# Towards Freshman Retention Prediction: A Comparative Study

Admir Djulovic and Dan Li

*Abstract*—**The objective of this research is to employ data mining tools and techniques on student enrollment data to predict student retention among freshman student populations. In particular, the goal is to identify freshman students who are more likely to drop out of school so that preemptive actions can be taken by the university. Through data analysis, we identify the most relevant enrollment, performance, and financial variables to construct learning models for retention prediction. The experiments have been conducted using Decision Trees, Naïve Bayes, Neural Networks, and Rule Induction models. These models have been compared and evaluated extensively. Our findings show that each model has its advantages and disadvantages and among all the input variables, students' GPA and their financial status have bigger impact on students' retention than other variables.**

*Index Terms*—**Classification, feature selection, freshman retention, prediction.**

## I. Introduction

Data mining tools and techniques have been extensively used in the private and business sectors but not so much in higher education [1]. Just recently, universities have recognized the power of such technology that they are now starting to invest time and resources into it. They are especially interested in exploring data mining power to help them improve student retention rates.

The motivation for this research is to find the most common factors that influence students to stay or leave the university. Most importantly, the goal is to identify potential financial, academic, and/or personal reasons that cause students to drop out of school. From university's perspective, it is very costly and time consuming to bring new students into the system. Therefore, the student retention and student academic success are top priorities for universities. On the other hand, the top priorities for students and their parents are to get into a good school and successfully fulfill their academic goals as quickly as possible. One way of achieving both students' and universities' goals is to provide the means to identify the at-risk student populations as early as possible, so that the institutions can employ additional resources to help student succeed [2]. This is where the data mining tools and techniques become useful.

However, there are many challenges related to the mining of enrollment data from the point of data collection all the way to model creation and deployment. This paper will address the issues and challenges encountered during the entire research process. The rest of this paper is organized as follows: Section II describes the current research work related to the prediction of student retention using data mining techniques; Section III discusses the main mining steps including data collection, feature selection, data preprocessing, and predictive model construction; The experimental results and analysis are provided in Section IV; Finally, concluding remarks along with directions for future improvements are presented in Section V.

Manuscript received April 10, 2013; revised June 19, 2013.

The authors are with the Computer Science Department, Eastern Washington University, Cheney, WA 99004 USA (e-mail: adjulovic@eagles.ewu.edu, danl@ewu.edu).

## II. Related Work

Data mining techniques have been commonly used in many areas including business, health, science and engineering, etc. However, it has been pointed out that data mining techniques have not been widely used in higher education, especially when it comes to the improvement of student retention [1]. To address this issue, the authors in [1] have used three years of data collected from the first year degree-seeking students to develop prediction models. The models are generated using several data mining algorithms including decision trees, neural networks, ensemble, and logistic regression. The authors have selected decision tree model for their final implementation due to its higher prediction accuracy, its ability to better handle missing data, and its intuitive representation of knowledge.

The authors in [3] have used three decision tree algorithms (ID3, C4.5, and ADT) to predict student retention probabilities. They have presented acceptable precision rate ranging between 68.2% and 82.8%. However, the recall rates range between 6.4% and 11.4%. This low recall range indicates that most positive cases have been misclassified by their prediction systems.

Eitel J.M. Lauria *et al.* have presented preliminary experimental results on the development of initial retention prediction model using several data mining algorithms [4]. Their preliminary findings indicate that both the logistic regression and the Support Vector Machine (SVM) algorithms considerably outperform the C4.5 decision tree in terms of their ability to detect students at academic risk.

While some of the research on retention prediction has focused on comparing and testing different classification models, there are other studies focusing on identifying crucial student attrition/retention factors. Rather than merely focusing on the analysis of students' academic standings, Chong Ho Yu *et al.* have examined other factors that affect student retention from sophomore to junior year using decision trees, multivariate adaptive regression splines (MARS), and neural networks [5]. Interestingly, the authors have found that among many potential predictors, transferred hours, residency, and ethnicity are three crucial factors affecting student retention rate.

DOI: 10.7763/IJIET.2013.V3.324

James N. Wetzel *et al.* have also focused their work on identifying key factors affecting student retention using logistic regression functions [6]. They have found that academic progress drives the attrition/retention decision largely, and student social integration also plays an important role in persistence decision. Among all the factors, financial considerations appear to be minor in importance.

The authors in [7] have primarily focused their research on the prediction of student retention in engineering programs. This is motivated by the fact of lower student enrollment in engineering programs and higher demand in industry for engineers. The authors have found the factors that cause high student attrition in their engineering programs. Based on their findings, students who are placed on the first term probation are likely to leave before they graduate. Interestingly, they have also observed that students who are placed in second term probation are even more likely to leave the program. This points out that students' pre-college education readiness could have significant influence on their collage success.

In this research, we will explore various data mining techniques to identify most important academic, personal, and financial factors that impact students' attrition/retention decisions at our university. The research in [3] shows very low recall values when decision tree approach is used to predict student retention rate. We will address this concern and evaluate decision tree approach using different number of input attributes. Besides decision tree model, we will also explore other predictive modeling approaches including Naïve Bayesian, neural networks, and rule induction, and evaluate these approaches extensively under different experimental settings.

## III. Methodologies

Fig. 1 shows the major components of our system, which includes two main branches. One branch is used to pre-process the training data and build different predictive learning models. The second branch focuses on the pre-processing of the unseen test data and the application of different learning models to generate comparable prediction results. The data pre-processing in both branches consists of data collection, feature selection, missing data handling, outlier removal, and data transformation.

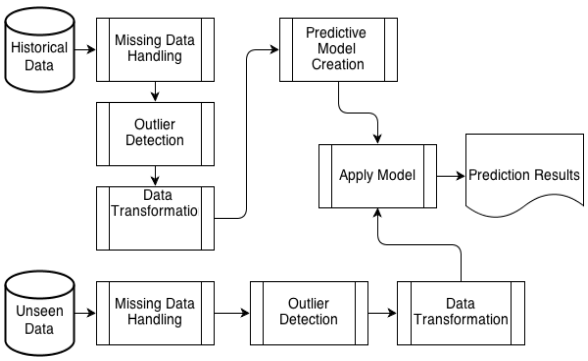



Fig. 1. System architecture.

### A. Data Collection and Pre-Processing

To conduct the research, the freshman enrollment data have been collected from 2006 to 2012 academic years. As suggested in [1], [2], [8], students' pre-college academic standings, gender, and residency status play important roles in the prediction of student retention. Therefore, we have included SAT scores, high-school GPA, gender, and living on/off campus information into our data set. In addition, we add two more attributes to the data set, financial aid status and the amount of balance due, because we want to identify the potential relationships between a student's financial status and his/her retention status. All of these attributes serve as the initial input/independent variables to build our predictive learning models. In addition, we use attribute RETAINED to denote the dependent or target variable which is set to 1 if a student is retained; otherwise it would be 0. Below is a list of variables being used in this study and their explanations:

Pre-enrollment variables include:

1) AGE: Student Age at the beginning of the academic year
2) GENDER: F(female), M(male), N(not disclosed)
3) PREV_ED_GPA: High school GPA
4) **RETAINED** (target variable): Student retained next year (0: No, 1: Yes)
5) SAT_READING: Student SAT score
6) SAT_MATH Student SAT score
7) SAT_WRITING: Student SAT score

Fall/Winter/Spring term-specific variables include:

1) FALL/WINTER/SPRING_BAL: Student term-specific financial balance
2) FALL/WINTER/SPRING_CUMULATIVE_GPA: Student cumulative GPA
3) FALL/WINTER/SPRING_GPA: Student term-specific GPA
4) FALL/WINTER/SPRING_LIVING_ON_CAMPUS: Term-specific living on campus status (0: No, 1: Yes)
5) FALL/WINTER/SPRING_RECEIVED_FINAID: Term-specific financial aid status (0: No, 1: Yes)

Note that we have used accumulated attributes for analysis. This means the analysis for Fall-term will use students' pre-enrollment data plus all the Fall-term related attributes. Similarly, the analysis for Winter-term will use pre-enrollment, Fall, and Winter related attributes, and the analysis for Spring term will use pre-enrollment, Fall, Winter, and Spring related attributes.

Among 7800 training records, 12% of them have missing values, and most of the missing values come from the fields of SAT scores and high-school GPA. To avoid biased analysis, these instances have been removed from the data set. In addition, outliers have also been identified and removed using distance-based clustering approach. To generate more condensed classification models, numerical attributes including GPA, SAT, AGE, and BAL have been discretized into categorical attributes based on domain knowledge. For example, the GPA number schema has been converted to the letter grading scale, and the attributes related to financial balance have been converted to categorical values with specific balance ranges.

### B. Identifying Important Retention Factors

Even though there exist research papers [1], [5], [6] discussing important factors affecting student retention probabilities, in this study, we would like to conduct our own research to identify the most important factors impacting freshman retention status at our institution. Four statistical

methods are adopted to determine the importance of each independent variable. These methods include Chi-squared test, information gain, gain ratio, and correlation analysis using local polynomial regression.

TABLE I: NORMALIZED WEIGHTS OF INDEPENDENT VARIABLES

| Variables | **Information Gain** | **Chi Squared** | **Correlation** | **Gain Ratio** |
|---|---|---|---|---|
| FALL_LIVING_ON_CAMPUS | 0 | 0 | 0 | 0 |
| GENDER | 0.001 | 0.001 | 0.024 | 0.001 |
| AGE | 0.001 | 0.001 | 0.029 | 0.001 |
| SAT_READING | 0.010 | 0.009 | 0.096 | 0.012 |
| SAT_MATH | 0.013 | 0.012 | 0.102 | 0.016 |
| SAT_WRITING | 0.015 | 0.014 | 0.119 | 0.019 |
| WINTER_LIVING_ON_CAMPUS | 0.018 | 0.017 | 0.132 | 0.024 |
| FALL_RECEIVED_FINAID | 0.020 | 0.019 | 0.142 | 0.036 |
| FALL_BAL | 0.029 | 0.029 | 0.169 | **1** |
| WINTER_BAL | 0.043 | 0.043 | 0.206 | **0.942** |
| PREV_ED_GPA | 0.065 | 0.064 | 0.259 | 0.077 |
| SPRING_BAL | 0.065 | 0.067 | 0.241 | **0.937** |
| SPRING_LIVING_ON_CAMPUS | 0.070 | 0.080 | 0.290 | 0.112 |
| WINTER_RECEIVED_FINAID | 0.103 | 0.102 | 0.328 | 0.172 |
| SPRING_RECEIVED_FINAID | 0.276 | 0.276 | **0.540** | 0.431 |
| FALL_GPA | 0.387 | 0.399 | **0.621** | 0.276 |
| FALL_CUMULATIVE_GPA | 0.389 | 0.402 | **0.620** | 0.281 |
| WINTER_CUMULATIVE_GPA | **0.585** | **0.600** | **0.729** | 0.440 |
| SPRING_CUMULATIVE_GPA | **0.711** | **0.715** | **0.786** | **0.538** |
| WINTER_GPA | **0.745** | **0.763** | **0.864** | 0.476 |
| SPRING_GPA | **1** | **1** | **1** | **0.605** |

Table I shows the normalized weight of each input variable generated by the above four methods and the weights higher than 0.5 are highlighted. From Table I, we have the following observations:

1) Chi-squared analysis and information gain generate the exact same ordering of input attributes with little variation in numerical values. The ordering of attributes generated from correlation analysis with local polynomial regression is almost the same as Chi-squared test and information gain but with larger variation in numerical values.
2) The results from information gain, Chi-squared, and correlation analysis indicate that students' first-year academic performance, especially their performance in Winter and Spring terms (represented by TERM_CUMULATIVE_GPA and TERM_GPA) is one of the key factors impacting freshmen's retention status.
3) Different from what have been found in [5], our study indicates that students' residency status (represented by LIVING_ON_CAMPUS) is not an important factor affecting students' retention status.
4) Different from what have been found in [1] [8], our study indicates that gender, age, and students' pre-college academic standings (represented by SAT and PRE_ED_GPA) contribute very little to students' retention probabilities.
5) In general, Spring-term attributes are more important than Winter-term attributes, and Winter-term attributes are more important than Fall-term and Pre-enrollment attributes. This suggests that helping students succeed in the last term of their first academic year could potentially improve university's freshman retention rate.
6) The ordering of attributes by information gain ratio indicates that financial balance (represented by TERM_BAL) could potentially be an important factor impacting freshman retention. Note that the gain ratio measure is to remove the potential biases of information gain measure when there are too many outcome values of an independent attribute. This finding suggests us to further evaluate the impact of students' financial situations. We will address this further in later sections.

### C. Classification Models and Their Settings

One major goal of this research is to develop and evaluate multiple classification models for the prediction of freshman retention. In this section, we will introduce four learning models we have constructed, their settings, and some of the results from each model.

#### 1) C4.5 decision trees

Among many decision tree approaches, we use WEKA's C4.5 algorithm [9] to build a binary decision tree. This algorithm uses pessimistic pruning to remove unnecessary branches to improve the accuracy of prediction and we set the confidence threshold to 0.25 for pruning. We generate this predictive model using 10-cross validation with stratified sampling to maintain the original data distribution.

```
SPRING_CUMULATIVE_GPA = D: 0 (96.0/6.0)
SPRING_CUMULATIVE_GPA != D
|   SPRING_CUMULATIVE_GPA = F: 0 (10.0)
|   SPRING_CUMULATIVE_GPA != F
|   |   SPRING_CUMULATIVE_GPA = C
|   |   |   FALL_GPA = F: 1 (16.0/3.0)
|   |   |   FALL_GPA != F
|   |   |   |   SPRING_BAL = $0.00 or less
|   |   |   |   |   SPRING_GPA = D
|   |   |   |   |   |   WINTER_GPA = A: 1 (3.0)
|   |   |   |   |   |   WINTER_GPA != A
|   |   |   |   |   |   |   WINTER_GPA = B: 1 (43.0/13.0)
|   |   |   |   |   |   |   WINTER_GPA != B
|   |   |   |   |   |   |   |   SAT_CRITICAL_READING_VERBAL = Low: 0 (9.0/1.0)
|   |   |   |   |   |   |   |   SAT_CRITICAL_READING_VERBAL != Low
|   |   |   |   |   |   |   |   |   FALL_CUMULATIVE_GPA = C: 0 (14.0/2.0)
|   |   |   |   |   |   |   |   |   FALL_CUMULATIVE_GPA != C
|   |   |   |   |   |   |   |   |   |   WINTER_GPA = C
|   |   |   |   |   |   |   |   |   |   |   SAT_CRITICAL_READING_VERBAL = Medium
|   |   |   |   |   |   |   |   |   |   |   |   SAT_WRITING = Low: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   SAT_WRITING != Low
|   |   |   |   |   |   |   |   |   |   |   |   |   GENDER = F: 1 (8.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   GENDER != F: 0 (8.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   SAT_CRITICAL_READING_VERBAL != Medium: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   WINTER_GPA != C: 0 (21.0/6.0)
```

Fig. 2. Decision tree by C4.5 algorithm.

Fig. 2 shows a portion of the tree starting from the root node. From this snapshot we can see that among many input variables, GPA-related variables (represented by TERM_CUMULATIVE_GPA and TERM_GPA) are the ones being selected early in the decision tree. This indicates that GPA-related variables are more important in determining the label of target variable RETAINED. In addition,

SPRING_BAL representing the amount of financial balance due is also one of the top selected attributes. These findings are consistent with the observations we have discussed in the previous section.

*2) Naïve bayes*

The second classification model we have constructed is the Naïve Bayes model which is a probability model based on the assumption that all the input variables are independent from each other. Even though this assumption does not completely hold on our data set, our correlation analysis shows that only a few variable pairs are highly correlated. For instance, FALL_CUMULATIVE_GPA and FALL_GPA have the highest normalized correlation value of 0.955, and WINTER and SPRING LIVING_ON_CAMPUS have the second highest correlation value of 0.813. This correlation analysis result suggests us to remove highly correlated and redundant variables from the data set for efficiency considerations.

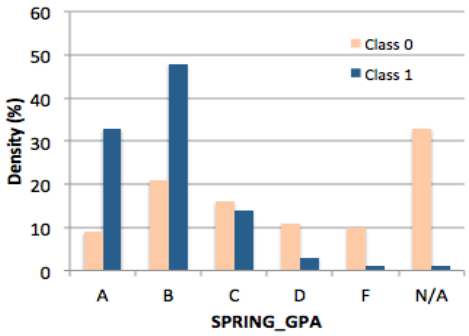



Fig. 3. SPRING_GPA *vs.* class distribution.

Fig. 3 shows the relationship between SPRING_GPA and the likelihood of class label generated by the Naïve Bayes model. Note that the class attribute RETAINED is a binary attribute and the value of 1 denotes a retained case. Fig. 3 shows the strong impact of SPRING_GPA on students' retention decisions. When SPRING_GPA is either A or B, there are more retained students than drop-out students. However, as SPRING_GPA gets lower into C, D, F or N/A (this denotes the cases when students do not receive any valid grades in Spring-term), the proportion of drop-out students to retained students increases significantly.

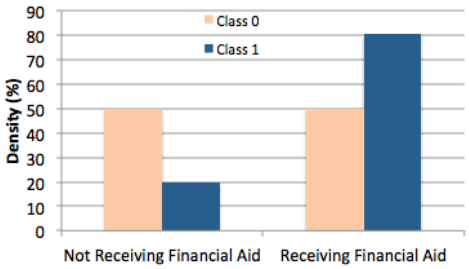



Fig. 4. SPRING_RECEIVED_FINAID *vs.* class distribution.

As mentioned earlier, we are particularly interested in identifying the impact of a student's financial status on his/her retention decision. Fig. 4 shows the impact of financial aid attribute. Among all the students who have received financial aid in Spring-term, the proportion of retained freshmen to not-retained freshmen is about 1:0.6. However, among all the students who have not received financial aid in Spring-term, the proportion of retained freshmen to not-retained freshmen is about 1:2.5. This result suggests that the university should investigate financial aid policy and take preemptive actions to help at-risk students stay in the university.

*3) Neural networks*

The third predictive learning model we have constructed is the Neural Network model. Since this model cannot handle polynomial and binomial values, we have transformed all the attributes into numerical values. Furthermore, we have identified the optimized settings for the Neural Network model. Accordingly, the network has one hidden layer with 13 nodes and the number of training cycles is set to 500, the learning rate is 1.0, the momentum is set to 0.74000001, and the error epsilon is 1.0E-5.

*4) Rule induction*

The last learning model we have investigated is the rule induction model using Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm [10]. The reason of choosing rule induction model is because rules are intuitive and relatively easy for people to understand. RIPPER starts with the less prevalent classes and the algorithm iteratively grows and prunes rules until there are no positive examples left or the error rate is greater than 50%. In the growing phase, for each rule greedily conditions are added to the rule until the rule has 100% accuracy. The procedure tries every possible value of each attribute and selects the condition with the highest information gain.

Fig. 5 list all the IF-THEN rules generated from the rule induction model. The numbers inside parenthesis are the number of negative (not-retained) cases versus the number of positive (retained) cases covered by each rule. Surprising, there are only a total of 14 rules and these 14 rules correctly cover 4268 out of 5104 training examples. From these rules we can see again, students' academic performance in terms of GPA is a key factor affecting students' retention decisions because there are 9 out of 14 rules having GPA as the antecedent or a portion of the antecedent.

```
if SPRING_RECEIVED_FINAID = 1 then 1  (660 / 3100)
if SPRING_GPA = B then 1  (54 / 332)
if SPRING_GPA = N/A then 0  (429 / 39)
if SPRING_GPA = A then 1  (28 / 172)
if SPRING_GPA = C and WINTER_CUMULATIVE_GPA = B then 1  (23 / 80)
if SPRING_CUMULATIVE_GPA = D then 0  (50 / 2)
if SPRING_GPA = F and FALL_LIVING_ON_CAMPUS = 1 then 0  (19 / 3)
if AGE = 19-22 and GENDER = F then 1  (9 / 26)
if WINTER_RECEIVED_FINAID = 1 then 0  (8 / 0)
if SAT_WRITING = Low then 1  (2 / 12)
if WINTER_GPA = D then 0  (9 / 2)
if SAT_MATHEMATICS = High then 1  (2 / 8)
if WINTER_GPA = C then 0  (13 / 5)
if WINTER_CUMULATIVE_GPA = A then 1  (0 / 3)
else 1  (7 / 7)
```

Fig. 5. IF-THEN rules.

## IV. Experimental Results

As suggested in [1], the decision tree model generates the increased performance as the number of input attributes for analysis increases. Therefore, we conduct our experiments in the same manner by gradually adding more independent attributes to our learning models. We start our experiments by using pre-enrollment and Fall-term attributes only. Then we add Winter-term attributes to the data set. Finally, Spring-term attributes are added to the data set. We test all four learning models in each case and the performance is evaluated using the five performance metrics defined below:

TABLE II: COMPARISON OF FOUR LEARNING MODELS

| Model | Overall Accuracy | Positive Precision | Positive Recall | Negative Precision | Negative Recall |
|---|---|---|---|---|---|
| Pre-enrollment + Fall-term Attributes | | | | | |
| C4.5 Decision Tree | 81.07 | 82.29 | 96.75 | 66.29 | 23.51 |
| Naïve Bayes | 77.58 | 84.50 | 87.53 | 47.25 | 41.04 |
| Neural Networks | 81.24 | 82.14 | 97.29 | 69.14 | 22.31 |
| Rule Induction | 80.73 | 81.69 | 97.29 | 66.69 | 19.92 |
| Pre-enrollment + Fall-term + Winter-term Attributes | | | | | |
| C4.5 Decision Tree | 83.12 | 85.28 | 94.90 | 68.03 | 39.84 |
| Naïve Bayes | 79.97 | 87.06 | 87.53 | 53.25 | 52.19 |
| Neural Networks | 78.09 | 87.49 | 84.16 | 48.95 | 55.78 |
| Rule Induction | 82.86 | 85.38 | 94.36 | 66.23 | 40.64 |
| Pre-enrollment + Fall-term + Winter-term + Spring-term Attributes | | | | | |
| C4.5 Decision Tree | 85.76 | 87.94 | 94.90 | 73.60 | 52.19 |
| Naïve Bayes | 80.48 | 88.29 | 86.66 | 54.10 | 57.77 |
| Neural Networks | 86.02 | 85.89 | 98.37 | 87.18 | 40.64 |
| Rule Induction | 86.27 | 85.96 | 98.81 | 90.18 | 40.24 |

$$Overall\ Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negaive + False\ Negative}$$

$$Positive\ Precision = \frac{True\ Positive}{True\ Positive + False\ Positive};$$

$$Positive\ Recall = \frac{True\ Positive}{True\ Positive + False\ Negative};$$

$$Negative\ Precision = \frac{True\ Negative}{True\ Negative + False\ Negative};$$

$$Negative\ Recall = \frac{True\ Negative}{True\ Negative + False\ Positive}.$$

From Table II, we have the following observations:

1) **The effect of increased number of input variables**: As more input variables being added to the data set, all four learning models demonstrate improved performance. For instance, the overall accuracy using rule induction model has increased from 80.73% to 86.27% when the full set of attributes is being used for analysis. Similarly, the negative precision has jumped from 66.69% to 90.18% using rule induction model. This observation is consistent with the conclusion drawn in [1]. Therefore, the best setting for our research is to use all the available attributes for the prediction of retention rate.
2) **The comparison of four learning models**: One of the initial goals of this project is to develop multiple classification models for retention prediction and then choose the best model for system deployment. Now, the question is: among the four learning models we have presented in this paper, which one should be selected as the final winner? Unfortunately, based on Table II, we cannot easily answer this question because the performance of these four models varies with regard to different performance metrics. If the overall accuracy of the prediction system is the major consideration, we can use any one of the top three models, i.e., C4.5 decision trees, neural networks, and rule induction, because these three models provide the overall accuracy of 86% when the complete set of independent variables is used. If the prediction accuracy for positive (i.e., retained) instances is the major concern, then the Naïve Bayes model slightly outperforms the other three models because the Naïve Bayes model has the positive precision of 88.3% which is the highest one among the four models. If the goal is to identify as many positive cases as possible, then the rule induction model is the winner because it has the highest positive recall value of 98.81% among the four learning models. If the prediction accuracy for negative (i.e., not-retained) instances is the major consideration, then again, the rule induction model is the winner because it has the highest negative precision of 90.18%, which is much higher than other three models. Finally, if the goal is to recognize as many negative cases as possible, then the Naïve Bayes model should be used because it generates the highest negative recall of 57.77%. Therefore, our conclusion is: we cannot simply say one model is better than another one because we need to take different performance metrics into consideration.
3) **Comparing with other related work**: Now we would like to compare our learning models with the models presented in other research papers. The authors in [3] have used three decision tree methods to predict the probability of students' retention. They show the highest accuracy of 74.4% with C4.5 decision tree algorithm and the highest precision of 82.8% and the highest recall of 11.4% with adaptive decision tree (ADT) algorithm. As mentioned earlier, with a low recall rate of 11.4%, the system can barely be useful because the model has misclassified most of the positive instances. In comparison, our predictive models correctly recognize 57.77% drop-out students and 98.81 % of retained students. In addition, the authors in [1] have used decision trees, logistic regression, neural networks, and ensemble models to predict freshman retention. They show the highest overall accuracy of 80%, the highest negative precision of 78%, and the highest negative recall of 52%. Again, based on Table II, our learning models outperform theirs regarding all the performance metrics.

After presenting the results from our learning models, now we would like to further examine the impacts of individual variables on the prediction of student retention. During our experiments we have observed that there is a certain level of correlation between students' financial balance and their retention status. The relation between the target variable RETAINED and the independent variable SPRING_BAL is shown in Fig. 6. The x-axis is our target variable RETAINED

and the y-axis is the Spring-term financial balance due variable SPRING_BAL. It is obvious that the students who have SPRING_BAL greater than zero are more likely to withdraw from school than those students who do not have balance due.

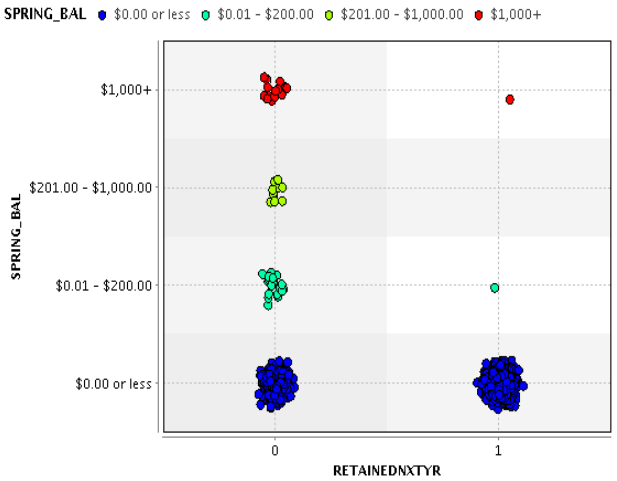



Fig. 6. SPRING_BAL *vs.* RETAINED.

Interestingly, we have also observed the relationship between SPRING_GPA and SPRING_BAL. As shown in Fig. 7, the students who have financial balance great than zero in Spring are more likely to have a SPRING_GPA lower than C. This implies that there is a certain degree of correlation between students' financial situation and their academic performance, and consequently, students' academic performance impacts students' retention status.

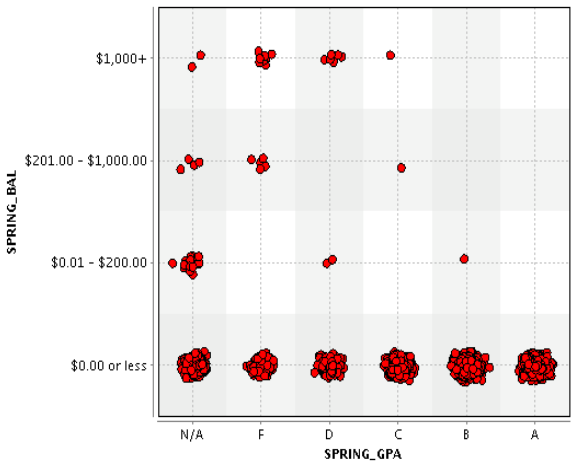



Fig. 7. SPRING_BAL *vs.* SPRING_GPA.

## V. Conclusions and Future Work

Educational data mining has played an increasingly important role recently. This research is to identify the most important factors impacting retention decisions and develop multiple predictive classification models for retention prediction. We have created and analyzed four predictive models including decision trees, Naïve Bayes, neural networks, and rule induction models.

Among all the independent input attributes, the attributes related to first-year academic performance contribute most to retention status. In addition, students' financial aid status and financial balance due also have impact on students' retention decisions. Our study also shows that the more independent attributes we use in analysis, the most accurate the system could be.

Comparing all four learning models, the rule induction model has the highest overall accuracy, and the IF-THEN rules generated from the model are easily understandable by users. The Naïve Bayes model presents the lowest accuracy among the four models. However, when the goal is to identify as many at-risk students as possible, the Naïve Bayes model is the winner because it has the highest negative recall rate. We also compare our predictive models with other models and demonstrate the improvements we have obtained with regard to all the performance metrics we have defined.

Even though the resulting classification models show the overall good prediction results, there is certainly room for improvement. For instance, additional personal background attributes such as parents educational background, high school rankings, and first-generation college status could be added to pre-enrollment attribute set to make better predictions. Similarly, if we add more attributes such as the number of credits taken, the number of class withdrawals, and student credit overload indicator to the term-based attribute sets, improvements on the prediction accuracy could be expected. Since most students in our data set are retained students, this implies an imbalanced data distribution on class attribute. This may consequently affect the precision and recall for negative (not-retained) cases. It is worth further investigation to find better classification models in handling imbalanced data sets.

## Acknowledgment

We would like to give our thanks to Eastern Washington University and Institutional Research Department for their support and making this research possible. Special thanks to Toni Habegger and Dennis Wilson for ongoing support.

## References

[1] M. Bogard, T. Helbig, G. Huff, and C. James, "A comparison of empirical models for predicting student retention," *Tech report*, Western Kentucky University, 2011.

[2] Z. J. Kovacic, "Predicting student success by mining enrolment data," *Research in Higher Education Journal*, vol. 15, pp. 1-20, March 2012.

[3] S. K. Yadav, B. Bharadwaj, and S. Pal, "Mining education data to predict student's retention: a comparative study," *International Journal of Computer Science and Information Security*, vol. 10, no. 2, pp. 113-117, 2012.

[4] E. J. M. Lauria, J. D. Baron, M. Devireddy, V. Sandararaju, and S. M. Jayaprakash, "Mining academic data to improve college student retention: an open source perspective," in *Proc. the 2nd International Conference on Learning Analytics and Knowledge*, April 2012, Vancouver, Canada, pp. 139-142.

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

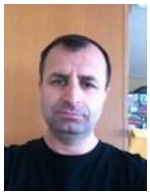
[6] J. N. Wetzel, D. O'Toole, and S. Peterson, "Factors affecting student retention probabilities: a case study," *Journal of Economics and Finance*, vol. 23, no. 1, pp. 45-55. Spring 1999.

[7] A. Scalise, M. Besterfield-Sacre, L. Shuman, and H. Wolfe, "First term probation: models for identifying high risk students," in *Proc. the 30th Annual Conference on Frontiers in Education*, pp. F1F/11-FiF/16, 2000.

[8] R. Alkhasawneh and R. Hobson, "Modeling student retention in science and engineering disciplines using neural networks," in *Proc. the 2011 IEEE Conference on Global Engineering Education*, April 2011, pp. 660-663.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10-18, 2009.

[10] W. W. Cohen, "Fast effective rule induction," in *Proc. the 12th International Conference on Machine Learning*, 1995, pp. 115-123.



**Admir Djulovic** was born in 1975 in Bosnia and Herzegovina. He received his B.S. degree in computer science from Eastern Washington University (EWU) in 2005. He is currently a Master student in Computer Science at EWU and he is planning to graduate in Fall 2013. His research interests include educational data mining, multivariate time series analysis, and biomedical data analysis.

While pursuing his master degree in computer science, Admir is also working full time as Information Technology Specialist IV at EWU. As a senior-level specialist in an assigned area of responsibility and as a team and project leader, Admir applies advanced technical knowledge and considerable discretion to evaluate and resolve complex tasks such as planning and directing large-scale projects, conducting capacity planning, designing multiple-server systems, directing or facilitating the installation of complex systems, hardware, software, application interfaces, or applications, developing and implementing quality assurance testing and performance monitoring, planning, administering, and coordinating organization-wide information technology training, and developing security policies and standards.

**Dan Li** was born in 1973 in China. She received her B.E. degree in computer science from Harbin Engineering University in 1996, received her master degree in computer science from Shenyang Institute of Computing Technology, Chinese Academy of Sciences in 1999, and received her Ph.D. degree in Computer Science from University of Nebraska - Lincoln, in 2005. She is currently an Assistant Professor in Computer Science Department at Eastern Washington University. Her current research interests include large-scale databases, spatio-temporal data mining, educational data mining, information security, and computer science education.