An Interpretable Predictive Framework for Students' Withdrawal Problem Using Multiple Classifiers

Nashat T. Al-Jallad, Xu Ning, Mergani A. Khairalla

Abstract— **Students' withdrawal problem is one of the main concentration of enrollment management at educational institutions as it negatively affects their performance and reputation. This paper discusses two types of students' withdrawals which includes long-term dropout and the short-term dropout and considers this problem as a multi-class classification problem rather than a binary classification problem. We first introduce a novel (RG*) method to generate ruleset using multiple rules learning classifiers including Decision Trees and Rule induction methods to improve the accuracy and interpretability of the classification. Then we propose a predictive framework based on the RG* to predict at-risk students and to address students' data problems such as imbalanced and high-dimensionality. Two groups of criteria are used to evaluate the proposed framework including: model performance and interpretability. The results revealed the possibility of a tradeoff between the performance and interpretability of the classification outputs through exploiting the ability of the multiple classifiers. In addition, the proposed framework shows a significant improvement in predicting both dropout and stopout students' compared with using individual classifiers.**

Index Terms— **Student withdrawal, Rule-learning method, Enrollment Management, Rule interpretability**

I. Introduction

The higher education institutions work in very competitive environments that create an urgent need to analyze students' data in order to make more informed decisions and come up with plans and strategies especially with regard to the students' enrollment problems [1].

Students' withdrawal problem has attracted an increased attention of universities due to the fact that one-third of students leave without receiving their certificates [2]. This problem results in wasting students' time, financial resources and self-confidence [3]. Several studies have distinguished two types of students' withdrawal behavior. The short-term stopout, which refers to the case of a temporary interruption of study[4, 5] and the long-term stopout, which reflects the case of total interruption of study. For example, Horn in [6] examined the ability to distinguish between these two types based on students' characteristics. She finds that there are several factors affects a student's enrollment decision such as the student's age and admission type. Stratton et.al. in [4] investigated the impact of the financial aid on student decisions. They find that the factors associated with stopout behavior are different from those associated with dropout behavior. Therefore, assuming the withdrawal problem is a binary classification problem (continue, and dropout) may lead to misleading results and inaccurate targeting of at-risk students (who may give up the study). Accordingly, the withdrawal problem should be considered as multi-class decision problem includes: continue, dropout and stopout.

that have addressed the problem of student retirement can be classified as: statistical-based approaches [6], and the machine learning based approach [7]. Under the statistical-based approach, the multiple linear regression, multinomial logit model [4], multilevel history analysis [8], propensity score matching[9], and logistical regression [10] are used frequently. On the other hand, several machine learning techniques (i.e., classification techniques) are used to determine the influencing factors that shape student decisions of whether to continue their studies or withdraw such as Decision Trees (DTs), Induction methods (IM), Artificial Neural Network (ANN), Support Vector Machine (SVM), Bayesian Net, and Random Forest. It is worth mentioning that the main objective of those studies is offering the universities a proactive system that enables them to make informed decisions with the assistance of the collaborative applications or recommendation systems [7, 11].

The classification techniques are divided into two groups: The Black-Box classifiers (such as ANN and SVM, and ensemble learning models) which cannot be interpreted their results, and the White-Box classifiers (such as Decision Trees and rule induction methods) or Rule-Based (RB) classifiers which produce interpretable rulesets in form of tress structure or IF-THEN, but these classifiers are not compete well as the Black-Box classifiers (in term of the accuracy of predictability) [12, 13].

Therefore, improving a white-box classifier to achieve a right tradeoff between performance and interpretability is our aim in this study. Thus, to improve the prediction of withdrawal of students using the data of the freshman students (the first-year students), and to provide a powerful

Manuscript submitted December 16, 2017; revised May 18, 2018. This work was supported by the Chinese Scholarship Council, China, Wuhan, under Grants CSC N0. 2015- 7360122.

Nashat. T. Al-Jallad, is with the Computer Science and Technology, Wuhan University of Technology, Wuhan, P.R. China 430070, Doctoral student. (corresponding author to provide phone: 15623897720, e-mail: jallad@whut.edu.cn). Also, he is with Computer Science School; Palestine Technical University, Tulkarem, Palestine, lecturer. (email: n.jallad@ptuk.edu.ps).

Xu-Ning is with Wuhan University of Technology, Wuhan, P. R. China 430070, Professor. (e-mail: xuning@whut.edu.cn).

Mergani A. Khairalla is with Wuhan University of Technology, Wuhan, P. R. China 430070, Doctoral student (mirgani2008@gmail.com.com).

predictive tool for the university directors.

Although, there is almost a consensus that using multiple classifiers (such as Ensemble Learning) can improve the learning results [14], most of these methods focused on the performance criteria (such as accuracy) and ignores interpretability criteria (such as rule-size and rules-overlap) which is important for many real-world applications.

The main contributions of this work are summarized as follows: First, we propose a novel method of combining the results of multiple classifiers with the aim of improving the accuracy of the model and the interpretability of the model results. Second, this paper proposes an interpretable framework based on the proposed model to predict different types for students' withdrawal decisions and address set of problems that students' data suffer from. Third, according to our knowledge, none of the prior studies disused the two types of withdrawal problems using data mining methods as addressed in this paper. Finally, although students' enrollment problems are the main concern of many academic institutions around the world, few studies have been conducted in Palestine with respect to predicting the enrollment behavior of students, thus this study comes to fill this gap in the academic literature.

The remainder of this paper is organized as follows: The second section elaborates on details of the proposed method and framework. Section three demonstrates the experimental framework setup and the details of the datasets and classifiers used. Section three provides experimental results and the corresponding discussions. Finally, the conclusion and recommendations for further research are explored.

II. Methodology

The proposed framework aims to produce a rule-set in form of IF-THEN because this form is more interpretable for users (e.g. managers' and instructors') who do not have experience in machine learning techniques. To achieve this goal, we propose RG* method based on multiple white-boxes classifiers includes Induction methods (IM) and Decision Trees (DTs'). Then, use a novel way of combining and filtering the rules that produced by individual classifiers to create a more accurate and interpretable subset. Figure 1, illustrates the framework components which are described as follows:

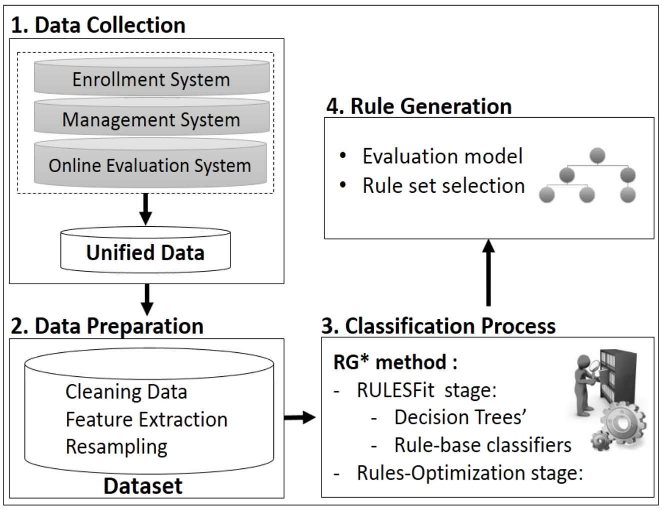


Fig 1. Interpretable predictive framework

1) Data Collection

This step aims to prepare students' data by collecting them from different sources in order to remove incompatibility and inconsistency. Thus convert it to an appropriate format in which to be used for the next stage of the framework. It is worth mentioning that researches in the field of education provided a basic understanding of the common student attributes that can be used, such as what being done by Bonaldo et al.[10]. Section three of this study includes the description of students' data and features used.

2) Data Preparation

The preparation process is performed to make data appropriate for classifiers. Data integration and cleaning processes are applied to remove outliers and incomplete data. In this context, we distinguish between data entry errors and unavailability of data due to student withdrawal. As the missing-data with respect to withdrawals are significant. Therefore, we choose to replace it with a zero value. In case that the subsequent records of the student are non-zero, he is regarded as stopout, whereas, if his records are of zero value he is regarded as a dropout.

Frequently, the students' data encompass a large number of characteristics with lead to increased dimensionality that affect the performance of classifiers. Therefore, the need to reduce the attributes size should be taken into consideration when applying an algorithm for the prediction in order to increase both its efficiency and effectiveness. On the other hand, the imbalanced data is a common problem in this kind of application domain. Therefore, we examined the presence of an imbalanced problem, then we used the SMOTE method as a one of the common algorithm used in this context to solve this problem before feeding data to classifiers.

3) Classification Process and the Description of RG*

The aim of the classification process is to predict unknown class label $Y \in \{y_1, y_2, \cdots, y_m\}$ on an object using known values of attributes $X = (x_1, x_2, \cdots, x_n)$. This prediction done by constructing a classification function $f(x)$ that predicts accurately the value of Y. If the class is correctly predicted, the classification function is not penalized, otherwise the unit penalty will be imposed. Most of RB classifier generate rules in form of IF-THEN or tree structure (such as IM and DTs' respectively). Due to it is simplicity of IF-THEN format we focus on it in the current study. IF-THEN rule consist of two part the (condition and class-label part of the rule). Let Ψ is the condition parts of the rule and Ψ (x) is an indicator function equal to 1 if x satisfies the condition part Ψ and equals 0 otherwise. The response of the rule is then defined as a vector of $\lambda \in \mathbb{R}^m \rightarrow \{0,1\}$ as defined by Ψ. Therefore, the decision function is defined as in the following Equation 1:

$$r(x) = \lambda \Psi (x) \qquad (1)$$

Let ruleset $\mathcal{R}_k$ is a collection of rules $r = \{r_1, r_2, \cdots, r_m\}$ that generated by classifier k where k $\in K$ which is the classifiers pool, and $\mathcal{R}$ is the collection of rulesets that are generated by these classifiers, where $\mathcal{R} \in \{\mathcal{R}_i, \mathcal{R}_{i+1}, \cdots, \mathcal{R}_K\}$ and k is a number of classifiers.

Although the RB classifiers aim to generate a rule that minimizes the loss function of prediction, the right tradeoff between prediction performance and model interpretability is still the subject of research.

To this end, we propose the RG* method based on the principle of the combination strategies [15] using multiple classifiers to find a small and simple set of subset of rules that lead to good generalization by getting the benefit from multiple RB classifiers. The RG* method consists of the RULESFit and Rules-Optimization stages. These stages are:

First, at the RULESFit stage, it is assumed that the classification function is a linear function that obtain conditional function from n rules:

$$f(\mathrm{x}) = \alpha + \sum_{i=1}^{n} r_i\ (x) \tag{2}$$

where α is a default rule covering whole attribute space. The max value obtain by $f(\mathrm{x})$ denotes the number of instances covered by the ruleset generated by a classifier.

Second, at Rules-Optimization, as shown in Algorithm 1, we evaluate the generated ruleset by define a weight function based on a both of interpretability ($\mathbb{I}$) and performance ($\mathbb{P}$) sub functions:

$$\ell = \mathbb{I}(\mathrm{r}).\mathbb{P}\ (\mathrm{r}) \tag{3}$$

Where $\mathbb{I}$ is a linear function denotes the interpretability of a rule, and $\mathbb{P}$ is a performance function that denotes the probability of correctly classifying each class-label.

Rulesets were generated using heterogeneous classifiers (including decision trees and rule-based algorithms), each of which produces ruleset weighted by ℓ function. They were then analyzed by the performance of the classifier using novel criteria that focuses on the accuracy and interpretably of the prediction results. Set of experiments were conducted to answer following questions:

--How well do decision trees and rule-based algorithms perform when they are evaluated using the traditional evaluation metrics such as recall?

--How well do decision trees and rule-based algorithms perform when they are evaluated using the Interpretability metrics such as rules size and rules overlap?

It is worth mentioning that, using algorithms that are able to produce rules in the form of Tree or IF-Then are more interpretable for end-users (e.g., educational directors). Moreover, these algorithms differ in the method used for generating their rule set.

Finally, we identify the "best" n rulesets that produce the most accurate and interpretable rules (in our experiments we assume $N = 2$). Therefore, the pool of rules includes all rules generated by the top two classifiers.

The objective is to maximize ℓ using rules from multiple rulesets (classifiers).

$$L = arg_{max} \sum_{i=1}^{N} \sum_{j=1}^{n} \mathbb{I}(\mathrm{r}).\mathbb{P}\ (\mathrm{r}) \tag{4}$$

Where N denotes the number of rulesets (we assume N =2) and n is the number of rules in the ruleset. For more details about interpretability and performance measurement, see the Evaluation section.

Next section discusses a brief presentation of the algorithms used in this study.

4) Evaluation Classifiers

This stage is concerned with the process of evaluating classifiers to obtain the appropriate classifiers for Optimization stage of RG* method. In order to avoid overfitting problem, the dataset is divided into two datasets to avoid data overfitting problem.

```
Algorithm 1 RG* method (
   A be the ruleset for best interpretable classifier,
   B be the ruleset for best model performance classifier,
   $K$ be the number of rules generated by the second-ranked interpretable classifier,
   $E$ examples in training set)
1. Let $R = \{\}$ be the initial rule set
2. $R \leftarrow A$
3. Remove all instances from $E$ that are covered by $A(r)$
4. For each class $C$ in B
   4.1. $A(C)_s$ be the Sensitivity of class $C$ in ruleset A
   4.2. $B(C)_s$ be the Sensitivity of class $C$ in ruleset B
   4.3. if $B(C)_s \leq A(C)_s$ remove all rules related to $C$ from B
5. Order the rules in $B$ ascending based on class sensitivity range in $A(C)$
   5.1. While not meeting the stopping criteria
      5.1.1. if $Evaluation(R) > Evaluation(R + \{B(r)\})$
         1. Add rule to ruleset: $R = R + \{B(r)\}$
      5.1.2. remove instance from $E$ that are covered by $B(r)$
```

The training dataset that includes 70% of students' data instances and the testing dataset that includes 30% of students' data instances. Two groups of measurements are used includes model performance and model interpretability measurements. The next section describes these two groups.

5) Rule generation

The set of rules are output from RG* method. This ruleset has the greatest ability to balance the model interpretability and performance characteristics [22]. Therefore, these rules are easy to understand by educational directors and can be easily used for developing an educational recommendation system.

III. Experimental Framework

In this section, we present the setup of the experimental Framework including, firstly, the details of the datasets are described; Secondly, the description of the RB classifiers is presented; Thirdly, the evaluation measurements are explained. Finally, the statistical tests are applied to compare the results obtained by experiments.

1) Datasets

In this section, the datasets are divided into two group. First, is the public datasets which used to evaluate the proposed RG* method. Second, is the students' data which used to evaluate the proposed framework. Next, is the description of these two groups of datasets.

-- The public datasets: seven datasets are used for evaluating the RG* method after applying SMOTE resampling method. These datasets are downloaded from the UCI Machine Learning repository. The description of these datasets is illustrated in Table 1.

-- The students' dataset: the data of 721 "Banking & Financial Management" freshman students enrolled during the academic year 2010-2016 in Palestine Technical University (PTUK) are used in this work. Three different

TABLE I
DESCRIPTION OF DATASET USED FOR THE SECOND EXPERIMENT

	#Attribute	# Classes	# Instances	Type
Glass	8	6	453	Numeric
Soybean	36	19	4838	Nominal
Page Blocks	11	5	24562	Numeric
Win	14	3	178	Nominal
Splice	62	3	4947	Nominal
Auto	26	6	159	Nominal
Contr[5]	10	3	1473	Nominal

sources of data have been used: First: The Enrolment System provides information about a social, economic, personal data and prior performance related to students. Second, the University Management System provides information about students' performance progress during their study semesters. Finally, the Online Evaluation System questionnaire provides information for three indicators that are used in this study, and it includes evaluation of the university, lectures, and courses. Table 2 and Figure A.1 illustrates 68 student features that are collected from the aforementioned sources. For more details about student attributes and data sources, see our forthcoming paper[16].

Since our focus is on the freshman students, the investigation period is divided into three sub-periods (P1, P2, P3). P1 and P2 cover the periods following the midterm exams of the first and second semesters respectively. While P3 is the last period which describes the students' status after the final exam in the second semester as shown in Figure 2.

TABLE II
DATA SOURCES AND STUDENT FEATURES USED

Source	Enrollment System	Management System	Online Evaluation System
Information Type	Demographical, Historical performance, Socio-economical.	Registration per semester, Student Performance.	University parameters, Lecturers variables, Courses variables.
Number of Attributes	22	16	30

2) The RB classifiers

As mentioned early, in this subsection we aim to describe the RB classifiers used for producing rulesets and for comparison it with our RG* method.

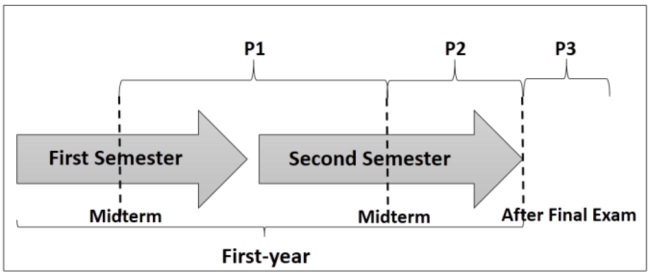


Fig 2. Steps in which data gathered

Firstly, Decision Tree (DT) Techniques is one of the popular classification techniques that classify a target attribute in a form of tree structure. The rule-set generated by the DT can be converted into the form of IF-THEN which is more interpretable [17, 18]. Three decision trees are used in this study including:

-- Decision tree (C4.5) which is the most common decision algorithms. This algorithm is an extension of the ID3 algorithm, and it generates an initial set of rules using the direct method. The C4.5 algorithm applies the process of normalization of information gain ratio which calculated using Equation 5.

$$GianRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{5}$$

$Gain(A)$ denotes to the information gain for attribute A, and $SplitInfo(A)$ denotes to the splitting information value. The splitting information value represents the probable information generated by splitting the dataset D into set of P partitions, corresponding to P outcomes on attributes A. The splitting information calculated using Equation 6.

$$SplitInfo_A(D) = \sum_{i=1}^{P} \frac{|D_i|}{|D|} \times \, log_2 \, (\frac{|D_i|}{|D|}) \tag{6}$$

Where $|D|$ denote the number of elements in the training dataset D, and $|D_i|$ denotes the number of elements into partition P.

-- Combining Decision Trees (CDT) uses the imprecise Dirichlet model and uncertainty of measure or credal sets as criteria for branching in order to reduce the complexity of the generated tree. The total uncertainty is calculated as:

$$TU\,(p) = IG(p) + GG(p) \tag{7}$$

where IG is the measurement of non-specify, and GG denote randomness function for credal sets.

-- Multi-class Alternating Decision tree (LADTree), which combines decision trees and Logit-Boost to produce a set of classification rules. In addition, it deals with the multi-class problem by splitting it into several binary-class problems.

Secondly, Rule-based classifiers include rule learner classifiers that generate interpretable rule-set in the form of IF-THEN [19, 20]. Four types of rule-based classifiers are investigated in this component including:

-- Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is based on associations rules. In the learning process, RIPPER orders a training dataset according to class frequencies then select one class as a default prediction. After generating rules set of that class, all instances covered by these rules are extracted from the dataset. This process is repeated for other classes. This algorithm uses direct method for generating the initial rule-set.

-- PART is an algorithm based on decision tree C4.5. It is a relatively a simple algorithm based on divide and conquer strategy that generates a tree in each iteration in order to produce a set of rules called "decision lists".

-- RIpple-DOwn Rule learner (RIDOR) starts by generating a default rule, then finding the exceptions to this rule in order to find the "best" exception with the least error rate. Those exceptions denote the list of alternative rules that can be used to predict classes other than the default one.

-- Fuzzy Unordered Rule Induction Algorithm (FURIA) is an extension of the RIPPER algorithm where it learns from fuzzy rules to generates unordered rule sets. Moreover, it makes use of an efficient rule stretching method to deal with uncovered examples.

3) Evaluation measurements:

Next is the description of these two groups of measurements includes Model Performance (MP) and Model Interpretability (MI). The MP measurements describe how well a classifier can predict students' withdrawal decision correctly. In this context, three metrics are used:

--True Negative rate or Specificity ($TNrate$) is the ratio to the negative prediction number that is truly classified as negative and the actual number of the negative class. It is calculated using Equation 8:

$$TNrate = \ (TN)/(FP + TN) \tag{8}$$

--True Positive rate, Recall or Sensitivity ($TPrate$) denotes to the preparation of positive predictions that correctly identified by classifier. It is calculated using Equation 9:

$$TPrate = \ TP/(TP + FN) \tag{9}$$

--Accuracy (Acc) which denote the ratio of correct prediction to the total number of instances evaluated, in other words, this measure specifies the degree of success for a classifier. It is calculated using Equation 10:

$$Acc = \ (TP + TN)/(TP + TN + FP + FN) \tag{10}$$

The MI measurements evaluate the simplicity of the rulesets that generated by classifiers for the end-user. Either these rulesets are generated in the form of IF-Then (such as in the Induction methods) or in the form of the tree structure (such as in the Decision trees techniques). It is worth mentioning that the tree structure can be converted to Form IF-Then easily. The following is a brief description of a ruleset.

Let $\mathcal{R}$ is the collection of rule sets that are generated by classifiers, where $\mathcal{R} \in \{\mathcal{R}_i, \mathcal{R}_{i+1}, \cdots, \mathcal{R}_N\}$ and N is number of classifiers. Each ruleset includes set of rules $r = \{r_1, r_2, \cdots, r_m\}$, $r \in \mathcal{R}$, and m is number of rules in $\mathcal{R}_N$. In addition, $r = (\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ is conditional part of the rule and $\mathcal{Y}$ is the class-label of that rule, where r is represented in a form of $r : Cond \leftarrow Class$. In addition, $Cond$ denotes to the conditional part, while $Class$ denotes to the class label given to the dataset examples that satisfy the $Cond$ part.

We adapt the interpretability assumptions that are proposed in [21] to identify matrices that are used to measure the model interpretability.

Assumption 1: Whenever a rule set includes less number of rules, it becomes more understandable.

--Size($\mathcal{R}$) denotes the number of rules contained in a rule set. Based on assumption 1, the "best" rule set is the one that contains less number of rules.

Assumption 2: Whenever a rule includes less number of predictors (conditions), the rule becomes easier for end-users.

-- Length ($L(r)$) denotes to the number of predicates used in the rules of a rule set divided by the maximum number of predicators can be used in that rule set.

-- Average Length ($AL(\mathcal{R}_i)$) : is the rate of the rule set length based on the overall generated rule sets, which is calculated using Equation 11:

$$AL(\mathcal{R}_i) = |\mathcal{R}|^{-1} \cdot \sum_{r \in \mathcal{R}_i} L(r) \tag{11}$$

Where $|\mathcal{R}|$ denotes maximum rules length achieved by the classifiers in the experiment. This matric assumes maximum value of 1.0 for a "worst" rule set based on its length.

Assumption 3: whenever the rule-set consists of less number of overlapping predicates, it becomes easy for a human to understand the relation between predicates and class labels.

-- Cover ($Co(\mathcal{X}, \mathcal{Y})$) is the set of data points in the training dataset that satisfy rule $\mathcal{X}$ in a rule set.

-- Overlap ($O(r, \acute{r})$) : let $r = (\mathcal{X}, \mathcal{Y})$ and $\acute{r} = (\acute{\mathcal{X}}, \acute{\mathcal{Y}})$, we consider r overlapped with $\acute{r}$ if there is any data point that satisfy with $\mathcal{X}$ and $\acute{\mathcal{X}}$ at the same time as the shown in Equation 12:

$$O\ (r, \acute{r}) = Co(\mathcal{X}, \mathcal{Y}) \cap Co\ (\acute{\mathcal{X}}, \acute{\mathcal{Y}}) \tag{12}$$

-- Overlap Fraction ($OF(\mathcal{R})$) represents the overlap between every pair of rules in the rule set, where a small value means more interpretability. Based on assumption 3, we favor a rule set with less overlapping. The overlap fraction is calculated using Equation 13:

$$OF(\mathcal{R}_i) = \frac{2}{|\mathcal{R}| \cdot (|\mathcal{R}| - 1)} \sum_{r_i, r_j \in \mathcal{R}_i\, r_i \neq r_j} (O((r_i, r_j)) \cdot N^{-1} \tag{13}$$

4) Statistical tests

Statistical tests have been used in this section to observe the significant difference within the studying methods. We adopt nonparametric hypothesis testing techniques that recommended by Detta et.al.,[22]. Specifically, we employed Friedman aligned rank test. The Wilcoxon signed rank test is used to perform the pairwise comparison [23].

IV. Results and Discussions

In this section, we developed extensive experiments to analyze the proposed RG* method using real-world datasets and then applied statistical tests to compare the results obtained by these experiments. Next, we analyze the proposed framework to address students' withdrawal problem. Finally, we explore the Framework output (generate rules) which improve the students' withdrawal prediction and can use by both an educational director and to improve recommendation systems.

A. Analyzing the performance of the proposed RG method*

Firstly, we examined the effectiveness of the proposed method (RG*) using real-world datasets in order to answer this question:

Is it possible to obtain a rules-set that is more interpretable and accurate by combining the results from different classifiers?

Therefore, we compared RG* with two state-of-the-art rule mining methods: RIPPER and FURIA.

As shown in Table 3, the overall ruleset size generated by FURIA algorithm is greater than those generated by each RIPPER and RG* algorithms.

In addition, the RG* improved the accuracy values for all datasets compared with the accuracy achieved by the RIPPER algorithm, while the interpretability measurements are still "better" than that achieved by the FURIA algorithm. For example, the accuracy achieved by applying the RIPPER algorithm on the "Glass" dataset is (79.69%). The accuracy value increased to (94.52%) after applying the RG* method. Although there is a decrease in interpretability, the ruleset

TABLE III
DESCRIPTION OF DATASET USED FOR THE SECOND EXPERIMENT

	DB[1]	Sen[2]	Spec[3]	Acc.	Size (Ri)	AL (Ri)%	OF (Ri)
FURIA	Glass	0.920	0.854	85.43	21	0.846	0.002
	Soybean	0.970	0.969	96.90	41	0.056	0.002
	PB[4]	0.996	0.987	98.72	95	0.554	0.003
	Win	0.949	0.974	94.94	7	0.038	0.002
	Splice	0.963	0.964	96.42	125	0.064	0.003
	Auto	0.786	0.931	78.61	16	0.401	0.002
	Contr[5]	0.546	0.731	54.58	10	0.605	0.003
RIPPER	Glass	0.782	0.797	79.69	13	0.741	0.002
	Soybean	0.966	0.966	96.62	34	0.084	0.002
	PB[4]	0.995	0.983	98.33	47	0.909	0.002
	Win	0.921	0.951	92.13	4	0.560	0.002
	Splice	0.782	0.797	79.69	13	0.741	0.002
	Auto	0.767	0.922	76.72	13	0.772	0.002
	Contr[4]	0.519	0.687	51.86	6	0.882	0.002
RG*	Glass	0.855	0.825	81.48	15	0.710	0.002
	Soybean	0.958	0.966	96.65	35	0.068	0.002
	PB[4]	0.995	0.984	98.35	50	0.680	0.002
	Win	0.932	0.955	93.04	6	0.128	0.002
	Splice	0.863	0.884	92.42	78	0.398	0.002
	Auto	0.779	0.928	77.04	14	0.624	0.002
	Contr[4]	0.538	0.692	52.89	7	0.781	0.002

1: Databases; 2: Specificity, 3: Sensitivity; 4: Page Blocks; 5: contraceptive

generated by RG* remains better than FURIA ruleset.

Hence, we cannot have obtained meaningful conclusion without applying the proper statistical analysis. The Wilcoxon test was performed to compare the proposed RG* method with RIPPER and FURIA methods. The completed statistical results are illustrated in Table 4. Based on these results, we can find that RG* method is significantly better RIPPER and FURIA since the correspondent p-value is less than (5%). Although there are no significant differences between RG* and RIPPER in term of model interpretability, we can highlight the effectiveness of the proposed RG* method.

TABLE IV
THE RESULTS OF WILICOXON TEST FOR COMPARING RG* WITH JRIP AND FURIA CLASSIFIERS

Measurement	Comparison	Hypothesis	p-value
Sen[1]	RG* vs. RIPPER	**Rejected for RG* at 5%**	0.046399*
	RG* vs. FURIA	**Rejected for RG* at 5%**	0.017756*
Spec[2]	RG* vs. RIPPER	**Rejected for RG* at 5%**	0.027708*
	RG* vs. FURIA	**Rejected for RG* at 5%**	0.017148*
Acc.	RG* vs. RIPPER	**Rejected for RG* at 5%**	0.017960*
	RG* vs. FURIA	**Rejected for RG* at 5%**	0.017960*
Size (Ri)	RG* vs. RIPPER	Not rejected	0.232508
	RG* vs. FURIA	**Rejected for RG* at 5%**	0.027281*
AL (Ri)%	RG* vs. RIPPER	Not rejected	0.498962
	RG* vs. FURIA	**Rejected for RG* at 5%**	0.027992*
OF (Ri)	RG* vs. RIPPER	Not rejected	1.000000
	RG* vs. FURIA	**Rejected for RG* at 5%**	0.045500*

A “*” mean there is statistical difference with $\alpha = 0.05$
1: Specificity, 2: Sensitivity

B. Analyzing the performance of the proposed Framework for student withdrawal problem

As shown previously, the RG* method improves the predictability of the model while saving an acceptable level of interpretability by adding selective rules that generated the multiple RB classifiers. Therefore, in this section, we evaluate the proposed framework uses the RG* method to predict students’ enrolment decisions. At first experiment, we evaluate seven RB classifier in order to identify the “best” classifiers for Optimization step of RG* method. Then at the second experiment, we investigate the results obtained by the framework with those resulting from the first experiment.

--Experiment I:

as shown in Table 5, under the decision tree techniques using the performance metrics for (P1), the C4.5 algorithm achieve the best value with respect to accuracy (86.16%). In addition, the rules-based methods for the same period (P1) show that the RIDOR algorithm obtains the lowest value with respect to accuracy (84.4%). Moreover, the FURIA achieve the best value with respect to accuracy (87.22%), not only within the same group but also in comparison with all other algorithms. However, this superiority of FURIA is not preserved when looking at the interpretability matrices. To be more specific, the FURIA generates eighteen rules with (51%) of the average overlap value. However, the RIPPER algorithm achieves the best values with respect to rule size (13), and overlap (3%).

TABLE V
EVALUATION METRICS FOR CLASSIFIERS IN THE THREE PERIODS (P1, P2, P3)

	Model performance metrics				Interpretability metrics		
	TP rate	TN rate	Recall	Acc.	Size (Ri)	AL (Ri)%	OF (Ri)
P1							
C4.5	0.91	0.95	0.91	86.16	57	1.00	0.00
CDTree	0.90	0.96	0.90	84.88	51	0.38	0.00
LADTree	0.88	0.97	0.88	85.15	31	0.35	0.00
RIPPER	0.92	0.96	0.92	85.25	13	0.31	0.03
RART	0.90	0.94	0.90	86.36	35	0.44	0.02
RIDOR	0.93	0.93	0.93	84.40	30	0.47	0.02
FURIA	0.92	0.97	0.92	87.22	18	0.51	0.02
P2							
C4.5	0.85	0.90	0.85	85.91	55	1.00	0.00
CDTree	0.85	0.90	0.85	83.90	35	0.25	0.00
LADTree	0.88	0.90	0.88	85.91	20	0.24	0.00
RIPPER	0.89	0.88	0.89	84.64	12	0.25	0.02
RART	0.85	0.92	0.85	86.33	30	0.41	0.01
RIDOR	0.84	0.90	0.84	83.42	20	0.29	0.02
FURIA	0.89	0.91	0.89	89.36	22	0.27	0.02
P3							
C4.5	0.93	0.95	0.93	93.58	18	1.00	0.00
CDTree	0.91	0.94	0.91	92.54	14	0.72	0.00
LADTree	0.92	0.95	0.92	93.66	14	0.48	0.00
RIPPER	0.90	0.94	0.90	92.23	5	0.44	0.02
RART	0.92	0.93	0.92	92.48	11	0.60	0.01
RIDOR	0.90	0.94	0.92	92.09	8	0.60	0.02
FURIA	0.92	0.97	0.92	94.44	12	0.70	0.01

In P2, we can observe a decrease in the overall accuracy compared to those shown in P1. This is due to the new variables added to the feature list that are not of significant importance. Consequently, the learning process is negatively affected. For example, under decision trees techniques, the LADTree and C4.5 achieve high value in term of accuracy (85.91%), while result obtained by CDTree in the lowest value in term of the same metric (83.9%). Under rule-based algorithms, the FURIA algorithm achieves the best value in term of accuracy (89.36%) while it generates 22 rules. In the same period (P2), LADTree algorithm obtains the lowest average rules length (24%), while RIPPER generates the lowest numbers of rules.

In P3, all accuracies values increase compared to the previous steps. For instance, the FURIA achieves the highest values with respect to TN (97%), recall (92%), and accuracy (94.44%). However, this does not give FURIA the superiority of other algorithms. The FURIA algorithm remains the best algorithm in term of model performance matrices, while it generates the highest number of rules. The RIPPER algorithm saved its superiority in term of interpretability.

Generally, the decision trees techniques generate high rules-size without overlap between rules, while rule-based algorithms produce the lower number of rules with more rule overlap values. Consequently, there is a negative relationship between the rules-size and rules-overlap obtained in all study periods (datasets). Based on this fact, we use rule-size and rules-average as indicators for selecting the algorithms. In addition, we can observe a negative relationship between MI (such as rules-size) and MB measurements.

Finally, the observations showed that the RIPPER is the "best" algorithm in term of interpretability and the FURIA algorithm is the "best" algorithm in term of accuracy

--Experiment II:

In this experiment, the goal is to investigate the improvement gained by applying the RG* method on the students' dataset during the three periods, then compare this improvement by those obtained in the first experiment.

As shown in P1 of Table 6, one rule is added to the ruleset, and it leads to an increase in the average rule values from (60.4%) to (62.5%). However, this addition leads to improvement in the accuracy from (85.2%) to (86.1%). In P2, only two rules succeeded to improve accuracy from (84.64%) to (86.07). Unfortunately, it increases the rule size value by 0.16%. Despite the increase, the superiority of the generated ruleset is saved in terms of interpretability. The best improvement appears in the last step by adding two rules to the generated ruleset that obtained 93.13.

TABLE VI
DESCRIPTION OF DATASET USED FOR THE SECOND EXPERIMENT

	Acc.	Size(Ri)	AL(Ri)	OF(Ri)
P1				
RIPPER	85.25	13	0.604	0.01
FURIA	87.22	18	0.747	0.03
RG*.	86.10	14	0.625	0.01
P2				
RIPPER	84.64	12	0.701	0.01
FURIA	89.36	22	0.812	0.01
RG*	86.07	14	0.797	0.01
P3				
RIPPER	92.23	5	0.25	0.01
FURIA	94.44	12	0.27	0.01
RG*	93.13	7	0.75	0.01

C. Analyzing the rules generated by the Framework

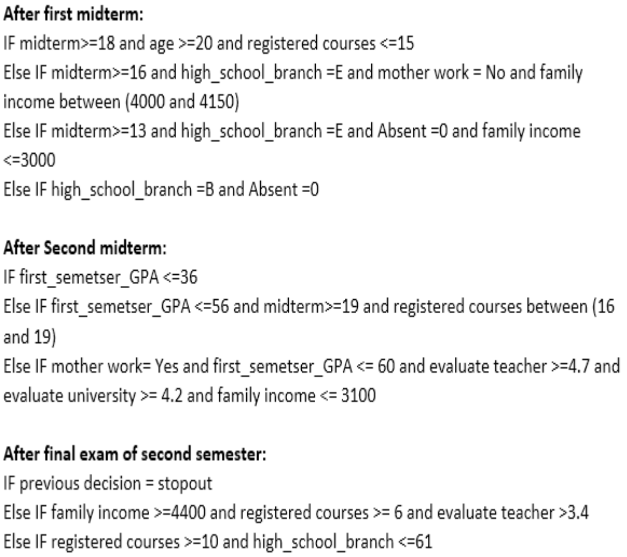
In this section, we illustrate' rules derived from the proposed framework to identify the features that are affecting the decision of the students and to understand the status that distinguishes the dropout and stopout.

Frequently, the dropout decision is taken after the first mid-term if a student has obtained less than 13 points in the mid-term exam and if he has absented more than once and if the age is less than 19, and the family member greater than 4. The dropout decision is taken after the second semester if the previous GPA is 39-61, and the number of registered hours is less than 10 credit hours, and family income is less than NIS4,000 (the average income in Palestine is NIS 3,000 [99]) and if the teachers' rating is less than 3.5. The dropout decision is taken after the final exam (of the second semester) if the student's previous mark is less than 43. On this basis, we note that students' assessments are key factors in making the decision to drop out. On the other hand, the stopout decision is taken after the first midterm-exam, if a student midterm mark is greater than 13, his age is higher than 19, and the number of registered courses is less than 15 credit hours; the family income of a student is less than NIS3,000, the high school branch for a student E or B, and the number of student absences is zero. The stopout decision may be made by a student after the second midterm exam if he achieves less than 19 points in the previous midterm exam, the GPA of the first semester is less than 60 and the family's income is less than NIS3,000.

For example, if a students' evaluation points of the teachers and/or university is greater than 4.2, he is more likely to stopout. Also, the students who stopout at the first semester are more likely to stopout again at the second semester, especially, if the family income is over NIS4,400, the number of registered courses (at the second semester) is greater than 6 credit hours and the evaluation score of teachers is larger than 3.5. (See the lists of rules below).

The results of this paper can be used by educational managers, and collaborative applications [11] for the development of a recommendation system where these rules are stored in the knowledge statements'.

2. List of rules that identify stopout students'.

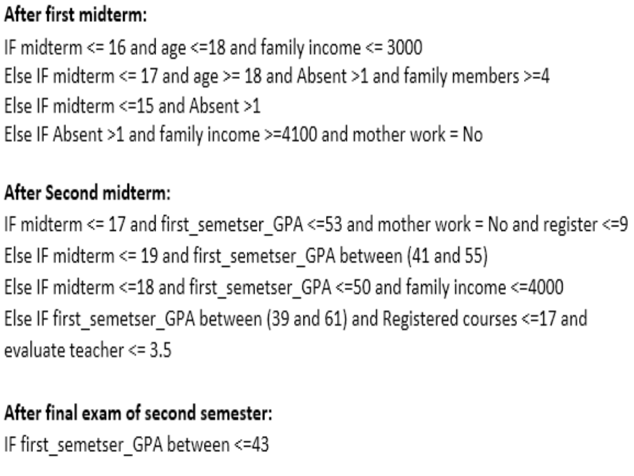


After first midterm:
IF midterm>=18 and age >=20 and registered courses <=15
Else IF midterm>=16 and high_school_branch =E and mother work = No and family income between (4000 and 4150)
Else IF midterm>=13 and high_school_branch =E and Absent =0 and family income <=3000
Else IF high_school_branch =B and Absent =0

After Second midterm:
IF first_semetser_GPA <=36
Else IF first_semetser_GPA <=56 and midterm>=19 and registered courses between (16 and 19)
Else IF mother work= Yes and first_semetser_GPA <= 60 and evaluate teacher >=4.7 and evaluate university >= 4.2 and family income <= 3100

After final exam of second semester:
IF previous decision = stopout
Else IF family income >=4400 and registered courses >= 6 and evaluate teacher >3.4
Else IF registered courses >=10 and high_school_branch <=61

1. List of rules that identify dropout students'.



After first midterm:
IF midterm <= 16 and age <=18 and family income <= 3000
Else IF midterm <= 17 and age >= 18 and Absent >1 and family members >=4
Else IF midterm <=15 and Absent >1
Else IF Absent >1 and family income >=4100 and mother work = No

After Second midterm:
IF midterm <= 17 and first_semetser_GPA <=53 and mother work = No and register <=9
Else IF midterm <= 19 and first_semetser_GPA between (41 and 55)
Else IF midterm <=18 and first_semetser_GPA <=50 and family income <=4000
Else IF first_semetser_GPA between (39 and 61) and Registered courses <=17 and evaluate teacher <= 3.5

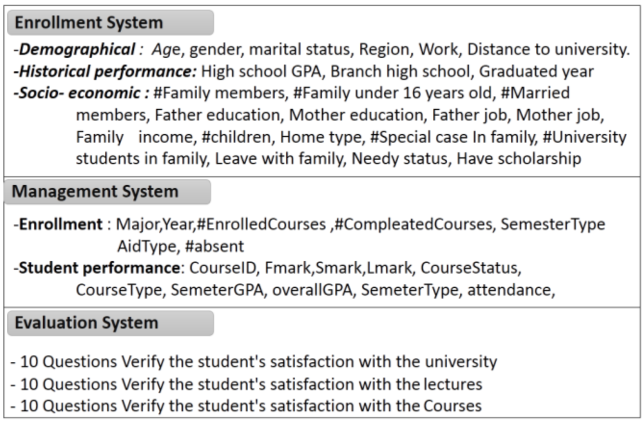
After final exam of second semester:
IF first_semetser_GPA between <=43

V. Conclusion

All educational institutions are interested in following up their students to make sure that they can graduate on a timely manner. Therefore, these institutions have interested in finding creative ways to deal with the withdrawal problem. The previous studies have proposed methodologies that estimate the number of enrolled students and to predict the number of drop-outs as well as the identification of the factors that affect students' decisions. However, those studies did not differentiate between the different types of withdrawals. In addition, the studies concentrated on measured the performance of their proposed methodologies based on the accuracy of the prediction without taking into consideration the interpretability of their results. We introduce a new methodology that fill the gap in the existing methodologies by implementation a two-step process: Ensemble Learning and Filtering. The results improve of the resulting rules as they become more interpretable and higher level of prediction accuracy. The resulting rules help mangers to identify the factors that affect students' withdrawal decision and to differentiate between the different withdrawal types. Our results improve the ability of the recommender systems' developers using the resulting rules.

Finally, it is worth to mentioning that the main limitation of the proposed method is its high complexity due to using multiple classifiers which is affected by the data size. This limitation is to be solved through future research work.

Appendix



Enrollment System

-***Demographical*** : *Age*, gender, marital status, Region, Work, Distance to university.
-***Historical performance:*** High school GPA, Branch high school, Graduated year
-***Socio- economic :*** #Family members, #Family under 16 years old, #Married members, Father education, Mother education, Father job, Mother job, Family income, #children, Home type, #Special case In family, #University students in family, Leave with family, Needy status, Have scholarship

Management System

-**Enrollment** : Major,Year,#EnrolledCourses ,#CompleatedCourses, SemesterType AidType, #absent
-**Student performance**: CourseID, Fmark,Smark,Lmark, CourseStatus, CourseType, SemeterGPA, overallGPA, SemeterType, attendance,

Evaluation System

- 10 Questions Verify the student's satisfaction with the university
- 10 Questions Verify the student's satisfaction with the lectures
- 10 Questions Verify the student's satisfaction with the Courses

Fig. A.1 The features used in this study

Acknowledgment

The authors would like to thank both the Chinese scholarship council for supporting and funding this study, and the Palestine Technical College for providing data and their faculty staff for providing insights and information that have greatly helped in this research.

References

[1] S.P. Karkhanis, andS.S. Dumbre, "A Study of Application of Data Mining and Analytics in Education Domain". International Journal of Computer Applications, 2015. 120(22).

[2] D. Shapiro, A. Dundar, X. Yuan, A.T. Harrell, J.C. Wild, andM.B. Ziskin, "Some College, No Degree: A National View of Students with Some College Enrollment, but No Completion (Signature Report No. 7)". National Student Clearinghouse, 2014.

[3] V. Tinto, "Research and practice of student retention: What next?". Journal of College Student Retention: Research, Theory & Practice, 2006. 8(1): p. 1-19.

[4] L.S. Stratton, D.M. O'Toole, andJ.N. Wetzel, "A multinomial logit model of college stopout and dropout behavior". Economics of education review, 2008. 27(3): p. 319-331.

[5] O.F. Porter, "Undergraduate Completion and Persistence at Four-Year Colleges and Universities: Completers, Persisters, Stopouts, and Dropouts". 1989.

[6] L.J. Horn, Stopouts or Stayouts? Undergraduates Who Leave College in Their First Year. Statistical Analysis Report. 1998: ERIC.

[7] S. Herzog, "Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression". New directions for institutional research, 2006. 2006(131): p. 17-33.

[8] R. Chen, "Institutional characteristics and college student dropout risks: A multilevel event history analysis". Research in Higher Education, 2012. 53(5): p. 487-505.

[9] M. Clark, andN.L. Cundiff, "Assessing the effectiveness of a college freshman seminar using propensity score adjustments". Research in Higher Education, 2011. 52(6): p. 616-639.

[10] L. Bonaldo, andL.N. Pereira, "Dropout: Demographic profile of Brazilian university students". Procedia-Social and Behavioral Sciences, 2016. 228: p. 138-143.

[11] M. Anzures-García, L.A. Sánchez-Gálvez, M.J. Hornos, and P. Paderewski-Rodríguez, "A Knowledge Base for the Development of Collaborative Applications," Engineering Letters, vol. 23, no. 2, pp65-71, 2015.

[12] S. Liu, R.Y. Patel, P.R. Daga, H. Liu, G. Fu, R.J. Doerksen, Y. Chen, andD.E. Wilkins, "Combined rule extraction and feature elimination in supervised classification". IEEE transactions on nanobioscience, 2012. 11(3): p. 228-236.

[13] J.T. Lalis, "A New Multiclass Classification Method for Objects with Geometric Attributes Using Simple Linear Regression". IAENG International Journal of Computer Science, 2016. 43(2): p. 198-203.

[14] A.O. Abuassba, D. Zhang, X. Luo, A. Shaheryar, andH. Ali, "Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines". Computational intelligence and neuroscience, 2017. 2017.

[15] R.P. Duin, andD.M. Tax. Experiments with classifier combining rules. in International Workshop on Multiple Classifier Systems. 2000. Springer.

[16] X.N. Nashat Al-jallad, Mergani A. Khairalla "Rule Mining models for Predicting Dropout/Stopout and Switcher at College using Satisfaction & SES Features". International Journal of Management in Education, forthcoming,http://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijmie

[17] P. Meedech, N. Iam-On, andT. Boongoen, Prediction of student dropout using personal profile and data mining approach, in Intelligent and Evolutionary Systems. 2016, Springer. p. 143-155.

[18] J. Abellán, andS. Moral, "Building classification trees using the total uncertainty criterion". International Journal of Intelligent Systems, 2003. 18(12): p. 1215-1225.

[19] J. Fürnkranz, D. Gamberger, andN. Lavrač, Foundations of rule learning. 2012: Springer Science & Business Media.

[20] W.N.H.W. Mohamed, M.N.M. Salleh, andA.H. Omar. A comparative study of reduced error pruning method in decision tree algorithms. in Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on. 2012. IEEE.

[21] H. Lakkaraju, S.H. Bach, andJ. Leskovec. Interpretable decision sets: A joint framework for description and prediction. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. ACM.

[22] J. Derrac, S. García, D. Molina, andF. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms". Swarm and Evolutionary Computation, 2011. 1(1): p. 3-18.

[23] J. Hodges, andE. Lehmann, Rank methods for combination of independent experiments in analysis of variance, in Selected Works of EL Lehmann. 2012, Springer. p. 403-418.