PREDICTING ACADEMIC SUCCESS THROUGH STUDENTS' INTERACTION WITH VERSION CONTROL SYSTEMS

GRUPO DE ROBÓTICA

universidad de león

ROBOTICA, LEARNING ANALYTICS Y TICS APLICADAS A LOS PROCESOS DE ENSEÑANZA/APRENDIZAJE (ULE ROBLATIC)

Miguel Ángel Conde, Gonzalo Esteban González, Camino Fernández Llamas, Ángel Manuel Guerrero Higueras, Vicente Matellán Olivera, Francisco Javier Rodríguez Lera, Francisco Jesús Rodríguez Sedano, Lidia Sánchez González
{mcong, gestc, camino.fernandez, am.guerrero, vicente.matellan, fjrodl, francisco.sedano, lidia.sanchez}@unileon.es

INTRODUCTION

- Version Control Systems (VCSs) allow monitoring programmers activity working in a project.
- These systems are commonly used by Information and Communication Technologies (ICTs) professionals, so they are also used by educational institutions.
- The aim of this work is to evaluate if the academic success of students may be predicted by monitoring their interaction with a VCS. In order to do so, we have built a model which predicts student results in a specific practical assignment of the second course of the degree in Computer Science of the University of León, through their interaction with a GIT repository.

GOAL

The goal of this work is to answer the following research questions:

Question 1 Are there any features that we can extract from the students' interactions with VCs that are related to academic success?

Question 2 Can we build a model that allows predicting students' success at a practical assignment, by monitoring their use of a VCS?

To answer the above questions, we have carried out an experiment in the Operating Systems Extension subject from the second course of the degree in Computer Science of the University of León.

METHODOLOGY

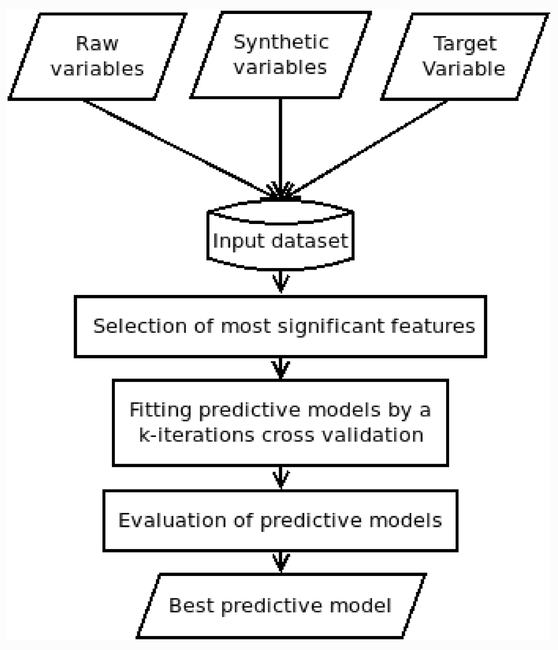

Figure 1. Methodology.

1. We start from an input dataset that contains two different type of data.
   - We have variables that directly come from a GIT repository (*raw variables*): *id*, *commits*, *additions*, *deletions*, *issues opened* and *issues closed*.
   - we also have variables that a researcher constructs based on the raw variables, or that they are provided by other sources (*synthetic variables*): *days*, *commits/day*, and *authorship proof*.
   - we also need a target variable (*class*); namely, a variable with the labels of classification that let us train and test the supervised learning model. Our target variable has two possible values: "AP", and "SS".
2. The following step is determining what are the most significant features in order to obtain a classification model based on the target variable. Feature selection is a procedure that selects the features that contribute most to the classification or the prediction.
3. Once selected the most significant features, several models are fitted to predict the target variable from input data. We work with the following well-known methods: Adaptive Boosting, Classification And Regression Tree, K-Nearest Neighbors, Linear Discriminant Analysis, Logistic Regression, Multi-Layer Perceptron, Naive Bayes, and Random Forest. To fit the above models, we perform a k-iteration cross-validation.
4. Finally, to get the most suitable learning algorithms, we evaluate the previous models. In order to do so, we compute some well-known KPIs:
   - The accuracy classification score is computed as follows, where $\sum T_p$ is the number of true positives, and $\sum T_n$ is the number of true negatives.

$$accuracy = \frac{\sum T_p + \sum T_n}{\sum \text{total data}}$$

   - The three models with the highest accuracy classification score have been pre-selected for in-depth evaluation by considering the following KPI: Precision (P), Recall (R), and F_1-score; all of which were obtained through the confusion matrix.

$$P = \frac{\sum T_p}{\sum T_p + \sum F_p} \qquad R = \frac{\sum T_p}{\sum T_p + \sum F_n} \qquad F_1 = 2\frac{P \times R}{P + R}$$

$\sum F_p$ is the number of false positives. $\sum F_n$ is the number of false negatives.

RESULTS

Table I. Accuracy classification score.

Classifier	Test score	Validation score
NB	0.8	**0.8**
RF	0.8	**0.8**
LDA	0.8	**0.7**
MLP	0.5	0.7
CART	0.4	0.6
AB	0.4	0.5
LR	0.7	0.5
KNN	0.6	0.5

Table II. Accuracy classification score without considering *authorship proof*.

Classifier	Test score	Validation score
RF	0.7	**0.7**
MLP	0.4	**0.7**
LDA	0.6	**0.6**
CART	0.6	0.6
NB	0.5	0.5
KNN	0.6	0.5
AB	0.4	0.5
LR	0.6	0.4

Table III. Precision, recall and F_1-score for the test dataset.

Classifier	Class	P	R	F_1-score	#examples
	AP	0.67	1.00	0.80	4
NB	SS	1.00	0.67	0.80	6
	avg/total	0.87	0.80	0.80	10
	AP	0.67	1.00	0.80	4
RF	SS	1.00	0.67	0.80	6
	avg/total	0.87	0.80	0.80	10
	AP	0.67	1.00	0.80	4
LDA	SS	1.00	0.67	0.80	6
	avg/total	0.87	0.80	0.80	10

Table IV. Precision, recall and F_1-score for the validation dataset.

Classifier	Class	P	R	F_1-score	#examples
	AP	0.79	0.90	0.84	21
NB	SS	0.88	0.74	0.80	19
	avg/total	0.83	0.82	0.82	40
	AP	0.76	0.90	0.83	21
RF	SS	0.87	0.68	0.76	19
	avg/total	0.81	0.80	0.80	40
	AP	0.86	0.57	0.69	21
LDA	SS	0.65	0.89	0.76	19
	avg/total	0.76	0.72	0.72	40

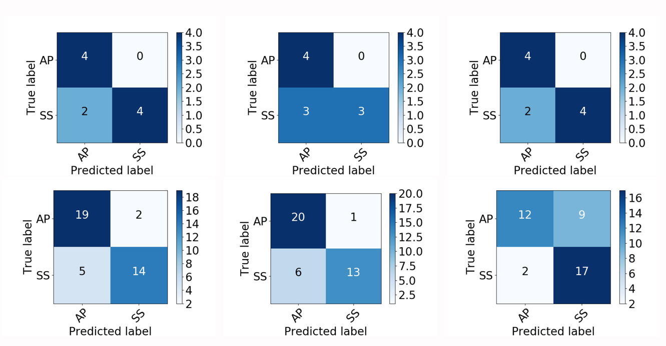

Figure 2. Top: Confusion matrix for the NB (left), RF (center), and LDA (right) classifiers evaluated using the test dataset. Bottom: Same data using the validation dataset.

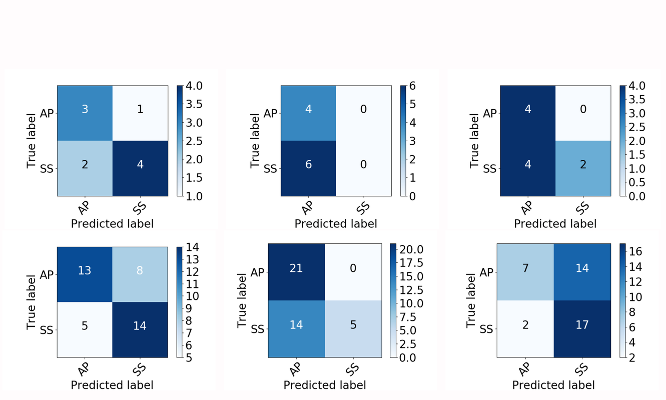

Figure 3. Top: Confusion matrix for the RF (left), MLP (center), and LDA (right) classifiers evaluated using the test dataset without considering *authorship proof*. Bottom: Same data using the validation dataset.

Table V. Precision, recall and F_1-score for the test dataset without considering *authorship proof*.

Classifier	Class	P	R	F_1-score	#examples
	AP	0.60	0.75	0.67	4
RF	SS	0.80	0.77	0.73	6
	avg/total	0.72	0.70	0.70	10
	AP	0.40	1.00	0.57	4
RF	SS	0.00	0.00	0.00	6
	avg/total	0.16	0.40	0.23	10
	AP	0.50	1.00	0.67	4
LDA	SS	1.00	0.33	0.50	6
	avg/total	0.80	0.60	0.57	10

Table VI. Precision, recall and F_1-score for the validation dataset without considering *authorship proof*.

Classifier	Class	P	R	F_1-score	#examples
	AP	0.72	0.62	0.67	21
RF	SS	0.64	0.74	0.68	19
	avg/total	0.68	0.68	0.67	40
	AP	0.60	1.00	0.75	21
RF	SS	1.00	0.26	0.42	19
	avg/total	0.79	0.65	0.57	40
	AP	0.78	0.33	0.47	21
LDA	SS	0.55	0.89	0.68	19
	avg/total	0.67	0.60	0.57	40

CONCLUSIONS

- With regard to the first question, the feature analysis carried out show the importance of each feature. This allows identifying which ones have a greater weight in the model. This is the first step to obtaining a classification model that allows predicting the academic success of students. Results show that some features related to students interaction with the VCS are discriminant. However, include more features, such as an authorship proof, increase models accuracy.
- Relative to the second question posed, we provide a prediction model by evaluating several classifiers. There are future works to do due to optimizing the selected model by tuning its hyperparameters, but results are enough to assert that we can predict students' results at a practical assignment with a success high percentage.
- Further works should face accuracy improvement. In order to do so, in addition to hyperparameters tuning, it would be desirable to increase training data. A first approach may be done by combining both test and validation dataset. However, a prior analysis is required in order to assert that there are not statistically meaningful differences between both datasets.

ACKNOWLEDGMENTS

This work has been partially funded by the "Plan de Apoyo a los Grupos de Innovación Docente (PAGID)" of the University of León.

PUBLICATIONS

- Guerrero-Higueras, Á. M., Conde, M. Á., & Matellán, V. (2017, October). Uso de sistemas de control de versiones para aplicar estrategias de evaluación por pares en contextos tecnológicos. In *IV Congreso Internacional sobre Aprendizaje, Innovación y Competitividad (CINAIC 2017)* (pp. 4-6).
- Guerrero-Higueras, Á. M., Matellán, V., Esteban-Costales, G., Fernández-Llamas, C., Rodríguez-Sedano, F. J., & Conde, M. Á. (2018, June). Model for evaluating student performance through their interaction with version control systems. In *Learning Analytics Summer Institute Spain*.
- Guerrero-Higueras, Á. M., DeCastro-García, N., Matellán, V., & Conde, M. Á. (2018, October). Predictive models of academic success: a case study with version control systems. In Proceedings of the *Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 306-312). ACM.
- Guerrero-Higueras, Á. M., DeCastro-García, N., Rodríguez-Lera, F. J., Matellán, V., & Conde, M. Á. (2019, Septembter). Predicting academic success through students' interaction with Version Control Systems. *Open Computer Science*.
- Guerrero-Higueras, Á. M., Sánchez-González, L., Fernández-Llamas, C., Conde, M. Á., Rodríguez-Lera, F. J., Rodríguez-Sedano, F. J., Esteban-Costales, G., & Matellán, V. (2019, October). Prediction of academic success through interaction with version control systems. In Proceedings of the *Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*.

TEEM'19

7^{th} **International Conference on Technological Ecosystems for Enhancing Multiculturality**
16–18 October 2019, León, Spain

Plan de Apoyo a los Grupos de Innovación Docente (PAGID)
Escuela de formación del profesorado

universidad de león