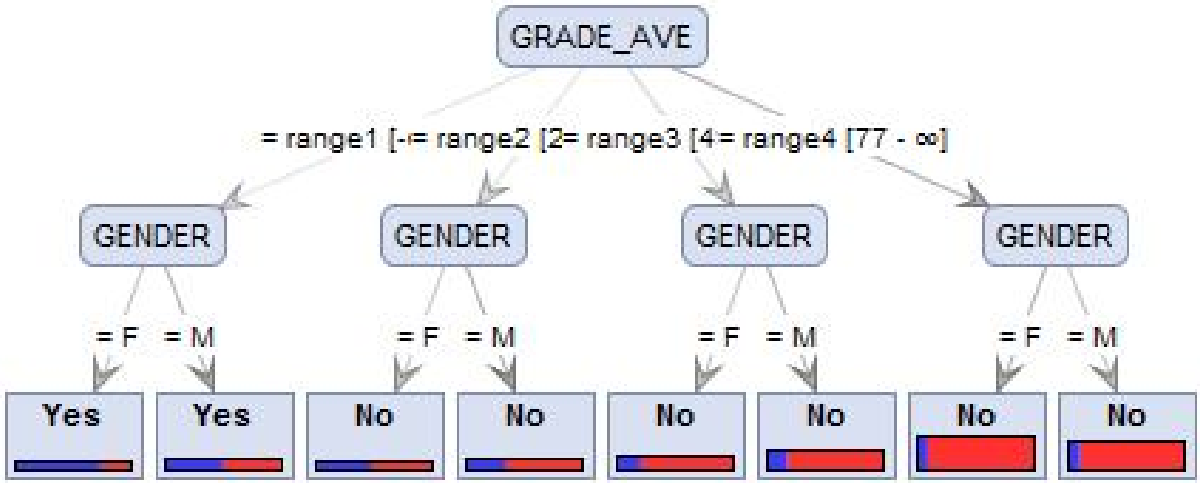


Figure 7 Decision Graph

```
GRADE_AVE = range1 [-∞ - 26]
|   GENDER = F: Yes {Yes=63, No=21}
|   GENDER = M: Yes {Yes=68, No=66}
GRADE_AVE = range2 [26 - 49]
|   GENDER = F: No {Yes=39, No=41}
|   GENDER = M: No {Yes=60, No=116}
GRADE_AVE = range3 [49 - 77]
|   GENDER = F: No {Yes=68, No=246}
|   GENDER = M: No {Yes=98, No=427}
GRADE_AVE = range4 [77 - ∞]
|   GENDER = F: No {Yes=121, No=1138}
|   GENDER = M: No {Yes=114, No=871}
```

Figure 8 Decision Tree

Naive Bayes Result

The distribution of student attrition according to average grade and gender are shownin Figures 9 and 10.

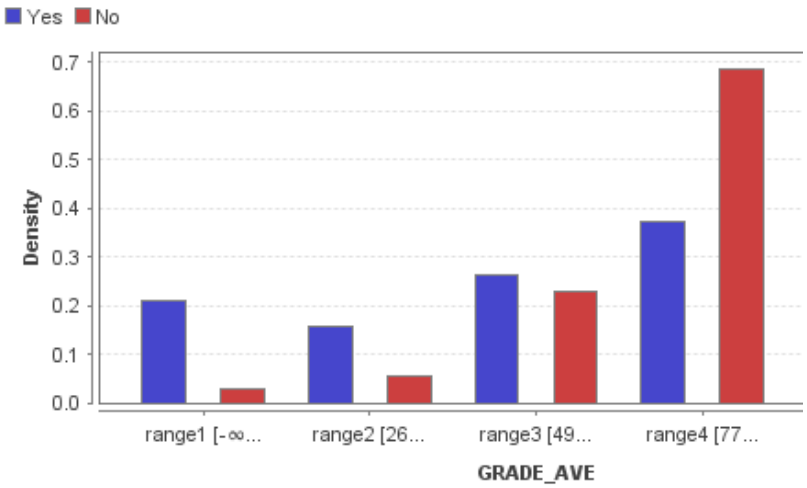


Figure 9 Student Attrition Distribution according to Grade Average

Although the highest number of student attrition falls under range 4 (77-100%), there are significant differencesbetween the "Yes" and "No" predictions in ranges 1 and 2 which means that students are more likely to attrite if their average grade in the first semester belongs to ranges 1, 2 and 3. Moreover, students under range 4 are more likely to stay and enroll for the second semester.

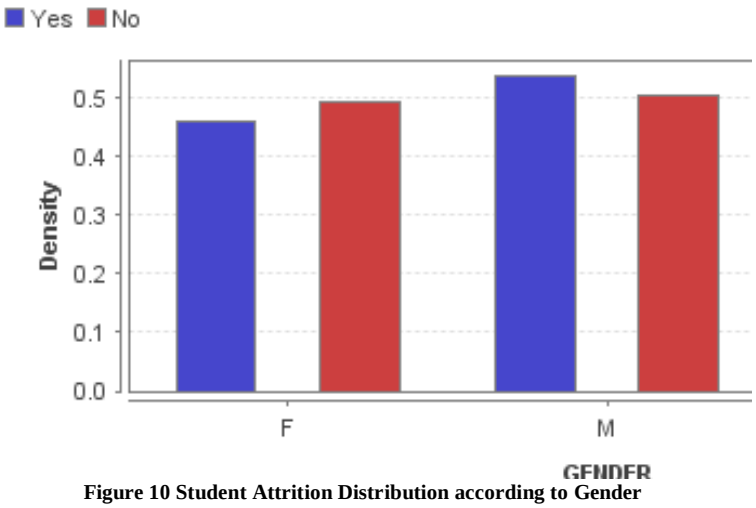


Figure 10 Student Attrition Distribution according to Gender

Based on the student attrition distribution, there are more male students who attrite than female students, in general.

Rule Induction Result

```
if GRADE_AVE = range4 [77 - ∞] and GENDER = F then No  (121 / 1138)
if GRADE_AVE = range4 [77 - ∞] and GENDER = M then No  (114 / 871)
if GRADE_AVE = range3 [49 - 77] and GENDER = M then No  (98 / 427)
if GRADE_AVE = range3 [49 - 77] and GENDER = F then No  (68 / 246)
if GRADE_AVE = range2 [26 - 49] and GENDER = M then No  (60 / 116)
if GRADE_AVE = range1 [-∞ - 26] and GENDER = F then Yes  (63 / 21)
if GRADE_AVE = range1 [-∞ - 26] and GENDER = M then Yes  (68 / 66)
if GRADE_AVE = range2 [26 - 49] then No  (39 / 41)

correct: 2970 out of 3557 training examples.
```

Figure 11 Rule Induction Result

Rule induction confirms the result of decision tree as shown in Figure 11.

CONCLUSION

This study has showed that academic performance can be used as predictor in determining student attrition and retention. As presented, if the student average grade is failing, he is at-risk of leaving the institution. Compared to the average grade attribute, gender, on the other hand, did not manifest significant results. Based from the outcomes of the three models, student attrition can be predicted regardless of the student gender, although male students have higher probability of attrition in general.

REFERENCES

[1] C. Demetriou und A. Schmitz-Sciborski, „Integration, Motivation, Strengths and Optimism: Retention TheoriesPast, Present, and Future," in 7th national Symposium on Student Retention, Charleston, SC, 2011.

[2] M. Villar, „College Education For Poor Students," science.ph, 5 December 2012. [Online]. Available: http://www.science.ph/full_story.php?key=7556:college-education-for-poor-students&type=latest. [Zugriff am 5 July 2017].

[3] S. Herzog, „Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-à-Vis Regression," in Data Mining to Improve Enrolment Management, Wiley Periodicals, 2006, pp. 17-33.

[4] L. Kalsner, „Issues in College Student Retention," Higher Education Extension Service Review, Bd. 3, Nr. 1, 1991.

[5] L. Caison, „Analysis of Institutionally Specific Retention Research: A Comparison Between Survey And Institutional Database Methods," Springer, Bd. 4, Nr. 48, pp. 435-449, 2007.

[6] D. Calbimonte, „Data Mining Introduction," SQL ServerCenter.com, 20 January 2017. [Online]. Available: http://www.sqlservercentral.com/articles/server/94235/. [Zugriff am 7 July 2017].

[7] M. M. Ali, „Role of Data Mining in Education Sector," International Journal of Computer Science and Mobile Computing, Bd. 2, Nr. 4, pp. 374-383, 2013.

[8] K. Surjeet und P. Saurabh, „Data Mining Application in Enrolment Management: A Case Study," International Journal Of Computer Application (IJCA) , Bd. 41, Nr. 5, pp. 1-6, 2012.

[9] L. Grebennikov und M. Shah, „Investigating Attrition Trends in Order to Improve Student Retention," Quality Assurance in Education, Bd. 20, Nr. 3, pp. 223-236, 2012.

[10] E. Yukselturk, S. Ozekes und Y. K. Turel, „Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program," European Journal of Open, Distance and e-Learning, Bd. 17, Nr. 1, 2014.

[11] R. K. Arora und D. Badal, „Predicting Students Attrition using Data Mining," International Journal of Computer Science & Engineering Technology, Bd. 4, Nr. 2229-3345, pp. 1338-1341, 2013.

[12] L. G. Moseley und D. M. Mead, „Predicting Who Will Drop Out of Nursing Courses: A Machine Learning Exercise," Elsevier Inc., Bd. 28, Nr. 4, p. 469–475, 2008.

[13] S. A. K. J. Rai, „Student's Dropout Risk Assessment in Undergraduate Course at Residential University," IJCA International Journal of Computer Applications, Bd. 84, Nr. 0975-8887, 2013.

[14] D. Delen, „Predicting Student Attrition with Data Mining Methods," Baywood Publishing Co., Inc., Bd. 13, Nr. 1, pp. 17-35, 2011.

[15] C. Shearer, „The CRISP-DM Model:The New Blueprint for Data Mining," Hournal of Data Warehousing, Bd. 5, Nr. 4, pp. 13-22, 2000.

[16] Tzu-Liang, C.-C. Huang, Y.-N. Fan und C.-H. Lee, „Quality Control Using Agent Based Framework," in Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 2015, p. 13.

[17] J. Vanschoren, J. N. v. Rijn, B. Bischl und L. Torgo, „OpenML: Networked Science in Machine Learning 10-fold Cross Validation," SIGKDD Explorations, Bd. 15, Nr. 2, pp. 49-60, 2013.

[18] L. Aulck, N. Velagapudi, J. Blumenstock und J. West, „Predicting Student Dropout in Higher Education," arXiv preprint, p. 1606.06364, 2016.

[19] M. J. Bani und M. Haji, „College Student Retention: When Do We Losing Them?," in Proceedings of the World Congress on Engineering and Computer Science 2017, 2017.

★★★