Analysis of Student Performance Data from a Computer Architecture and Organization Course

Luciano Moraes da Luz Brum - Graduate Program in Applied Computing/UNIPAMPA - lucianobrum18@gmail.com
Milton Roberto Heinen - Graduate Program in Applied Computing/UNIPAMPA - milton.heinen@unipampa.edu.br
Sandro da Silva Camargo - Graduate Program in Applied Computing/UNIPAMPA - sandrocamargo@unipampa.edu.br

Abstract. Some known problems in the literature are the lack of interest and difficulty of students in Computer Architecture and Organization courses. We analyzed the students' performance in tests and semipresential learning activities in three editions of Introduction to Computer Architecture course from the Computer Engineering program of the Federal University of Pampa. We also analyzed the National High School Exam grades of these students, by knowledge area. The purpose of this study is to identify, through analytical processes using data from the course and grades of the NHSE, the most influential factors in the students' first and final grades and subsidize the proposal of effective actions to solve the problem of high percentages of students' failure.

Keywords: Learning Analytics, Computer Engineering, Data Mining.

Resumo. Alguns problemas conhecidos na literatura são a falta de interesse e dificuldade dos estudantes em disciplinas de Arquitetura e Organização de Computadores. Nós analisamos o desempenho dos estudantes em avaliações e atividades semipresenciais em três edições da disciplina de Introdução à Arquitetura de Computadores do curso de Engenharia de Computação da Universidade Federal do Pampa. Também analisamos as notas do Exame Nacional do Ensino Médio destes estudantes, por área de conhecimento. O propósito deste trabalho é identificar, através de processos analíticos utilizando os dados da disciplina e as notas do ENEM, os fatores mais influentes nas notas da primeira avaliação e na nota final dos estudantes e subsidiar a proposição de ações efetivas para a solução do problema dos altos percentuais de reprovação dos alunos.

Palavras-chave: Análise de Aprendizado, Engenharia de Computação, Mineração de Dados.

1. Introduction

The report No. 136/2012 prepared by the National Education Council (NEC), which deals with the National Curriculum Guidelines (NCG) for computer careers, provides a set of knowledge units about Computer Architecture and Organization (CAO) in computer engineering baccalaureate degree curricula in Brazil (Brasil, 2017). In the Computer Engineering (CE) program of the Federal University of Pampa (UNIPAMPA), three courses in the first three semesters approaches topics about CAO. This subject has fundamental importance in courses of computer science, considering it composes a significant part of the knowledge (Shackelford et al., 2006). In the specific case of CE, this subject is essential prerequisite for students' entry into more advanced and applied courses, which involve microcontrollers programming, digital hardware design, projects development, among others (NDE, 2017).

DOI: 10.22456/1679-1916.89249

Figure 1 presents the students' success percentages from the Introduction to Computer Architecture (ICA) course. ICA is the first CAO course that students have contact. The history of the ICA's success percentages in Figure 1 reinforces the need for a joint dedication of researchers and students to overcome difficulties in the teaching-learning process in CAO courses, especially in the introductory ones, which have a high failure rate (Woszczynski et al., 2005).

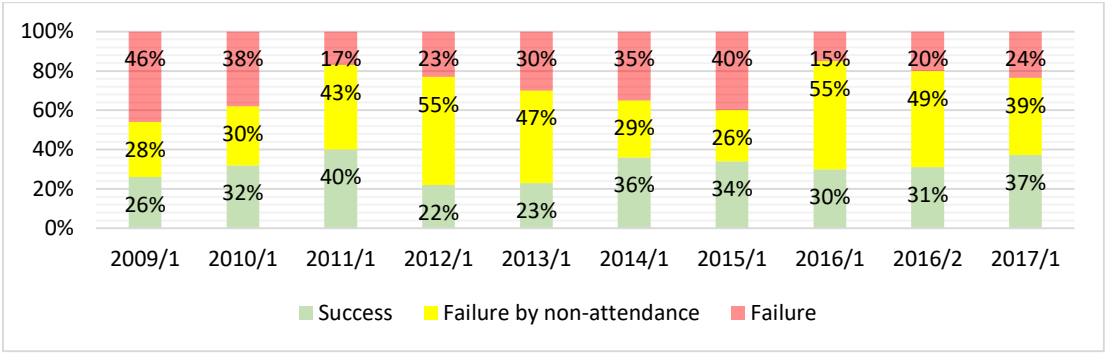

Figure 1: Percentage of ICA success. Source: Prepared by the authors, 2017.

In the period from 2011/01 to 2016/01, the same teacher taught the course and from 2016/02 to 2017/01, other teacher started to teach it. There were no significant changes in the assessment and pedagogical methods in both periods. External factors to the teaching-learning process can explain some of the variations in these percentages. These analyses are beyond the scope of this work. One important detail is that students who reproved in ICA in previous semesters compose the class of semester 2016/02.

The course of ICA has 60 hours of workload in the classroom and 30 hours of workload in the semipresential mode, totaling 90 hours. According to Report 4059 (Brasil, 2004, p. 34, Our Translation), semipresential activities are characterized as:

> [...] any didactic activities, modules or teaching-learning units focused on self-learning and with the mediation of didactic resources organized in different information support that use remote communication technologies.

Before proposing any initiative to solve the problem, it is necessary to carry out an investigation of the problem itself through an analysis of students' data from this course and detect the factors that are more influent in the success or not of these. These data may reveal some pattern of students that approve, disapprove or even those that have dropout trends from higher education.

The purpose of this work is to apply data mining techniques in performance and frequency data of the students from ICA course and in their grades in the National High School Exam (NHSE), and carry out the analysis of the results. The objective is to identify standards and extract relevant and useful information to subsidize decisions to solve the problem of high level of failure in the course. This work is in the context of learning analytics, a recent research area (Barbosa, 2015).

The rest of this work is organized as follows. Section 2 shows related works about CAO education and learning analytics in this context. Section 3 covers the entire systematic approach of the methodology adopted in this work. Section 4 shows the results generated by the algorithms, the evaluation of their accuracy and correlations between the variables. Section 5 presents the results obtained with the applied techniques and a brief discussion. Section 6 presents the final considerations of this study.

DOI: 10.22456/1679-1916.89249

2. Related works

In the context of learning analytics, Camargo, dos Santos and Camargo (2012) and dos Santos, Camargo and Camargo (2012) made a similar study in the same course and institution, but considering formative evaluations grades and other summarized variables. In addition, only one class was analyzed, from 2012. Their results showed a tendency that students' total frequency and presential frequency are more influent variables to predict the student's success in the ICA course, if compared to formative evaluations. REPTree and J48 decision tree algorithms were used to train the models.

Considering works in the context of CAO learning and education, some possible causes of high level of failure of students in this particular subject are presented below:

- The teaching-learning process based only on expository-dialogue classes and theoretical activities ends up hindering the students' ability to abstract concepts. This complicates the absorption of content and compromises the understanding of a computational system as a whole, from high level programming to the hardware level (Esmeraldo and Lisboa, 2017).
- Ristov et al. (2011) and Stolikj et al. (2011) said that had a reorientation of computer courses that favor disciplines of high-level abstraction, while those of low-level abstraction are minimized. This can cause a feeling that students do not need to know how the computer works, but rather how they can use it to execute their software solutions.
- High-level programming does not reveal how these commands are executed on the computer and, therefore, students' interest in CAO concepts is hampered (Atanasovski et al., 2013).

The teachers can solve the first cause by using tools that make the students more active in the classroom, like simulators. This approach have been used in ICA course. The other causes are more complicated to address and depends on numerous factors, being a subject to be discussed and analyzed in other works.

3. Materials and Methods

This research is descriptive, explanatory, quantitative nature and follows the single case study method (Yin, 2001). The data collection was done through documentary research (electronic spreadsheets with performance data, frequency and grades in the NHSE) and the theoretical basis through bibliographic research.

In the ICA course editions analyzed, the teacher modified the minimum of factors in his control. The three editions analyzed, 2016/02, 2017/01 and 2017/02, had 45, 51 and 43 students, respectively, and were taught by the same teacher, using the same pedagogical method, with the same presential and semipresential workload. The teacher made modifications only in the number of semipresential activities and in the weight of some evaluations, but always maintaining three written evaluations.

The Knowledge Discovery in Databases (KDD) process (Fayyad et al., 1996) is the methodology that this work is based on. In order to execute this process, all data related to the course are necessary, such as students' performance data in assessments; data on the semipresential activities performed in the course; and frequency data.

The first step was to extract IAC students' performance and frequency information from three consecutive semesters from the databases to three different .csv files. The NHSE grades of these students were also included. Due to the different amounts of semipresential activities in the different editions of the course, we applied data mining

DOI: 10.22456/1679-1916.89249

techniques separated by semester, to avoid distortions in the results. The Table 1 shows the input variables.

The second step was to import the data into the RStudio tool. RStudio is an open source Integrated Development Environment (IDE) that has libraries for statistical analysis, data mining algorithms, generation of different types of graphs, among other possibilities, through the R language.

Table 1: Input variables of the study.

2016/02	Assessments grades 1 and 3 (A1 & A3), semipresential grades (SP1 – SP6 & SP8), presential frequency and semipresential frequency (SPs).
2017/01	Assessments grades 1, 2 and 3 (A1, A2 & A3), semipresential grades (SP1 – SP9), presential frequency, semipresential frequency (SPs) and NHSE grades in each knowledge area (Natural Sciences and Technologies (NST), Mathematics (MT), Human Sciences and Technologies (HST) and Languages and Codes (LC)) and Essay.
2017/02	Assessment grades 1 (A1), semipresential grades (SP1, SP23, SP4, SP5678) and NHSE grades in each knowledge area and Essay.

The third step was the data normalization using the min-max method. Normalization allows analyzing data at different scales on the same graph without distortion, and facilitates the execution of specific data mining algorithms. Prior to the normalization process, we added samples with maximum and minimum scores on each attribute to maintain the proportionality of the data, considering that the scores in assessments can range from 0 to 10 and the scores in the semipresential grades can range from 0 to the number of questions. We made the classification in two ways:

- In three categories, without NHSE grades: Approved (grade 6 or greater), insufficient (grade between 2 and 6) and poor (grade less than 2), to classify the final average.
- In two categories, including NHSE grades: Approved (grade greater than or equal to 6) and disapproved (grade lower than 6), to classify the grade in the first assessment.

We removed from all analysis the students who did not take the first assessment, who in all cases were infrequent students (17.55% of all samples removed from 2016/02, 13.73% from 2017/01 and 27.91% from 2017/02). These samples could interfere in generated model and did not aggregate any new useful information. We also removed samples of students that had not available the NHSE grades in second analysis (22.73% of remaining samples removed from 2017/01 and 32.26% from 2017/02). The students' sex was not considered in analysis because we had a very small number of female students (16% in 2016/2, 7.84% in 2017/1 and 16.28% in 2017/2). This fact, combined with the small number of samples, turns difficult to extract a pattern that consider this information.

The fourth step was the application of data mining techniques in these two approaches. For each of them, we used two classifiers, both of decision tree: the CART (Classification and Regression Tree) and C5.0. These algorithms are included in supervised learning scope. The objective is analyze the results of the algorithms under two different perspectives to check which variables are more relevant in students' final performance and in their grade in the first assessment in the ICA course. We also checked the accuracy of the models.

Decision tree algorithms are interesting for these reasons (Han, Kamber, Pei, 2011):

- Not require knowledge of the domain;
- They are widely used in exploratory knowledge analysis;
- Work well with multidimensional data;
- The representation in the form of decision tree is simple and intuitive;

DOI: 10.22456/1679-1916.89249

- The training and classification tasks are fast and;
- They have a good accuracy.

There are more accurate and efficient classifiers, such as neural networks. However, these classifiers are often regarded as black box. Their predictions are not as interpretable as those of decision trees. The objective was to perform an exploratory analysis to verify the most relevant attributes in the students' final performance in ICA. We compared the accuracy of the classifiers and verified the correlation of the variables under study with the output variables.

Finally, the fifth step was the process of analysis and discussion of the results. In section 5, we will cover this step in more details.

4. CART and C5.0 algorithms

At first, we performed the CART and C5.0 algorithms with all ICA students' performance data for the prediction of average in the semesters of 2016/02 (37 samples) and 2017/01 (44 samples). It were used the rpart (Therneau et al., 2018), C5.0 (Kuhn et al., 2018) and caret (Kuhn, 2008) packages from R. The data from the second evaluation (2016/02) were not included because it was a seminar, not a written evaluation. The evaluation of seminars tends to be more subjective, which may end up harming the accuracy of the generated model. This was the first analysis.

In Figure 2, a graphical representation of CART results in 2016/02 and 2017/01, respectively, is shown. In Figures 3 and 4, a graphical representation of the results of C5.0 in 2016/02 and 2017/01, respectively, is shown.

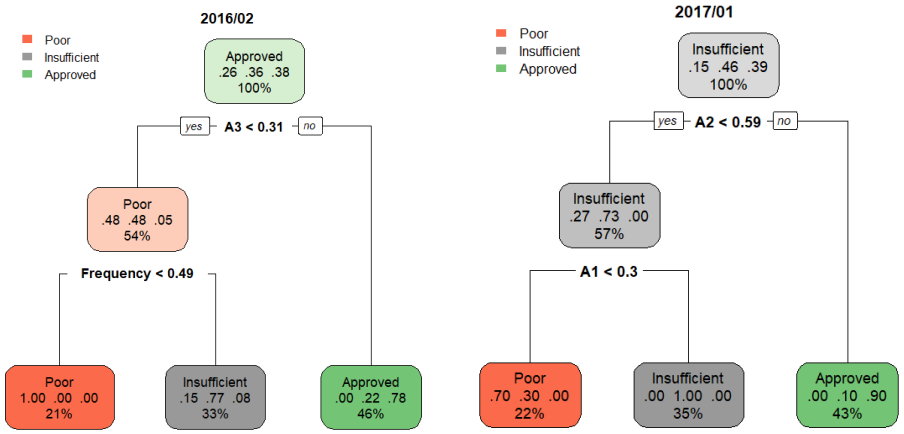

Figure 2: Results generated by CART in 2016/02 and 2017/01. Source: Authors, 2017.

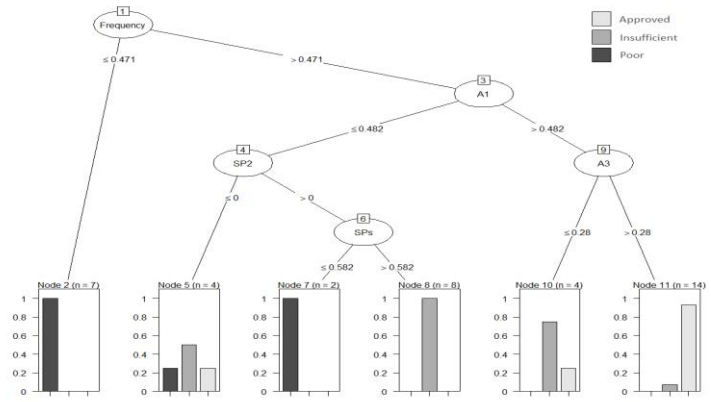

Figure 3: Results generated by C5.0 to 2016/02. Source: Authors, 2017.

DOI: 10.22456/1679-1916.89249

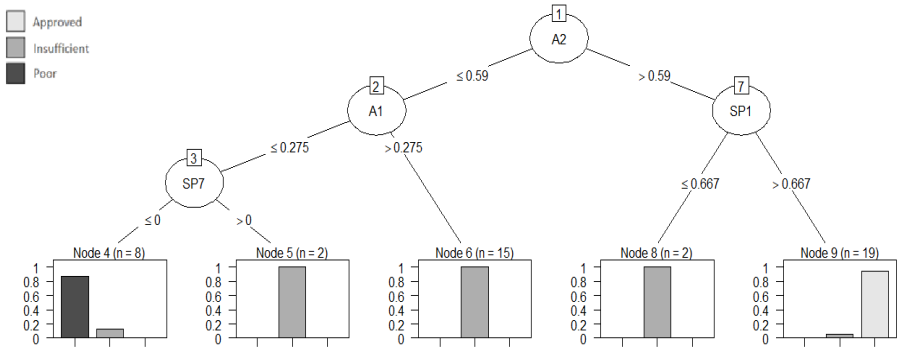

Figure 4: Results generated by C5.0 to 2017/01. Source: Authors, 2017.

In the second analysis, we performed the same algorithms with the performance data prior to the first test in the semesters of 2017/01 (34 samples) and 2017/02 (21 samples), for the prediction of student's success in the first evaluation. The NHSE grades of each student were included in this analysis, by knowledge area, to verify the importance of such indicator in the performance in the first evaluation. Some students' grades were not available, so we removed them from the analysis. Figures 5, 6 and 7 present the results generated by the algorithms.

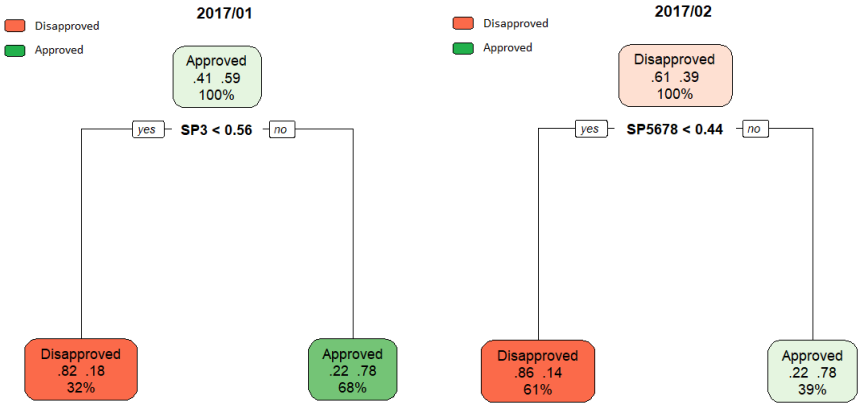

Figure 5: Results generated by CART to 2017/01 and 2017/02. Source: Authors, 2017.

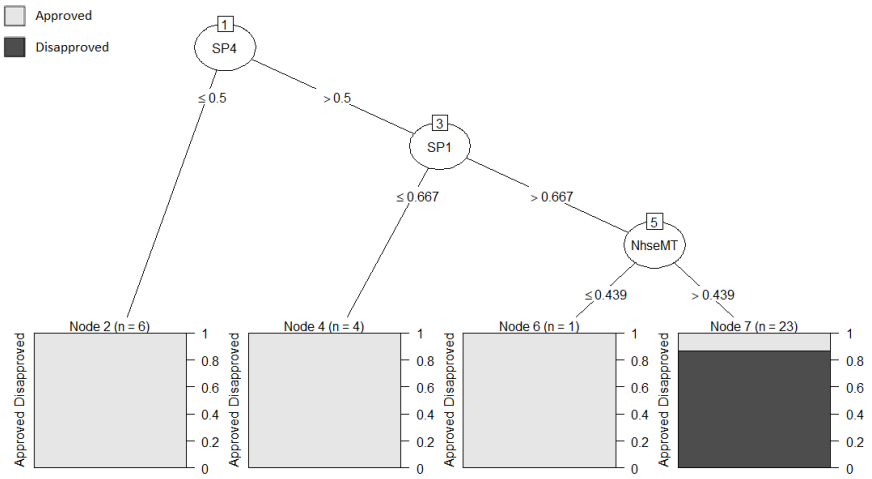

Figure 6: Results generated by C5.0 to 2017/01. Source: Authors, 2017.

DOI: 10.22456/1679-1916.89249

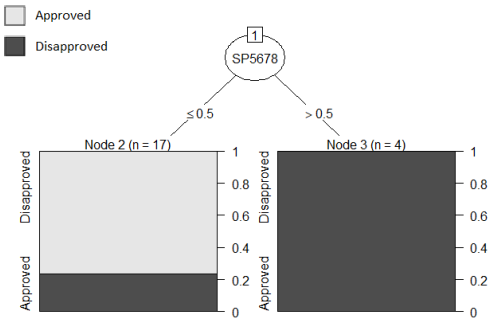

Figure 7: Results generated by C5.0 to 2017/02. Source: Authors, 2017.

In order to measure the accuracy of the models, we used the methods Leave-One-Out Cross Validation (LOOCV) and 10-Fold Cross Validation (10-CV). Both methods of evaluation produce good accuracy results. In the case of LOOCV, although computationally expensive, there are a small number of samples in this case study. Therefore, the runtime does not become a limitation.

The Table 2 shows the accuracy obtained from the models with all performance data for 2016/02 and 2017/01 and Table 3 shows the accuracy obtained from the models with the performance data up to the first evaluation and NHSE grades of 2017/01 and 2017/02. Tables 4 and 5 present the correlations of the variables with the averages of 2016 and 2017, respectively. Tables 6 and 7 present the variables correlation with the first evaluation grades in 2017/01 and 2017/02, respectively. We inserted only the most relevant correlations (up to 50% in 2016 and up to 60% in 2017).

Table 2: Accuracy of the models using LOOCV and 10-CV.

	CART		C5.0	
	Accuracy	**Kappa**	**Accuracy**	**Kappa**
2016/02(LOOCV)	61.53%	0.4067	69.23%	0.5333
2016/02(10-CV)	53.17%	0.2995	64%	0.4454
2017/01(LOOCV)	82.61%	0.7246	84.78%	0.7512
2017/01(10-CV)	77.83%	0.6461	85.66%	0.7604

Table 3: Accuracy of the models using LOOCV and 10-CV.

	CART		C5.0	
	Accuracy	**Kappa**	**Accuracy**	**Kappa**
2017/01(LOOCV)	76.47%	0.4925	76.47%	0.5142
2017/01(10-CV)	68.33%	0.34	75.00%	0.47
2017/02(LOOCV)	66.66%	0.2383	61.90%	0.1923
2017/02(10-CV)	65%	0	68.33%	0.175

Table 4: Correlation between the variables and the final average in 2016/02.

	A3	**Frequency**	**SPs**	**A1**	**SP6**	**SP8**	**SP3**
Average 2016/02	0.84	0.82	0.75	0.74	0.63	0.57	0.53

Table 5: Correlation between the variables and the final average in 2017/01.

	A2	**A3**	**A1**	**SPs**	**Frequency**	**SP2**	**SP4**	**SP7**	**SP3**	**SP6**
Average 2017/01	0.93	0.92	0.82	0.76	0.67	0.62	0.55	0.53	0.53	0.50

DOI: 10.22456/1679-1916.89249

Table 6: Correlation between the variables and the first evaluation grade in 2017/01.

2017/01	NST	SP2	MT	HST
A1	0.60	0.58	0.55	0.51

Table 7: Correlation between the variables and the first evaluation grade in 2017/02.

2017/02	SP5678
A1	0.63

4. Results and Discussions

In the first analysis, the grades in the evaluations can help to predict the students' success or failure in the course. Such an outcome was expected. This is because the subject of the first evaluation, computer numeration systems, is a prerequisite subject for a thorough understanding of the later issues. The same for the second assessment, where students begin programming in assembly language and studying computer components, which are also essential prerequisites.

We also noticed the relevance of the semipresential activities 1, 2 and 7 (conversion between numeric bases, addition in different bases and programming in assembly language), indicators that show itself as relevant in the classification in 2016/02 and 2017/01. Students who are performing poorly may be follow-up during the semester through monitoring or mentoring. We suggest a more cautious follow-up with the students who obtained an unsatisfactory performance in the first evaluation of the course, since this was an indicator of great relevance in both semesters (2016/02 and 2017/01). In addition, the relevance of frequency was noticed (Figures 2 and 3, 2016/02), matching with the results previously founded in Camargo, dos Santos and Camargo (2012).

The NHSE grades were not relevant in final grade according to algorithms used, so we proposed the second analysis, which the results between the semesters were quite different. In addition, we had a low kappa in all analysis of 2017/02. This is because, in this semester, we had less samples compared to other semesters and they are from repeating students from any semester. The data in this particular case are more heterogeneous than the data from other semesters.

In 2017/01, we noticed a great relevance of the semipresential activities that involve the numbering systems in computation, the operation of multiplication and division of binary numbers and fix and floating point systems. If the student did not perform these activities or did not obtain a good rate of correctness, it is very probable that the student will take an insufficient grade in the first evaluation, since these are the most fundamental concepts of the beginning of the course. Therefore, the teacher can monitor these students differently.

The influence of the grades in mathematics (from NHSE) on student performance in the first assessment is also apparent (Figure 6). This shows that if students have been struggling since high school, they are more likely to encounter difficulties in ICA course.

In 2017/02, we perceived a greater relevance of the grades in semipresential activities than the performance in the NHSE, in both algorithms. The model shows that last semipresential activity before the evaluation is more relevant, where are approached the subjects of multiplication and division with binary numbers and operations with fixed-point and floating-point numbers. Such issues encompass all concepts previously seen, from computer numeration systems, arithmetic operations, and conversions between numerical bases.

As the NHSE grades are external factors to the course, a more detailed analysis of their impact is required, for all courses. It is possible to measure if the performance in the NHSE is relevant in the performance of students in all courses. It is up to commissions of

DOI: 10.22456/1679-1916.89249

computation courses to take initiative to analyze and solve the problem of high percentages of failure in initial courses. One possibility is to adopt, for example, the use of different weights in the grades, by knowledge area, of NHSE for incoming students. Currently, the NHSE grades in all areas of knowledge have the same weight at UNIPAMPA.

Some important points that deserve emphasis:

- Analyzes were performed in only three editions of ICA;
- There were no NHSE grades available for all students. Therefore, we removed some samples from the analysis.
- There was no complete homogeneity in the way these courses were worked. That is why we performed separated analyses, by semester.
- Peer semesters are from repeating students, which can generate completely different results from classes of incoming students.
- The lower kappa in peer semesters is because of a greater heterogeneity between students. The models had poor agreement in these cases.

5. Final Considerations

Through this case study, it was possible to understand and analyze objectively some of the factors that influence the success percentages in initial courses of CAO in UNIPAMPA CE course. The figures 2, 3 and 4 confirms what the literature says about the students' difficulty in learning abstract concepts.

We also verified the importance of achievement in the semipresential activities and the grades in evaluations of ICA course in the student's final performance.

It was possible to verify that students' grades in the NHSE also exerts some influence in the performance of the first evaluation, a fact verified by the result of Figure 6. We highlight the importance of mathematics as indicator that exert positive influence on students' performance in the semester of 2017/01.

We suggest, as future work, the application of other data mining techniques, different approaches in the pre-processing of the data, obtain more samples of different semesters, to obtain a greater level of precision and certainty in the results obtained. We also suggest analyzing external factors that can affect the students' performance, like the adaptation to the city, travel problems, lack of student assistance, lack of knowledge from high school, among others.

References

ATANASOVSKI, B.; RISTOV, S.; GUSEV, M.; ANCHEV, N. Educache simulator for teaching computer architecture and organization. In: **IEEE Global Engineering Education Conference (EDUCON)**, 2013, Berlin. Proceedings. Berlin: IEEE, 2013, p. 1015-1022.

BARBOSA, M. W. Identificação de Experiências da Adoção de Learning Analytics no Ensino de Engenharia de Software. **Revista Novas Tecnologias na Educação (RENOTE)**, v. 13, n° 2, dec. 2015.

BRASIL. Ministro da Educação. Portaria n° 4059 de 10 de dezembro de 2004. **Introdução da oferta de disciplinas integrantes do currículo que utilizem modalidade semipresencial**. Diário Oficial da União, dec. 2004, section 1, p. 34.

_____. Ministério da educação. Conselho Nacional de Educação. **Parecer CNE/CES 136/2012, 2012**. Available at: <http://www.mec.gov.br>. Accessed in: sep. 2017.

DOI: 10.22456/1679-1916.89249

CAMARGO, F. N. P.; DOS SANTOS, H. L.; CAMARGO, S. S. Utilização de avaliações formativas no ensino de arquitetura de computadores: Um estudo de caso. **Workshop sobre Ensino em Arquitetura de Computadores (WEAC)**, 2012, Petrópolis. Anais do WSCAD-WEAC, 2012, p. 5-10.

DOS SANTOS, H. L.; CAMARGO, F. N. P.; CAMARGO, S. S. Predizendo o sucesso de estudantes através do uso avaliações formativas em AVAs. In: **Workshop sobre avaliação e Acompanhamento da Aprendizagem em Ambientes Virtuais**, 2012, Rio de Janeiro. Anais dos Workshops do Congresso Brasileiro de Informática na Educação, 2012.

ESMERALDO, G.; LISBOA, E. B. CompSim: Um Ambiente para o Ensino Integrado de Arquitetura e Organização de Computadores. In: **II Congresso Sobre Tecnologias na Educação (Ctrl+E)**, 2017, Paraíba. Proceedings. Paraíba: Universidade Federal da Paraíba - Campus IV Mamanguape, 2017, p. 697-703.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, 1996.

HAN, J.; KAMBER, M; PEI, J. **Data Mining: Concepts and Techniques**. 3° ed. Morgan Kauf. Publishers, 2011.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, n. 5, p. 1–26, 2008.

KUHN, M.; WESTON, S.; CULP, M.; COULTER, N.; QUINLAN, R. 2018. **Package "C50"**. Available at: <https://cran.r-project.org/web/packages/C50/C50.pdf>. Accessed in: Aug. 2018.

NÚCLEO DOCENTE ESTRUTURANTE DO CURSO DE ENGENHARIA DE COMPUTAÇÃO (NDE). **PPC - Projeto Pedagógico de Curso**. Technical report, Universidade Federal do Pampa, 2017.

RISTOV, S.; STOLIKJ, M.; ACKOVSKA, N. Awakening curiosity—Hardware education for computer science students. **Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics**, 2011, Opatija. Proceedings. Opatija: IEEE, 2011, p. 1275-1280.

SHACKELFORD, R.; MCGETTRICK, A.; SLOAN, R.; TOPI, H.; DAVIES, G.; KAMALI, R.; CROSS, J.; IMPAGLIAZZO, J.; LEBLANC, R.; LUNT, B. Computing curricula 2005: The overview report. **Special Interest Group on Computer Science Education Bulletin**, v. 38, n. 1, p. 456–457, Mar. 2006.

STOLIKJ, M.; RISTOV, S.; ACKOVSKA, N. Challenging student's software skills to learn hardware based courses. In**: Proceedings of the 33rd International Conference on Information Technology Interfaces**, 2011, Dubrovnik. Proceedings. Dubrovnik: IEEE, 2011, p. 339 –344.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. 2018. **Package 'rpart'**. Available at: <http://cran.r-project.org/web/packages/rpart/rpart.pdf>. Accessed in: Aug. 2018.

WOSZCZYNSKI, A. B.; HADDAD, H. M.; ZGAMBO, A. F. Towards a model of student success in programming courses. In: **Proceedings of the 43rd Annual Southeast Regional Conference**, v. 1, 2005, Kennesaw. Proceedings. New York: ACM, 2005, p. 301–302.

YIN, R. K. **Estudo de caso. Planejamento e métodos**. Tradução Daniel Grassi. 2ª ed. Porto Alegre: Bookman. 2001.

DOI: 10.22456/1679-1916.89249