CIID
I CONGRESO INTERNACIONAL DE INNOVACIÓN DOCENTE
20 y 21 de FEBRERO DE 2014
MURCIA - CARTAGENA (ESPAÑA)

Data Mining for Exploring E-learning in a Computer Science Course Using On-line Judging

Autor/res/ras: José Luis Fernández-Alemán[1], David Gil[2], Ginés García Mateos[3], Juan Carlos Trujillo[2], Ambrosio Toval[1]

Institución u Organismo al que pertenecen: [1]GIIS, DIS, University of Murcia, [2]Lucentia, DLSI, University of Alicante, [3]GIIA, DIS, University of Murcia,

Indique uno o varios de los seis temas de Interés: (Marque con una {x})

{ } Enseñanza bilingüe e internacionalización

{ } Movilidad, equipos colaborativos y sistemas de coordinación

{X} Experiencias de innovación apoyadas en el uso de TIC. Nuevos escenarios tecnológicos para la enseñanza y el aprendizaje.

{ } Nuevos modelos de enseñanza y metodologías innovadoras. Experiencias de aprendizaje flexible. Acción tutorial.

{ } Organización escolar. Atención a la diversidad.

{ } Políticas educativas y reformas en enseñanza superior. Sistemas de evaluación. Calidad y docencia.

Idioma en el que se va a realizar la defensa: (Marque con una {x})

{ } Español **{ X}** Inglés

Resumen. En el Espacio Europeo de Enseñanza Superior emergen nuevas metodologías de enseñanza basadas en el proceso de aprendizaje de los estudiantes, que promueven el interés de los estudiantes y ofrecen retroalimentación personalizada. Los sistemas de enjuiciamiento en red son métodos prometedores para estimular la participación de los estudiantes en el proceso de aprendizaje. La enorme cantidad de datos disponible en un sistema de enjuiciamiento en red ofrece la posibilidad de explorar qué parámetros son relevantes para el aprendizaje de la programación de computadores. En este artículo, se identifican los factores que afectan a la corrección de los programas a partir de las actividades de programación en un curso de algoritmos y estructuras de datos. Se utilizan tecnologías de minería de datos como los árboles de decisión, que han demostrado ser muy efectivos como predictores en algunos dominios de aprendizaje electrónico. Los resultados muestran que los parámetros Lenguaje de programación, Número de problema y Titulación pueden ser utilizados como predictores de la corrección de un programa, con una precisión del 60,1%. Como trabajo futuro, pretendemos estudiar los factores que afectan al rendimiento de los trabajadores en un entorno de desarrollo global del software, aplicando la minería de datos a actividades de programación colaborativas.

Palabras Claves: Minería de datos, aprendizaje electrónico, enjuiciamiento en red

Abstract. New teaching methods based on the students' learning process are being developed in the European Higher Education Area. Most of them are oriented to promote students' interest in the study and offer personalized feedback. On-line judging is a promising method for encouraging students' participation in the e-learning process. The great amount of data available in an on-line judging tool provides the possibility of exploring some of the most indicative attributes for learning programming concepts and techniques. In this paper, the results of programming activities carried out in a course on "Algorithms and Data Structures" has been used to identify the factors that affect the program correction, by using powerful data mining technologies taken from artificial intelligence domain. Concretely, our study uses a decision tree because it has been identified as the best predictor in some e-learning domains. An overall accuracy of 60.1% in the prediction of the program correction was achieved with three input parameters (Programming language, Number of problem and Degree). In future work, we aim to analyze collaborative activities in order to identify the factors or predictor variables that affect workers' performance in a global software development context.

Keywords: Data mining, e-learning, experience report, on-line judging

1. Introduction

More than four years after the birth of the European Higher Education Area (EHEA), we are starting to observe profound changes in the processes of teaching/learning, thanks to the development of new methodologies that it boosted. It is well known that the purpose of the EHEA was to develop new teaching methods based on the students' learning process rather than the teacher's point of view. Under this perspective, the new methods should stimulate students' interest and offer appealing material, fair assessment and relevant feedback.

Based on these ideas and principles, we developed an innovative experience with a course on "Algorithms and Data Structures" (ADS) in the second year of a Computer Science Degree, using a web-based automatic judging system called Mooshak (García-Mateos & Fernández-Alemán, 2009). The course was organized as a series of activities in a continuous evaluation context. Many of these activities used the on-line judging system; some of them were done individually, while others were done collaboratively by a group of students. These three elements –on-line judging, continuous evaluation and collaborative work– were the keys to generate motivation and enthusiasm among students, who were used to traditional teaching methods. The effect of that experience was an important descent on the dropout rates and a general improvement on marks and pass rates.

Thus, in subsequent years, the proposed methodology was adopted as the new basic organization of the referred course on ADS with some minor changes. Now, with six years of experience using on-line judging in class, we have available a great amount of data in order to research different aspects of the learning process. In particular, the goal of this study is to look into the results of programming activities done in ADS, by using powerful data mining technologies taken from artificial intelligence (AI) domain. These tools are able to extract relevant information from a huge amount of data which are able to find out hidden correlations among data. In our case, this data refers to the submissions done by the students to the programming activities of ADS to the on-line judge. More specifically, we want to investigate whether or not it is possible to predict the result of each submission (accepted/rejected) using only the context where that submission is done.

The rest of the paper is organized as follows. Section 2 presents a brief review of some related work. Then, in Section 3, we present the proposal of identifying valid, useful, and understandable patterns in data based on a Knowledge Discovery in Databases (KDD) process. Section 4 introduces and justifies the methodological approach of the proposal. Section 5 offers the main results of the research. In Section 6, we analyze and discuss the results achieved after applying data mining on a great amount of programs stored in the on-line judge. The last section presents some concluding remarks and outlines the efforts to be made in the future.

2. Related work

A recent review analyzes and discusses strengths, weakness, opportunities, and threats of 240 educational data mining works describing approaches and tools (Peña-Ayala, 2014). Another work reviews how the data mining was tackled by previous scholars and identifies limitations and latest trends on data mining in educational research (Mohamad & Tasir, 2013). Data mining has been applied to learning management systems (Lara, Lizcano, Martínez, Pazos, & Riera, 2014). The authors analyze data generated by the interaction of students with the Moodle platform. Clustering, association, classification, and time series analysis were some of the data mining techniques used. The system generated built historical reference models of students that both dropped out of and completed the course. The proposal was validated on real academic data, by using data from the informatics engineering courses at Madrid Open University (UDIMA).

3. Data mining

Data mining, which very often is called "Knowledge Discovery in Databases" (KDD), is the process of analyzing data from different perspectives and summarizing it into a useful and understandable structure for further use. This implies to discover patterns in large data sets involving methods at different areas such as AI, machine learning, statistics, and database systems (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) (Hastie, Tibshirani, & Friedman, 2009). Data mining is a relatively new term but the technology is not. Many companies have used for years some software packages to retrieve a lot of information (e.g. volumes of supermarket scanner data and market research reports to discover customer’s choices). However, in this area good things are yet to come since there are continuous innovations in computer processing power, disk storage, and statistical as well as AI software tools. This will make possible to increase the accuracy of analysis whereas will drive down the cost.

Data mining consists of four major elements: (1) extract, transform, and load transaction data onto the data warehouse system; (2) store and manage the data in a multidimensional database system; (3) analyze the data by application software, which is the procedure to apply AI or statistical methods; and (4) visualization of data.

Note that the third step can cover different levels of analysis. Several methods included in a software package can be applied. Just to name a few of them:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure, mainly with a few layers of neurons from the input towards the output.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. They provide a set of rules that can be applied to a new (unclassified) dataset to predict which records will have a given outcome.
- K-Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

In this paper we have used a decision tree (DT) not only for prediction of the output program but also to find out the correlation among the different input parameters. A Decision tree is a classifier in the form of a tree structure, where each node is either a leaf node, which indicates the value of the target attribute (class) of examples, or a decision node, which specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. DTs are powerful and popular tools for classification and prediction. In contrast to other AI methods, such as Artificial Neural Networks, the advantages of DTs are due to the fact that they represent rules as well as their very representative visualizations. These are the reasons for choosing this AI method besides the good accuracy, in general, for classification. Rules can readily be expressed so that persons can understand them or even directly use them in a database. In some areas of applications, the accuracy of a classification or prediction is the only thing that matters. In such situations we are not necessarily concerned about how the model works. However, in other situations, the ability to explain the reason for a decision is crucial. We really believe that in our case by distinguishing between the different program outputs we can obtain a successful prediction. The classification and regression trees (CART or C&RT) method of Breiman, Friedman, Olshen, and Stone (Breiman, 1984) generates binary decision trees. In the real world, the chances of biased binary outcomes are few, but the binary method allows for easy interpretation and analysis (Ture, Tokatli, & Kurt, 2009; Yang et al., 2003). Thus, the study uses the binary method in the DT to classify the different program outputs between Error and Correct.

4. Methodology

4.1 Participants

Students were recruited from a course on ADS which was present in three different degrees. Students were distributed in three groups: TECS (Technical engineering in computer systems), TECM (Technical engineering in computer management) and CSE (Computer science engineering). ADS is an advanced course in programming, which stresses issues of algorithms and data representation. This course introduces such topics as data structures, abstract data types and formal specifications. The course also includes techniques for performance analysis and design of algorithms. The programming languages used to illustrate these concepts are C, C++ and Maude. ADS was organized by weekly lectures, laboratory sessions, a final exam and a programming project for each semester.

4.2 Instruments

The learning environment used to evaluate cooperative work of the students was an on-line judging system. Mooshak (Leal & Silva, 2003) is a free and publicly available automatic tool which is able to evaluate the correctness of computer programs, based on a predefined set of pairs input/output. It has a web-based interface, which is different for the students, teachers, guest users and the system administrator. This system was originally created to manage programming competitions. However, Mooshak has been applied to both computer programming and nursing learning (Fernández-Alemán, 2011; Fernández Alemán, Carrillo de Gea, & Rodríguez Mondéjar, 2011; García-Mateos & Fernández-Alemán, 2009; Montoya-Dato, Fernandez-Aleman, & Garcia-Mateos, 2009).

The application used to analyze the data collected and apply data mining was Weka (Waikato Environment for Knowledge Analysis). This is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. It contains a collection of machine learning algorithms for data mining tasks (Holmes, Donkin, & Witten, 1994; Witten & Frank, 2005) . The algorithms can either be applied directly to a dataset or called from an own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes.

5. Results

In this section, we provide detailed information about the results of the experiment conducted at the Computer Science Faculty of the University of Murcia (Spain). Up to 273 of the 337 enrolled students (81%) participated in some activity of the judge; 268 of them (79,5%) solved at least one problem. In total, the on-line judge received 16054 submissions. This makes an average of 59 submissions per student: 44 C/C++ programs, and 15 Maude programs. The on-line judge classified 6427 submissions as correct (40,1%). Each student attempted an average of 20 problems and managed to solve 18,5. In this paper we present the results of mining the programs submitted by the students. To illustrate the viability of the proposal, the example set contained 2895 submissions. Metadata of each submission (programming language, number of problem, degree, time, day and output result) are shown in Table 1.

Name	Type	Description
Programming language	C, C++, TAR_File	Type of language used
Number of problem	Numeric	Problem identifier
Degree	ITIS, ITIG, II	The degree in which the student is enrolled
Time	Morning, Afternoon, Night	Time of the day when the submission was done
Day	Numeric	Number of days after the beginning of the activity
Output result	E, A	Error/ Accepted

Table 1. Input parameters for data mining

A bar diagram for each variable is presented in Figure 1 in order to show the correlations between the input variables (Programming language, Number of problem, Degree, Time and Day) and the output result. Note that problem 20 in degree ITIG achieved higher percentage of incorrect programs than the rest of values. Moreover, this percentage is also higher last days of the activity, when deadline is approaching. During this period of time, students dedicate less time to think and reflect on the problems.

The prediction results of the method used (decision tree) are presented in Figure 2. Since the output variable had two nominal Values (accepted/error), the confusion matrix shows 2x2 square matrix where the correct predictions are places at the diagonal from upper left to lower right corner (values 649 and 1066). In the confusion matrix the columns represent the actual whereas the rows represent the predictions. The prediction accuracy for the output variable value Accepted was 48,07%, whereas the prediction accuracy for the output variable value Error was 68,99%.The overall accuracy was 59,24%.

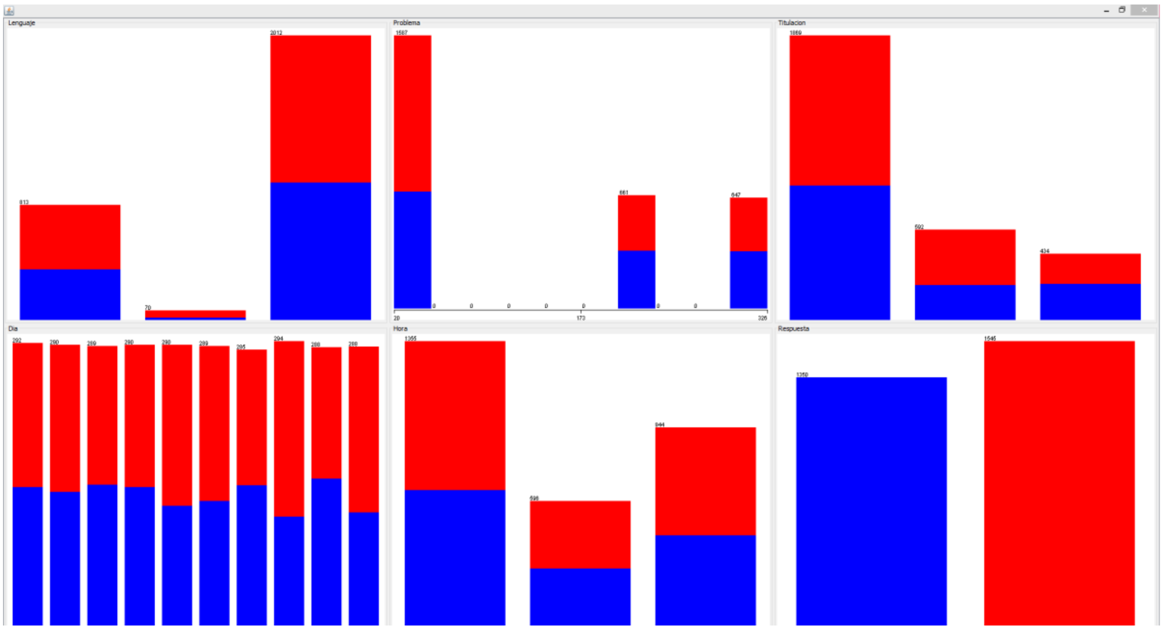
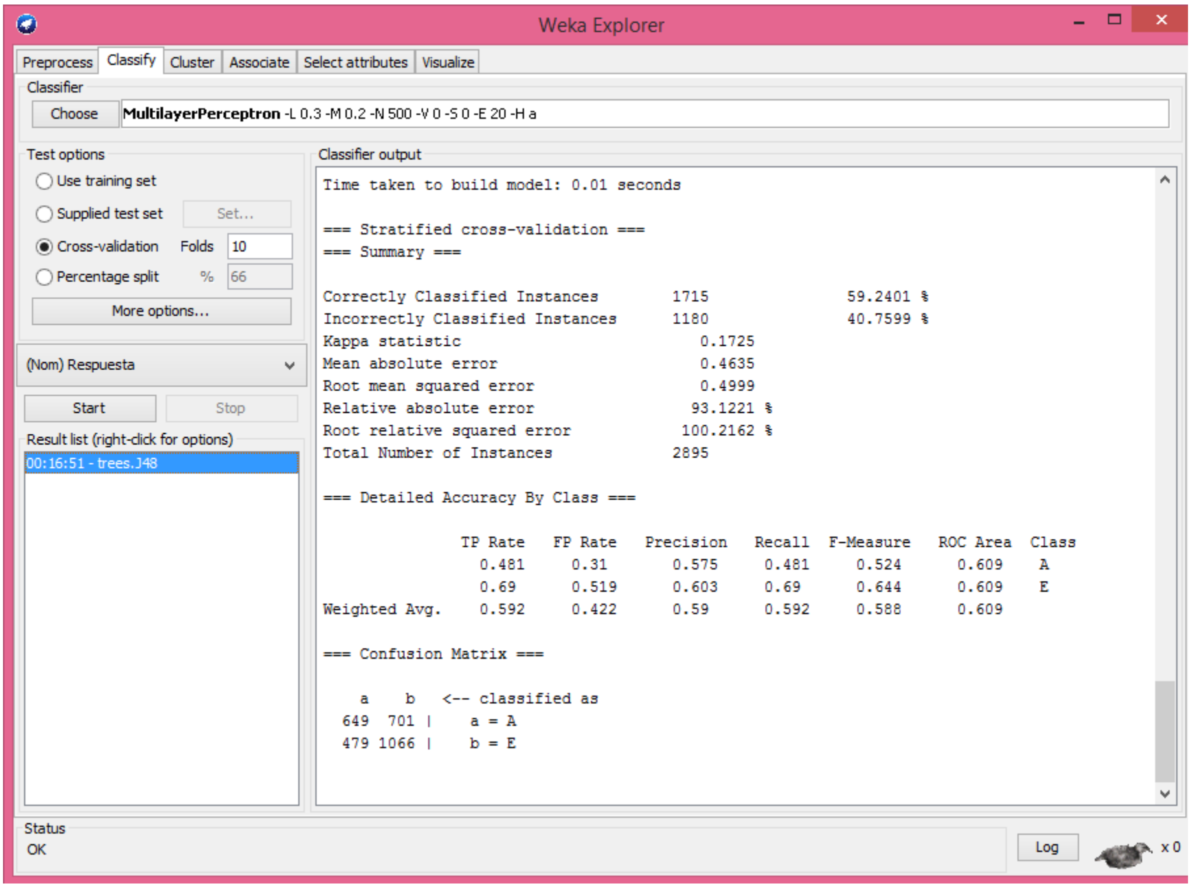


Figure 1 Visualization for every parameter related to output result.



MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Respuesta

00:16:51 - trees.J48

Status: OK

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1715               59.2401 %
Incorrectly Classified Instances      1180               40.7599 %
Kappa statistic                          0.1725
Mean absolute error                      0.4635
Root mean squared error                  0.4999
Relative absolute error                 93.1221 %
Root relative squared error            100.2162 %
Total Number of Instances             2895

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.481     0.31      0.575     0.481     0.524      0.609    A
                 0.69      0.519     0.603     0.69      0.644      0.609    E
Weighted Avg.    0.592     0.422     0.59      0.592     0.588      0.609

=== Confusion Matrix ===

    a    b   <-- classified as
  649  701 |    a = A
  479 1066 |    b = E
```

Figure 2 Prediction results for the classification method based on a decision tree algorithm.

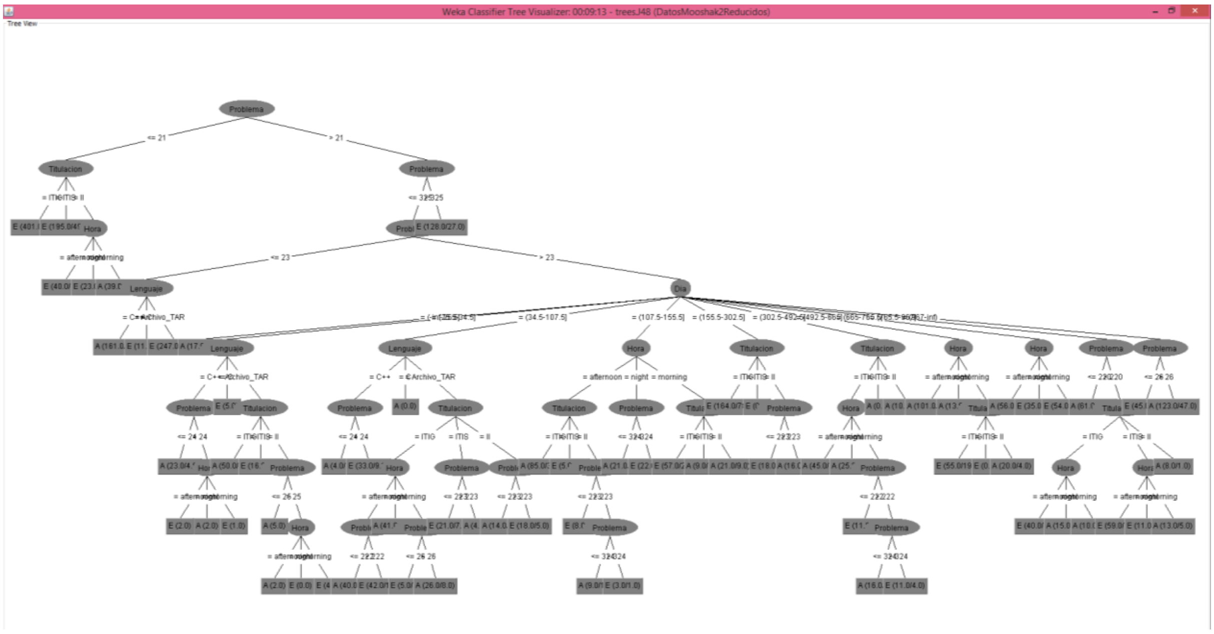


Figure 3 Classification type decision tree output for the on-line judge.

Figure 3 presents the decision tree produced in this study, which has been created by using the 5 input parameters and one output parameter which identifies the output program. Observe that the decision tree algorithm identifies the split values of the most discernible variables to construct branches recursively until the leaf nodes (the predictions) are homogenized to a satisfactory level. For example, the root of the tree shown in Figure 3 identifies Problem as the most discernible variable, with two branches: problems with an identifier higher than 21, and problems with an identifier lesser or equal than 21. The path provided by the answers to the questions in the internal nodes must be followed until a leaf is reached. The DT algorithm family includes classical algorithms, such as ID3 (J.R Quinlan, 1986), (J.R. Quinlan, 1993) C4.5, and CART (Breiman, 1984). The DT shown in Figure 3 is complex, so we searched combinations of input parameters to simplify the DT, but maintaining the overall accuracy. We found that with three input parameters (Programming language, Number of problem and Degree) a slightly higher overall accuracy (60,1%) was achieved.

6. Discussion

DTs are able to provide a good explanation among the diverse input parameters and the program output and, therefore, they interpret problems according to mathematical and statistical principles (Brida & Risso, 2010). DT algorithms have been identified as the best predictors in some e-learning domains (Şen, Uçar, & Delen, 2012). DT algorithms are powerful for analyzing the relationship between independent variables and dependent variables because of the tree searching schema (Lin, Yeh, Hung, & Chang, 2013). Moreover, DT along with Bayes theorem are preferable methods for predictive approaches (Peña-Ayala, 2014). DT is transparent to the end user; since decision tree outputs produce easily readable and understandable models, good and clear parameters visualizations, and they are easy to convert into mathematical expressions (i.e., a series of nested if-then rules) which facilitates their integration into a decision support system.

The application of data mining techniques, especially the decision tree technique, allows educators to tailor the programming activities to students' needs, thus optimizing personalized learning in an on-line judge. In particular, when using the Programming language, Number of problem and Degree in the decision tree algorithm, the prediction of the result of each submission was 60,1%, a result that is similar than that obtained when five input parameters are employed. The programming language has been recognized as a decisive factor in learning ADS (Saito & Yamaura, 2013), as our results confirm. Moreover, each academic discipline has its own learning objectives and selects appropriate pedagogical approaches to help students acquire programming skills needed for their careers. Our findings show that the degree must be taken into account when designing programming activities.

7. Conclusions

Data mining is an important tool for a wide variety of real-world domains (e.g., education, medicine, biology, banking and marketing), which collect and store large amounts of data. Our experiment illustrates the viability of applying data mining techniques to accurately predict and classify on-line judge results from parameters such as programming language and submission time. This may help in designing more effective programming activities and educational methodologies. In addition, this approach can be used for some Massive Open Online Course (MOOC) courses which are very common and extended nowadays.

In future work, we aim to conduct a sensitivity analysis on the prediction model to obtain the most important predictor variables. The credibility and reliability of the sensitivity analysis results will depend on the prediction accuracy of the prediction model. Since in our study the prediction model performed well (accuracy between 59% and 61%), the sensitivity results could not be reliable. Moreover, we purport to analyze collaborative activities in order to identify the factors or predictors variables that affect workers' performance in a global software development context.

8. Acknowledgments

This research is part of the PEGASO-PANGEA projects (TIN2009-13718-C02-02) financed by the Spanish Ministry of Science and Innovation (Spain), and the GEODAS project (TIN2012-37493-C03) financed by both the Spanish Ministry of Economy and Competitiveness and European FEDER funds. This work was also supported by the Spanish MINECO, as well as European Commission FEDER funds, under grant TIN2012-38341-C04-03.

9. Conflict of interest

The authors declare that they have no conflict of interest.

10. References

Breiman, L. (1984). Classification and regression: Chapman.

Brida, J. G., & Risso, W. A. (2010). Hierarchical structure of the German stock market. Expert Systems with Applications, 37(5), 3846-3852.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), Advances in knowledge discovery and data mining (pp. 1-34): American Association for Artificial Intelligence.

Fernández-Alemán, J. L. (2011). Automated Assessment in a Programming Tools Course. IEEE Transactions on Education, 54(4), 576-581.

Fernández Alemán, J. L., Carrillo de Gea, J. M., & Rodríguez Mondéjar, J. J. (2011). Effects of competitive computer-assisted learning versus conventional teaching methods on the acquisition and retention of knowledge in medical surgical nursing students. Nurse Educ Today, 31(8), 866-871.

García-Mateos, G., & Fernández-Alemán, J. L. (2009). A Course on Algorithms and Data Structures Using On-Line Judging. SIGCSE Bulletin, 41(3), 45-49.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition: Springer.

Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: a machine learning workbench. Paper presented at the Second Australian and New Zealand Conference on Intelligent Information Systems.

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA. Computers & Education, 72(0), 23-36.

Leal, J. P., & Silva, F. (2003). Mooshak: a Web-based multi-site programming contest system. Software: Practice and Experience, 33(6), 567-581.

Lin, C. F., Yeh, Y.-c., Hung, Y. H., & Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. Computers & Education, 68(0), 199-210.

Mohamad, S. K., & Tasir, Z. (2013). Educational Data Mining: A Review. Procedia - Social and Behavioral Sciences, 97(0), 320-324.

Montoya-Dato, F. J., Fernandez-Aleman, J. L., & Garcia-Mateos, G. (2009). An Experience on Ada Programming Using On-Line Judging Reliable Software Technologies - Ada-Europe 2009 (Vol. 5570, pp. 75-89).

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 41(4, Part 1), 1432-1462.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

Quinlan, J. R. (1993). C4. 5: programs for machine learning: Morgan Kaufmann.

Saito, D., & Yamaura, T. (2013). A new approach to Programming Language education for beginners with top-down learning. Paper presented at the IEEE International Conference on Teaching, Assessment and Learning for Engineering.

Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 39(10), 9468-9476.

Ture, M., Tokatli, F., & Kurt, I. (2009). Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Systems with Applications, 36(2 PART 1), 2017-2026.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition: Elsevier Science.

Yang, C.-C., Prasher, S. O., Enright, P., Madramootoo, C., Burgess, M., Goel, P. K., & Callum, I. (2003). Application of decision tree technology for image classification using remote sensing data. Agricultural Systems, 76(3), 1101-1117.