Department of Information Science and Technology

Constructed Response or Multiple-Choice for Evaluating Excel Questions? That is the Question

Yolanda Vidigal Belo

A Dissertation presented in partial fulfillment of the Requirements for the Degree of Master's in Computer Science and Business Management

Supervisor:

Doctor. Sérgio Moro, Assistant Professor, ISCTE - Instituto Universitário de Lisboa

Co-supervisor:

Doctor. António Martins, Assistant Professor, ISCTE - Instituto Universitário de Lisboa

January 2018

Dedication

To my dear parents, as Rick & Renner said: "Daughter wherever you go, there may not be a place for your parents, but we are sure that we will always be by your side wherever you go”. I make these words mine "I will be the one taking you wherever I go”

Aos meus queridos pais, já cantavam Rick & Renner “Filha onde você vai, pode não sobrar um lugar pro seus pais, mas temos certeza que vamos sempre estar perto de você onde quer que vá”. Faço destas palavras as minhas “Eu que vou vos levar onde quer que eu vá” …

Acknowledgement

The biggest challenge besides writing this thesis, was having only one page to thank the people who took part on my two-year trajectory at ISCTE-IUL.

Thank you so much to my parents, who have always prioritized my education. Thank you, Mr. Belo and Mrs. Tilinha, for, besides offering me the opportunity to study abroad, you have always been present and patient, most of the time like my best friends than as my parents, and I am very grateful to God for that. My sisters Denise and Tainara, for always crying with me and encouraging me to overcome at this stage of my life, particularly for being away from home.

I express my gratitude to Professor Sérgio Moro, who guided this thesis, for his sympathy since our first meeting, for his full support, criticism and advice, but above all for his encouragement and help in achieving the project.

I thank Professors António Martins and Pedro Ramos for their trust and for being always available to answer my questions, providing me very significant pedagogical experiences.

To my colleagues Daniela Almeida, Sabrina Alves and Hugo Santos, among others, who were always present during this phase, for companionship, strength and support me in several difficult times.

Finally, to all who have been directly and indirectly involved, there is my great Khanimambo [Mozambican dialect meaning "Thank you"].

Abstract

Evaluation plays a fundamental role in education, with a view to improve the teaching-learning process, which helps to identify factors that can contribute not only to the teacher in developing pedagogical methods and evaluation tools, but also to an academic evolutionary process of the student, and to achieve the objectives defined in the course or curricular unit.

In this dissertation project, it is proposed to develop explanatory models using Data Mining techniques and tools to predict the results obtained by students in performing Excel exams, more specifically, to verify if there is a difference in student performance when performing exams with Constructed Response questions and for exams containing Multiple Choice Question equivalent to the questions of the previous format. The samples were obtained in Advanced Excel exams performed at ISCTE-IUL, to verify the difference in the exams as stated before, and identify which factors influence this, extracting knowledge from them, and using them to decision making (to assist teachers improving the exams' preparation, either by the format of the question or by the content of each one).

Using CRISP-DM methodology, the students' responses were organized in the data set, where it was used to construct 6 predictive models from regression techniques, such as support vector machines and neural networks (other identified during the research), and for training and tests errors calculations.

The results show that the SVM model is the one with better performance, indicating the MCQ format as the one in which the students are most likely to succeed.

Keyword: Essay questions, Multiple Choice Questions, Educational Data Mining, Evaluation, Support Vector Machine, Neuronal Networks.

Resumo

A avaliação desempenha um papel fundamental na educação, numa perspetiva de melhorar o processo ensino-aprendizagem, pois auxilia na identificação de fatores que possam contribuir na elaboração de métodos pedagógicos e instrumentos de avaliação, e num processo evolutivo académico do aluno, atingindo os objetivos definidos na unidade curricular.

Neste projeto de dissertação, propõe-se desenvolver modelos explicativos usando técnicas de Data Mining para avaliar resultados obtidos pelos alunos na realização de exames de Excel Avançado, ou seja, verificar se existe diferença na performance do aluno ao realizar exames compostos por questões abertas e por exames com questões de escolha múltipla equivalentes às do formato anterior. As amostras foram obtidas em exames realizadas no ISCTE-IUL com o objetivo de além de se pretender verificar tal diferença nos exames, identificar quais fatores influenciam para que isto ocorra, e extrair conhecimento a partir destes, conduzindo-os à tomada de decisão (auxiliar os docentes na melhoria na elaboração dos exames, seja pelo formato da questão como pelo conteúdo de cada uma).

Seguindo a metodologia CRISP-DM, organizaram-se as respostas dos alunos dando origem ao *data set* que foi usado para a construção de 6 modelos preditivos a partir de técnicas de regressão, algumas como máquinas de vetores de suporte e redes neuronais (outras identificadas durante a pesquisa), e para cálculo de erros de treinos e testes.

Os resultados obtidos mostram que o modelo de máquinas de vetores de suporte é o melhor dos modelos construídos, indicando o formato de exame em múltipla escolha como aquele em que os alunos têm maior probabilidade de acertar.

Palavras-chave: Perguntas Abertas, Múltipla Escolha, Mineração de Dados Educacionais, Avaliação, Máquinas de Vetor de Suporte, Redes Neuronais.

Index

List of Tables

List of Figures

List of Abbreviations

BI	Business Intelligence
CR	Constructed Response
CRISP – DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
DSA	Data-Based Sensitivity Analysis
DT	Decision Trees
ICT	Information and Communications Technology
K-NN	K-Nearest Neighbors
KDD	Knowledge Discovery in Databases
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MCQ	Multiple Choice Questions
MED	Mining Educational Data
MLP/MLPE	Multilayer Perceptron
MSA	Measurement System Analysis
MSE	Mean Squared Error
NB	Naive Bayes
NN	Neural Network
REC	Regression Error Characteristic
RF	Random Forest
RMSE	Root Means Squared Error
RSC	Regression Scatter Plot Characteristic
SEMMA	Sample, Explore, Modify, Model
SSE	Sum Squared Error
SVM	Support Vector Machine

TF	True/False

Chapter 1 – Introduction

1.1.Topic and Research Problem

Assessment is essential in education to evaluate the acquired knowledge when dealing with problems, questioning and reflection on action (França & Amaral, 2013). Therefore, it plays a fundamental role in promoting learning, producing information that can help students and teachers. Thus, assessment is not merely an instrument of certifying learning but one which acts directly in the process of teaching and learning, permeating it and aiding it as if it was an activity at any one moment (Cerny, 2001 cited by França and Amaral, 2013).

It should be noted that the evaluation of student learning is one of the critical components of the educational process. If it is used in an appropriate way, it can be a decisive factor for achieving the objectives of the subject, or even of the course. Otherwise, it may put at risk any efforts to innovate and improve the quality of pedagogical methods and techniques, since the tests on the one hand are a source of motivation and evaluation, on the other, students will tend to study only what they believe will be asked in the tests (Camilo & Silva, 2008).

Thus, there are several test models, from the most traditional paper-based ones up to electronic format (using computer materials), composed by questions requiring Constructed Response (CR) where questions are directly asked, also called essay questions/ open questions/ open ended and for Multiple Choice Questions (MCQ) with the presence of several alternatives with only one of them is correct, in true-false format, open space and so many other ways to express this type of test format.

Nowadays, many teachers tend to pass assessments on paper, using Constructed Response (CR) to electronic platforms using Multiple Choice Questions (MCQ) as a means of evaluation, highlighting the **central problem of this study**: when students use MCQ, do they obtain the same results (identical) as CR? So, if the results are not identical, what would be the reasons to explain this discrepancy? These questions compose a sample of tests from the academic year 2016/2017, carried out on the curricular unit of Advanced Excel of ISCTE-IUL.

For example, Kuechler and Simkin (2003) consider that, since most teachers have a greater preference for CR over MCQ, the fact is that students with a high level of performance in the subject must assess the questions themselves, a task which is more protracted than MCQ which requires more subjectivity (Zeidner, 1987, cited by Kuechler and Simkin, 2003). If students' responses to MCQ correlate with CR results and such a relation is sufficiently high, lecturers may conciliate both methods, opting for MCQ, should they so choose, for example in exams in which there is a high number of students in the group, and the questions are easy to classify, and results can be given equitably.

On the other hand, it is necessary to question how converging is the commitment of the student in MCQ in relation to the commitment of CR. For these and other reasons, for many years researchers have tried to respond to these types of questions, in a way to take advantage of the teaching-learning process in its entirety, both for students as well as teachers.

1.2.Topic Motivation

Teachers in academic institutions can use a large variety of testing methods to assess the student concerning the topics in the course or the key curricular unit, including MCQ, CR, filling in in blank spaces, essay question or experimental and observations methods (Miranda, 2015). This author also considers that, usually, tests and/or exams tended to be composed just of CR, so that students created their own answers, instead of selecting the correct answer from a set of alternatives offered; however, currently the scenario is different, with some tests being composed of both models for questions, others only of MCQ.

As previously stated, in using MCQ, would the results be the same, or would there be any difference with the results obtained from assessments done by CR? Still working with the same analogy, would it be possible for a teacher to measure objectively the learning of his students if the choice of one method or another (on the same content), and would it have an impact on the involvement of the students and on the results at the end of the assessments?

In fact, concerning the quality of the assessments, the formulation of questions, the pedagogical quality of the teacher and all the factors involved in the teaching-learning process, there would be a greater flexibility if the teacher had help in the choice of one of the testing methods, if the model chosen is adapted to the intended evaluation technique and stimulates the part played by students, whether in CR or in MCQ.

1.3. Research Objectives and Methodology

The choice of the research theme to be elaborated revolves around the following investigative question to be answered as the work progresses:

Is there a difference in assessing by CR or MCQ on the same subject in terms of the results performed by the student? That is, in using Multiple Choice Questions (MCQ), do students obtain the same (or similar) results to those obtained by Constructed Response (CR)? And if these are not identical, what are the reasons that can explain this divergence? This question is associated to a generic objective analyzing to what point there is a discrepancy in the performance of students assessed by MCQ and by CR in the same content. Still on this objective, the following tasks are highlighted:

- Measure the existing discrepancy between the results obtained by students from CR and MCQ tests;
- If necessary, explain the possible reasons that may have influenced the existing discrepancy.

For obtaining of results and optimizing research, the samples' data will be analyzed using Data Mining techniques (this method will help the execution of the first task previously described). The intention is to obtain an explanatory model that illustrates the presence of variables relevant to the study, the implication of each one of these in the results obtained, and to confront similarities and differences in the two assessment methods, and with the aid of revision of the literature, one will be able to reach the second task associated with the principal objective.

Therefore, to validate the study, the results are presented using Data Mining techniques from data collected from a sample of 300 Excel Advanced assessment tests in which 50% are open and 50% are MCQ carried out at ISCTE-IUL. The questions are paired, that is, a topic of the subject is used in the two assessment methods in the same test.

1.4. Structure and Organization of the Dissertation

The present study is organized in four chapters that aim to reflect the different phases as far as their conclusion.

The first chapter introduces the theme of the investigation and contextualization of the problem, the motivations and relevance of the research, as well as the description of the objectives and corresponding methodology.

The second chapter reflects the theoretical context, designated revision in the literature, involving the synthesis effected from the researched literature, grouping the theme with its respective involving aspects, from a vision of teaching and academic assessment in a general sense, followed by the educational objectives of Bloom, details related to CR and MCQ, while certain concepts related to teaching will be explained, creating a bridge with that learnt using an Excel Spreadsheet. Afterwards, study cases related to the same theme of the present research will be approached, ending with a theoretical approach to the concepts of BI, Data Mining and respective application techniques and modulation of data.

The third chapter is dedicated to the Methodology used in the process of gathering and treatment of data, obtaining results, description of research cases and of all the work done, as well as the tools and techniques used.

Finally, the verification of results obtained during the study carried out, the conclusions that can be arrived at, and what the pertinent questions are, limitations and contributions of the dissertation and a declaration of the starting point for future projects.

Chapter 2 – Prior Literature

2.1. Teaching and Academic Evaluation

In the last decades, there has been an increase in the student population at all levels of education, and with this, educational problems related to the high number of pupils are common in several geographical areas, which does not dispense the university level as well. Solutions to these problems have been sought through diversified education systems, whether by individualized teaching or by group teaching. Related to these systems and their constant changes and updates, we find the evaluation, which can be from a set of methods to test performances, promote learning or even add grades to students (Buchweitz, 1975).

Starting from "*Guia De Elaboração E Revisão De Questões E Itens De Múltipla Escolha* (2011)[1]", learning is a cognitive process, inherent to the human being, but not observable directly. To evaluate it, it is necessary to have visibility, being one of the roles of assessment instruments, such as school tests, which act as stimulants whose function is to provoke responses that are the expression of learning and manifestation of the knowledge and skills that constitute it. In fact, knowledge assessment is an important aspect of the educational process to determine the extent to which learning outcomes are achieved (Čandrlić, Katić, & Dlab, 2014).

However, it is important to consider the clear and structured definition of educational objectives, since the acquisition of knowledge and skills appropriate to a professional profile to be acquired should be directed from a teaching process with adequate choices of strategies, methods, delimitation of specific contents, assessment tools, and consequently it will lead to effective and lasting learning (Ferraz & Belhot, 2010).

As a result, will it be possible to plan an evaluation and its set of objectives so that the result reflects something more than a simple memory of what is written in the manuals or was mentioned in the classes? To answer this question, it is necessary to know which the highest thought processes are, as well as the specific measurement methods of these

[1] Manual available in http://www.adventista.edu.br/_imagens/area_academica/files/guia-de-elaboracao-de-itens-120804112623-phpapp01(3).pdf

processes that can be evaluated in the school environment (Pinto, 2001). This requires a typology of processes and objectives of learning, among the several, it stands out that the greater of the academic contribution to educators over time, is the Bloom Taxonomy (Ferraz & Belhot, 2010).

2.2. Bloom's Taxonomy Applied to Evaluation

Bloom's Taxonomy is one of the instruments to help the identification, declaration and control of educational objectives linked to a set of processes from the acquisition of knowledge, skills and attitudes, to the planning of teaching and learning, although few educators take the maximum advantage of this tool because they do not know an appropriate way to use it (Ferraz & Belhot, 2010).

In fact, it proposes a more effective way for educators to dominate their abilities/capacities and to draw strategies from the simplest to the most complex ones, aiming that the student also acquire specific skills or to be required without distancing himself from the previously proposed instructional objectives (Ferraz & Belhot, 2010). They also consider that, although the three domains (cognitive - encompasses intellectual learning, affective - related sensibility and values, and psychomotor - skills of performing tasks that involves movements and actions) have been widely discussed and disseminated by several researchers, the cognitive domain has been the best known and used when addressing issues related to teaching and evaluation.

However, Bloom's Taxonomy suffered some revision (despite being used for four decades) in 1999 by Anderson (2001) cited by Ferraz & Belhot (2010) and his collaborators, who made a retrospective to the original version, changing some existing categories and maintaining a balance with the new positions declared, as you can see in Fig. 1, the levels organized in hierarchy.

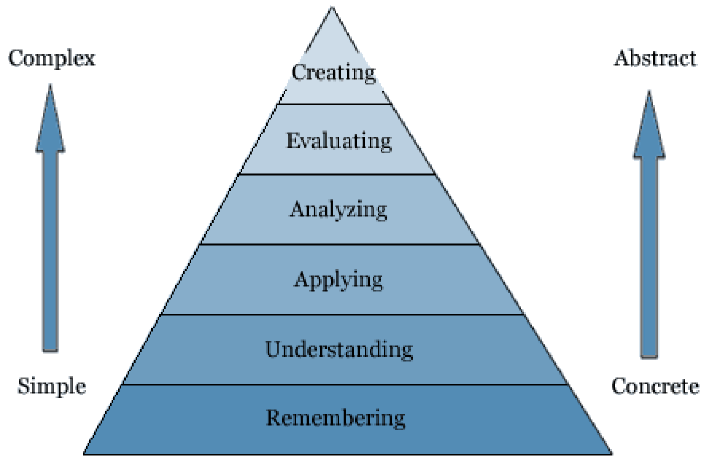

Figure 1: Bloom's Taxonomy. [Adapted from: Using a Learning Taxonomy to Align Your Course][2]

In this case, in the view of Salume et al. (2012) this taxonomy is intended not only to be a classification of the behavior expected by the students, as they must think, feel or act, but also brings with it other advantages according to Ferraz and Belhot (2010), such as:

- Provide a basis for the development of differentiated assessment strategies and instruments, stimulating students' performance and their acquisition of knowledge;
- Encourage educators on helping their students in a structured and consistent way, the process of acquiring specific skills linked to their simpler skills and transition to the most complex domain.

Note that, related to the taxonomy categories, the first four Bloom´s objectives (knowledge, understanding, application and analysis) can be applied in MCQ format exams, while the last two (synthesis and evaluation) would be better evaluated in essay questions format, not discarding the possibility of being also indicated for the previous objectives (Gronlund, 1998, cited by Salume, et al. 2012) (see appendix 4 for more details).

---

[2] Available in: http://www.ucdenver.edu/faculty_staff/faculty/center-for-faculty-development/Documents/tutorials/Assessment/module2/index.htm

2.3. Constructed Response *versus* Multiple Choice Questions

Evaluations by CR and MCQ are particularly included in a teaching - learning process in which a student is submitted throughout the school process.

Generally, CR have been the preference than MCQ by several educators, based on the belief that the first method measures a greater number of skills (and even the most important) on students' comprehension and ability to apply their knowledge, while MCQ reflects less cognitive aspects regarding the application of knowledge and the art of producing a response according to Chan and Kennedy (2002), although there is a frequent illusions on students in choosing MCQ than CR because they find it more "easy" to solve.

However, these can become complex, for example, when the number of alternatives is higher and the distractors very similar to each other.

MCQs are usually referred to as "objective" tests, although the purpose of these tests is limited only to correction systems that are quite reliable in relation to alternative systems that can be done "mechanically" (Pinto, 2001). These differ in their most common type of presentation: questions with 3 to 5 options (only 1 correct); paired items; double choice TF items (true / false); questions with spaces to complete, and much more. Particularly in this study, the form used by the samples are the questions with 4 options.

While the CR consisting of a set of tests, in the form of a question or statement, the student is asked to answer in writing by evoking knowledge and proceeding in greater or lesser degree to an analysis, description, explanation, comment and synthesis of a content, topic or subject (Pinto, 2001). Note that the number of cognitive skills is greater and memory recall by evocation can be also more complex and difficult, for example, in daily tasks it is easier the familiarity of a face and a voice than the memory of the person´s name (Pinto, 2001). Pinto (2001) also considers that a MCQ test is apparently a simpler task by making an analogy to the simple recognition of people, beings, objects, and events whose familiarity or lack of it is soon detectable.

Therefore, it is necessary to describe and analyze in more details the advantages and disadvantages of one type of evaluation in relation to the other, that is, multiple choice evaluations regarding the alternative for constructed response.

2.3.1. Constructed Response

CR brings an indirect benefit, since its content can provide suggestions to teachers about what is important to teach and to students about what it is important to learn, according to Lukhele, Thissen, and Wainer (1994). In fact, stimulation for organization and cooperation, propitious environment for comments, clarifications, examples and explanations that allow content analysis and interpretation, are some beneficial aspects of CR. In this case, the student only produces a response (instead of selecting an option from a set of items), whether short, explanatory, filling some blank spaces, represent graphs, among others, although in some cases there is difficulty in expressing the correct information to be provided to the examiner.

However, even for Lukhele, Thissen, and Wainer (1994), many students dislike open-ended assessments, since they require a high level of skills to organize, framing responses, formulate ideas, knowledge about the subject area, and finally, a convincing answer that correctly expresses what is questioned and that does not leave the subject, although, giving a wrong or not so convincing answer, the examiner can deduce the score of the question. In this sequence, there is no clear standard of possible answers, and there is great difficulty in interpretation and subjective compilation by the teacher (because they consume more time to be analyzed in relation to the MCQ), then the CR carries limitations to ensure a uniform and quality evaluation, especially if there are several teachers.

Also, another disadvantage of CR is related to the fact of the large amount of time needed to elaborate and classify, thus adding a cost to the examiner's time and effort according to Simkin and Kuechler (2005). In the opinion of Hickson and Reed (2009), although the evaluation of the student's answers in CR may consume time to the examiners, they should not be neglected, since it is possible in this way measure the understanding of the subject and the students' abilities to solve some types of problems and express ideas in writing.

In this way, according to Čandrlić, Katić, and Dlab (2014) with online tests, there has been a transition from evaluations on paper - based and using Constructed Response, to electronic platforms using Multiple Choice Questions tests as their evaluation tools, returning to the problem of this research, being necessary understand first how the electronic platforms teaching, or evaluating, and what repercussions or advantages they will have on student learning.

2.3.2. E-Learning

Information and Communication Technologies (ICT) raise challenges and at the same time offer to the teachers numerous tools that allow the creation of differentiated learning opportunities for students as reported by Azevedo (2017). Concerning the use of ICT in the evaluation process, this author considers to be an added value, in which they are used throughout the evaluation process from the design of the tests to the correction and storage of the results, through the electronic format/e- assessment, where the application of MCQ (despite portfolios or discussions) is verified.

Furthermore, a major advantage of e-assessment is the automatic correction that makes work easier for teachers (Hickson & Reed, 2009), obtaining privilege in relation to CR in teacher's involvement as stated above. On the other hand, this form of evaluation carries with it the guarantee that the tests are fair and do not harm the student in the evaluation process, when it comes to the scoring factor for example. In contrast, Čandrlić, Katić, and Dlab (2014) also adds the fact that the subjective factor of the teacher plays a crucial role in the evaluation of the answers coming from CR, since it has the possibility to evaluate this as partially correct or to identify the "hidden knowledge" in the given answer, although it suggests a form of combining online tests with the at least 30% of subjective questions.

A research based on the analysis and comparison of results based on traditional paper and online tests that the students solved in computer held in the Department of Informatics of the University of Rijeka by Čandrlić, Katić, and Dlab, (2014), had the objective to determine if and which online tests can replace the traditional paper-based ones. Using a model based on MudRi, objective and subjective (test and short answer) questions were used in evaluations in the courses of Data Modeling, Process Modeling and Information Systems, in which the results obtained for the first two courses were that, there was no significant difference in the mean values of both types of tests, whereas for the last course (consisting only of objective questions), the scenario was different, with better results in the tests performed online. Consequently, it was concluded that online tests can replace traditional paper-based tests to assess student knowledge but giving special attention to their composition.

By choosing the questions that will be used in an e-assessment test, the teacher can define a fixed set of questions, allowing the system to randomly choose items from a category or combination of several, being appropriate grouping by chapter, difficulty of question, or other criteria (Hickson & Reed, 2009).

Considering the above, E-Learning systems must be based on objective answers due the limitations described and the researches performed until then, giving rise to the evolution and increasing use of the multiple-choice questions (MCQ).

2.3.3. Multiple Choice Questions

The use of MCQ dates to the beginning of the 20th century, with Azevedo (2017) stating that even before the existence of the e-assessment, in which the objective was to reduce the ambiguity and consequent differentiation in the evaluations made by the teachers for the students. Since then, these have been gaining space in research fields, bringing with them advantages in relation of the limitation on the use of CR in terms of objectivity and consistency.

In this follow-up, Xu, Kauer, and Tupy (2016) in their research showed that there are ways to optimize the construction and use of MCQ to benefit the instruction and assessment in classrooms, learning and student performance, and yet using instructor's time and energy more efficiently.

In addition, the use of MCQ in the format of evaluations, are easy to apply and analyze because they do not require elaborate student responses as it happens in the CR, they offer quick response (depending on the degree of difficulty of the question), present several options that ends on assisting the student in choosing the best alternative through a process of eliminating alternatives that do not seem very satisfactory until it is only approximately one correct, and consequently can be objectively registered and classified by the teachers. However, concerns about academic dishonesty are common in these cases in relation to CR, since it is simpler for students to copy responses (from another student, book, or online resource), but this problem can be reduced by using alternative forms of testing, with variants, paired questions, alternating seats (if the test is realized in the classroom), among others, as reported by Xu, Kauer, and Tupy (2016). Sometimes there is not only a limitation in the expression of ideas and concrete and ideal solutions, but also these kinds of questions are susceptible to guessing.

Therefore, in addition to its ability to evaluate appropriate cognitive capacities, any assessment tool should be able to withstand content examination and build validity, reliability, fidelity, and at the same time discriminate the student's performance levels. Thus, a well-constructed, peer-reviewed MCQ is suggested to meets multiple educational requirements and is, above all, is considered a serious format to evaluate student (Palmer & Devitt, 2007).

There have been many discussions recently about the best choice for evaluation method, by CR or MCQ, in which the latter is preferred, in cases where, for example, the damage in the CR process is considered for both students and teachers. In fact, if there is an anticipation of which objectives an evaluation wants to achieve, whether by CR or MCQ, dealing into its advantages and disadvantages, both methods are effective, but the latter is still preferred because it provides greater objectivity to an evaluation and making it trustworthy.

2.3.4. Constructed Response versus Multiple Choice Questions: where do they differ?

In educational literature research on MCQ versus CR appears in the form of various debates generating controversy from the results of assessments that contain MCQ while for other students obtaining more positive results when they undergo assessment by MCQ, and finally authors that defend traditional CR, tending to show greater capacity of knowledge in relation to the previous.

On the one hand, comparing the results of categories MCQ and CR in the point of view of Simkin and Kuechler (2005), there is difficulty in constructing MCQ that reach a high level of learning in relation to the CR when referring to the level of application of Bloom's Taxonomy. In addition, his research concludes, the results tend to be positive in MCQ if these are developed around the understanding of the level of taxonomy, and CR respectively for a higher level as well as for application to higher levels. On the other hand, according to Martinez (1999); Hancock (1994) cited by Simkin and Kuechler (2005), and Kastner and Stangl (2011), CR and MCQ measure the same level of knowledge only in the first four dimensions of the taxonomy (knowledge, understanding, application and analysis), the difference being revealed in the last two categories, in which the degree of difficulty is greater.

In this point of view, Miranda (2015) also sought to analyze test formats, comparing MCQ to CR in computer and paper on the performance and satisfaction of a sample of 31 students of the Universidade Federal de Goiás coursing Administration. Therefore, the results that showed the differences between the means of presentation (by computer and paper) are small, but between the formats (MCQ and CR) is evident, presenting the best results in favor of the format of MCQ performed in paper followed by the MCQ computer model.

According to the empirical studies of Hickson and Reed (2009), using a set of data compiled from Introductory Microeconomics and Introductory Macroeconomics classes in the years 2002-2007 with approximately 8400 students, the CR does not measure the same as the MCQ, because for these authors, the student’s performance in a subsequent exam in the same course, and the academic performance in other courses, are factors that influence significantly so that there is a difference in the results obtained by CR and in MCQ, since the CR provide extra information to the student’s perception, as well as in the correction for the teacher that is not contained in MCQ for example. In the same way as Čandrlić et al. (2014) and Pinckard et al. (2009) cited by Mullen and Schultz (2012), these authors also propose the combination of CR components with those of MCQ to achieve better results and a deeper learning than just evaluations performed only in MCQ.

Sheaffer and Addo (2012) conducted a research on the Pharmaceutics II Course at the Bernard J Dunn School of Pharmacy at Shenandoah University, with the object of measuring and comparing the performance of students in CR and MCQ in exams done in this course, being possible, on the one hand, to conclude that they responded more accurately to MCQ in relation to CR (77.4% vs. 70.4%) and they felt more confident respectively, and on the other hand, that there must be other methods and techniques that can be added to the previous CR, in such a way as to improve learning in pharmacy education.

The work of Chan and Kennedy (2002), on data collected from the two types of exam done by 196 students, using the t-statistics method on expected difference, concluded that students have higher results in MCQ than in the equivalent assessment by CR and, in an unexpected difference, they measure the same. It is important to stress here that the main conclusion was that for certain types of MCQ, students have better results than those in the equivalent CR, even if the adjustments have been done by guesswork, memory or deduction.

However, Buchweitz (1975), in his studies, compared the results of MCQ with those of CR applied to the students of the General Physics course in UFRGS, concluding that there is no significant difference between evaluating by the first type or the other. Initially, the comparison of the results of the average of the tests was performed using the t-test (Spiegel, 1971 cited by Buchweitz, 1975), because the population presented a normal distribution, concluding from the studies, that there was no difference in the evaluation method has said previously for all educational levels described by Bloom’s Taxonomy, although considering a level of significance below 1%, the difference of the means tends to increase in favor of the MCQ and finally, both types of questions are considered good instruments of different levels of behavior and learning in General Physics. Additionally, similar studies that consider that there is no difference between CR and MCQ, or even if they evaluate the same were performed by Ackerman and Smith (1988); Van den Bergh (1990); Wainer and Thissen (1993) cited by Kastner and Stangl (2011).

It also happens to the conclusions of the Kastner and Stangl (2011) survey at the Vienna University of Economics and Business in order to compare the CR and MCQ tests using three different scoring rules, such as NC (Number Correct - only correct responses are counted while incorrectly ignored, NA (All-or-Nothing) and US (University Specific - subtracting a of each false alternative) and aided by the FACETS software and the MFRM approximation, the results indicated that both tests are equal, that is, they measure the same thing when it is considered the NC rule, since it does not penalize the wrong answers and considers them to be partially correct, which does not happen with other rules, so CR tests can be replaced by MCQ in these cases.

Given the above, there seems to be no consensus as to the best method to best assess the student’s learning, considering the opinions of each author supported by his research, particularly this research, which aims also to measure the distance between the types of questions; the samples are based on Excel Advanced assessment tests in which 50% of the questions are open and 50% are MCQ, being necessary therefore an exposition on this type of teaching material.

2.4. Teaching and Excel Learning

Spreadsheets are commonly used, whether in accounting, health, marketing or in areas requiring a little more programming, such as engineering, which include a set of design activities, documentation, debugging, testing, maintenance, storage and qualities (Maresca, 2016).

Maresca (2016) also adds that the most commonly used calculation tool every day is Microsoft Excel, although its users only use a fraction of the many Excel features, it has a very strong impact on the way companies apply in their business.

As an example, the work of Nossa and Chagas (1997) focused on the usefulness, efficiency and effectiveness of the linear programming technique demonstrating with practical examples of different situations in the daily life of professionals in Accounting for decision-making purposes, as these can benefit this technique especially when associated with the use of spreadsheets, specifically the use of the "SOLVER" MS Excel feature.

Thus, in an educational formation that involves a combination of practical knowledge and abstraction, using the computer and Excel, Silva (2009) states that there is a contribution to the establishment of an educational process that allows both the student to understand about the importance of knowledge as a new process of evaluation that allows the replacement of calculator, paper, pencil, pen and other materials, where the student faced with situations and problems, will learn to develop strategies that acquire spirit to research, experiments , data organization, systematization of results, validation of solutions as well as the expansion of new knowledge.

For example, Denari, Saciloto, and Cavalheiro (2016) in their article, evaluate Excel as a teaching tool in the discipline of Qualitative Analytical Chemistry for higher education students using calculations of concentrations of species in equilibrium in acid solutions, with the intention to observe some form of learning, where the students criticize the data and correlate it with the graphs also generated by the spreadsheet and other computational forms. In a critique and repetition process, where the graphs with conceptual or formal errors were returned for correction, the authors noticed a significant improvement at each iteration, and then arriving at the conclusion that Excel proved to be a motivating software for the content of the subject and an alarm clock for learning interest in students, although it is little used in teaching.

According to Cherinda (2016), on a survey carried out to a group of students about satisfaction with the use of the Excel tool, they realized that the tool has many functionalities that allow them to manipulate formulas, identify variables, and managing academic work. Nevertheless, for a sample that did not have a conscience of the potentiality of Excel, these students confirmed by the questionnaire that they were developing positive attitudes regarding the model and form of learning in the course in which they were enrolled, more concretely, in the Mathematics and Statistic calculations.

In this perspective, Cymrot (2006) used Excel to analyze students' understanding of content, learning and ease of calculation in the discipline of Statistics for Engineering II by using some statistical techniques commonly used in the Six Sigma quality management program, such as Multiple Regression, Statistical Quality Control, Measurement System Analysis (MSA) and Factorial Experiments, concluding that there is no difference in the learning of these techniques, considering the use of Excel as imperative, but the latter demonstrated a different behavior, dissociating from the rest. However, students consider it important to use Excel in the topics covered by making calculations easier.

2.5. Data Mining and Knowledge Extraction

2.5.1. KDD Process and Data Mining

Since the earliest times, from the massive use of paper to the present day in the age of digitization, companies have large amounts of data and information, stored either physically (as the first scenario) or as using (devices, tools, software, etc.). Quintela (2005) believe that, the process of extracting information and knowledge, to create new strategies and to continue business operation, consequently to decision-making, an aspect that has been considered an essential element of BI. Thus, the term BI is commonly defined as a process of collection, organization, analysis and all other information management that supports the organization's business, and since there may be scenarios with an immense amount of data, where the Data Mining arises (as a step of the KDD) to make this process efficient.

The terms KDD and Data Mining are confused in some literatures, described as being synonymous, but most of them points to Data Mining as one of the activities of the KDD, since the first one is related to a robust process of development of methods and techniques that help in the extraction of knowledge, while the DM phase encompasses the process of data visualization, with the objective of automatically inferring models and rules that have an implicit knowledge of the data studied (Quintela, 2005).

In this sequence, Thomé (2017) consider also that Data Mining as an activity of the KDD, and by the latter being known as interactive, involving several loops within a same stage and going through phases until a result is concluded, there is a need to graphically represent and describe all KDD steps including Data Mining, always remembering the distinction between the two terms.

According to Quintela (2005), these are the phases of KDD process (see Fig. 2):

- Selection: where there is a collection of useful data after a definition of the purpose and objectives of the work;
- Pre-processing: the whole process of transforming the data, eliminating the noise, omissions, errors, etc.;
- Transformation: identification of variables with greater value, as well as modification of the same in a way that becomes important;
- Data Mining: selection of appropriate methods and algorithms to extract data patterns;
- Interpretation and Evaluation: by visualization of the knowledge, and aid of graphs/ diagrams or another form of representation, make possible its interpretation, and evaluation, finishing the goal or initiating a new iteration.

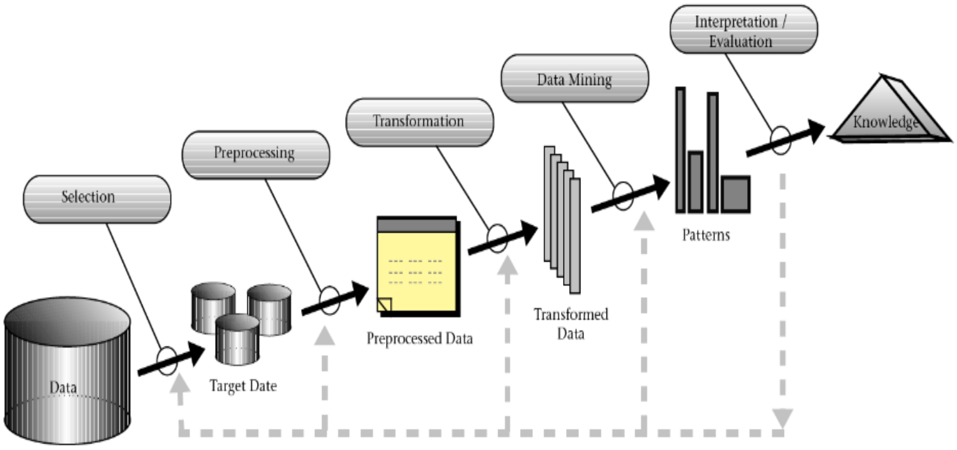

Figure 2: KDD Processes. (adapted Quintela, 2015)

2.5.2. Data Mining Methodologies

Currently, there are two main Data Mining methodologies: CRISP-DM (Cross Industry Standard Process for Data Mining) and SEMMA (Sample, Explore, Modify, Model) as the methodologies in which projects in this area have been most developed (Nogueira, 2014 cited by Almeida, 2017), which aid in organizing phases to achieve results, since the DM has an iterative life cycle, and its phases do not have a rigid sequence, only dependent on the result of each phase.

CRISP-DM

The CRISP-DM is a model that has iterative processes, of non-mandatory sequence, having also a life cycle (Fig.3), which occurs in the phases that have their tasks respectively (Fig. 4).

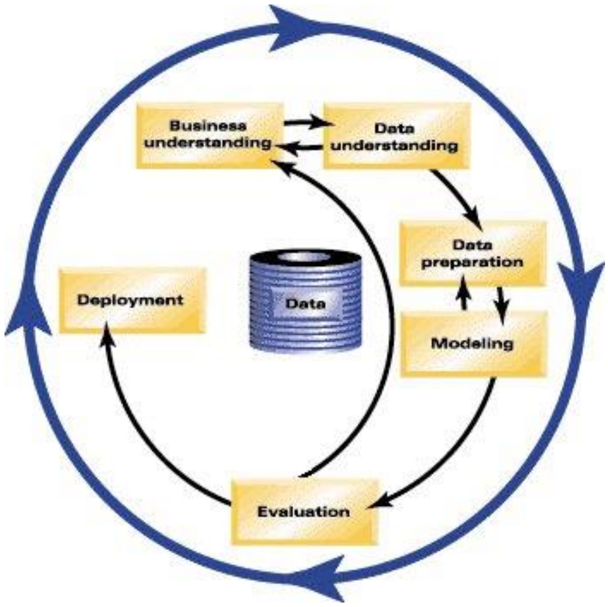

Figure 3: CRISP - DM Phases[3]

[3] Source: https://decisionstats.com/2013/04/10/visual-guides-to-crisp-dm-kdd-and-semma/

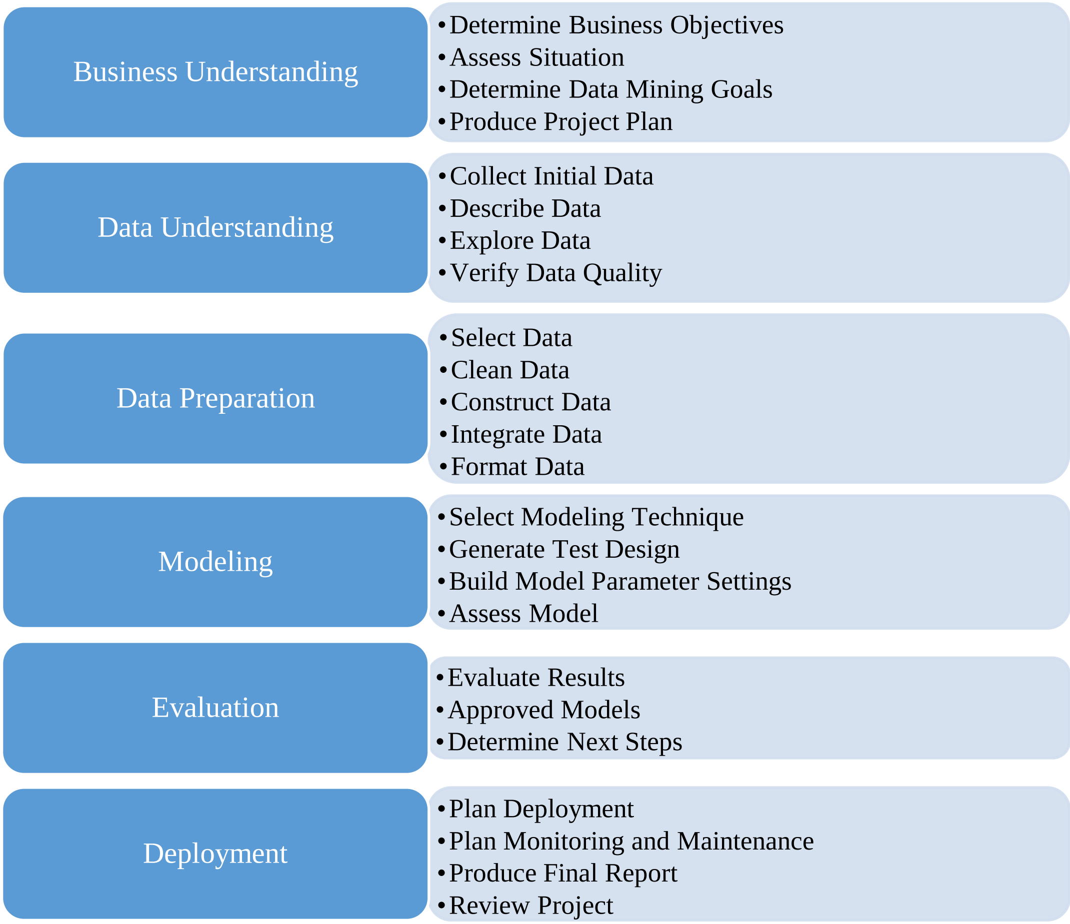

Figure 4: CRISP - DM Phases and Tasks[4]

The first phase of this methodology, according to Quintela (2005), is Business Understanding, which seeks to understand the objective to be reached with the DM and serves as a starting point for next phases. Next, Data Understanding, identifying the relevant data for the case, which problems exist and formulation of hypotheses. It follows, Data Preparation, which involves the cleaning of the data, combination of these and treatment of errors, in short, the production of the dataset to be used. The Modeling phase, involving the modeling techniques according to the initially defined objective. Then the Evaluation of the results obtained, where several graphical tools assist in the visualization process, and finally Deployment to produce reports and turning knowledge accessible to the others involved in the process.

[4] Adapted in: https://decisionstats.com/2013/04/10/visual-guides-to-crisp-dm-kdd-and-semma/

SEMMA

The SEMMA methodology, developed by the SAS Institute (which defines DM as the process of extracting valuable information and complex relationships from large volumes of data), divides the DM process into 5 stages in which the nomenclature composes the SEMMA acronym: Sample (step where a sample is selected, which corresponds to the subset of data belonging to a universe in which the assumptions of completion of each element must be the same), Explore (exploitation of data with the help of techniques, the search for trends unforeseen and anomalies on data), Modify (transformations necessary to correct the anomalies of the previous phase), Model (according to the defined objectives and the results expected to be achieved, Data Mining techniques are applied), and Assessment (evaluation of the model performance, presentation of the test results in the data and completion of all work done), which correspond to a cycle, where internal tasks are performed repeatedly until the objective is verified (Quintela, 2005).

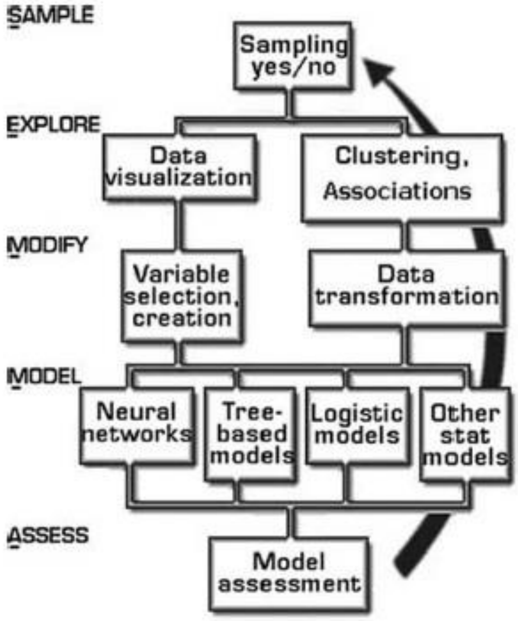

Figure 5: SEMMA Methodology (source: Ohri, 2013)

In this work, CRISP-DM was chosen as the model because it is considered one of the most used and widely accepted methodologies, as well as an extensive literature available on the model (e.g., Moro et al. 2011).

As stated by Quintela (2005), there is no universal Data Mining model that efficiently solves all problems (Harrison, 1998 cited by (Quintela, 2005)). The choice of an algorithm is somehow an art (Fayyad et al, 1996 cited by (Quintela, 2005)), since there are different models for the same tasks of DM with intrinsic advantages and disadvantages, and it is necessary to choose the techniques according to the objective of Data Mining that we intend to solve the problem.

2.5.3. Methods and Techniques of Data Mining Modeling

Data Mining Objectives

A DM goal influences the choice of algorithms and modeling techniques to use. Therefore, Quintela (2005) among the various types of DM objectives, highlights the following:

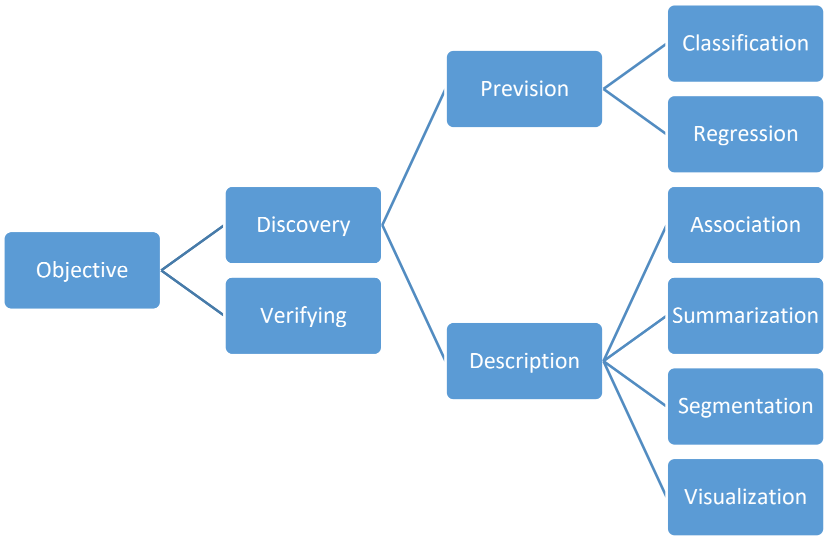

Figure 6: Data Mining Methods and Techniques. Adopted in (Quintela, 2005)

- Classification: As one of the most common and commonly used objectives in supervised learning, it corresponds to the discovery of a function that associates a case with a class within several discrete classification classes, that is, to identify a class in relation to the group that it belongs to.
- Regression: or prediction, consists in predicting future or unknown values of a dependent variable, from samples, there is usually the presence of numerical and non-categorical values.
- Association: or dependency, searches for a model that describes significant dependencies between variables, starting from a group of strongly identified and associated data.

- Summarization: Use of methods to find a robust description for a subset of data.
- Segmentation: one in which allows the identification of a finite set of categories to describe data.
- Visualization: use of graphs, diagrams, or another visual form to present results (final or intermediate) of DM.

Methods and Techniques

The DM methods are usually divided into supervised (predictive) and unsupervised (description) learning, where the supervised methods are provided as a set of data that have a predefined target variable and the records are categorized in relation to it, and the latter does not require a target attribute (in some cases the classification can also fit into this group). Once there is this separation in advance, it is also important to highlight the techniques associated with each objective (or task) of DM, presenting the main ones in the Table 1:

Table 1: Data Mining Techniques and Tasks

Techniques/ Tasks	***Classification***	***Regression***	***Segmentation***	***Association***	***Summarization***	***Visualization***
Decision Trees	✓	✓	✓		✓	✓
Rule Induction	✓	✓	✓	✓		
Neural Networks	✓	✓	✓	✓		
Genetic algorithms	✓	✓	✓		✓	
Neighborhood Roughs			✓		✓	

Source: Camilo & Silva (2009)

Starting from the beginning of the previous table, there are several techniques that can be applied for each task, and therefore, as the research problem is related to the Regression method, will be described the Decision / Regression (DT) Trees, Random Forest (RF), the Neural Networks (NN) on their variances MLP (Multilayer Perceptron) and MLPE, K-NN (K-Nearest Neighbors) and finally the Support Vector Machines (SVM), considering the scientific work of Moro, Cortez, and Rita (2015).

It is important to note that, one of the methods to assess and validate the built model consists in applying cross-validation, that is, if the original data set is very large, there is no problem in splitting it into training and a test set. However, available datasets are always “too small”, so that we need to make sure we use the available data most efficiently, using a process known as cross-validation (Janert, 2010). The objective of cross-validation is to estimate the expected level of fit of a model to a set of data that was used to train the model.

The model to be built works as an estimator looking for the best model. There are large mechanisms to measure the estimation of the error, being MAD (Mean Absolute Deviation), SSE (Sum Squared Error), MSE (Mean Squared Error), RMSE (Root Means Squared Error), MAE (Mean Absolute Error) and NMAE (Normalized MAE) (more information at Quintela, 2005). Particularly to this research, it will be used only two last metrics, where according to Silva et al. (2018), as far as error metrics are concerned, MAE is one of the most frequently used metrics for assessing forecast accuracy and it consists of the mean of the absolute difference between the total of predicted values (Predi) for a given output variable and its actual values (Truei) for all its n observations. The same with NMAE, which consists in entailing the distribution of the MAE through the difference between the interval of the values of the output variable.

According to Abreu (2016), it is important to compare errors metrics in order to evaluate the models, where the difference between the actual value and the predicted value (designated by error or residue) is as less as possible. Thus, all statistics compare true values with their predicted, despite the different formats, all illustrate "how far" are the predicted values from the true values. It is still important that the model with highest correlation and with estimates of smaller errors is the candidate with the best performance.

Decision Trees automatically test all values of a given data to identify those that have a strong connection to the output records selected for the test. Graphically, like a tree, consisting of a structure that connects a set of nodes through branches resulting from a recursive partition of the data, from the root node to the leaves, each branch representing a conjunction of conditions, as well as the leaves (pure nodes) correspond to classes / objects, internal nodes to attributes, and branches to attribute values (Quintela, 2005). Still, decision trees can be categorized in both classification trees and regression trees, the latter estimating the value of a numeric variable while the former qualifies the records and tries to associate them with a correct class.

The RF model is based on building a series of DT and use them in combination. Thus, it works by creating multiple decision trees with random distribution of the attributes in the nodes and selects the one that has the best result, that is, as if each tree in the forest was a decision tree, voting for the class returned by it, in the end, the forest chooses the most voted class as its decision (Barbosa & Rolim, 2017). In fact, the RF cannot be directly interpretable as it happens for DT, although it is still possible to provide explanatory knowledge in terms of its input variable relevance (Cortez & Silva, 2008).

Quintela (2005) states that NN are strongly associated with the nervous systems of living beings, the human being, where many researchers believe that these sub-symbolic models offer a more promising approach in the construction and operation of real intelligent systems. During the learning process, the NN, through a learning or training algorithm, adjusts the connection weights until a satisfactory result is achieved.

An alternative to the linearity problems that are identified in NN, we can search for great accuracy of the results by adopting more intermediate layers of neurons and an output layer, also called MLP. Although the perceptron network is simple, with only one layer of neurons (and MLP for several layers), they are best used in classification problems with good predictive capacity, but they are also applied in a regression context, changing the fact that there is no discretization imposed by the choice of the neuron with the highest output in the prediction (Gama et al. 2012).

Next, according to Gama et al. (2012), the k-NN algorithms are called paradigms where objects with similar characteristics belong to the same group. Although the prediction is considered costly in k-NN algorithms (because in a large set of training objectives this process can be time consuming), as it is affected by the presence of redundant and irrelevant attributes, these algorithms are applied in complex problems, where its algorithm training is very simple, consisting of memorizing training objects, and naturally increment them (Gama et al. 2012). These authors also consider that this algorithm is widely used by the knowledge extraction community because it is simple to apply and presents a good predictive rate in several datasets.

On the other hand, the SVM are based on algorithms that have their variants, initially created for classification problems, and nowadays also applied in regression, and has as objective the implementation of distance between the classes of a dataset (Quintela, 2005).

As reported by Gama et al. (2012), the most commonly used NN in practice are the MLP networks (although the research also benefits from the MLPE variation - ensemble architecture of neural networks), as well as the SVM, these techniques are considered "black box", that is, the extraction of knowledge is encoded in equations with difficulty on interpreting, in contrast to the models generated by symbolic techniques such as decision trees.

Therefore, according to Baker et al. (2011) cited by Camilo and Silva (2009), Data Mining techniques can be applied in different decision-making contexts, whether in finance, education (currently called MED), health, or marketing, logistics and manufacture. In fact, it is possible through the technique, the process of obtaining student data to verify the relationship between their learning on a pedagogical approach, thus allowing the teacher to understand whether their approach has been or not effective in both the development of the student, as in the elaboration of their teaching methods.

2.5.4. Mining Educational Data

Mining Educational Data has been considered a research area that is concerned with the search for methods that explore educational data, or when it comes from educational environments, in which exist an objective on perceiving students and their academic performances, as well as to explore better ways of learning for the same.

For example, Brito, Júnior, Queiroga, and Rêgo (2014) which aimed at the identification of students who needed didactic support in science disciplines in the course of Computing Sciences, which through a set of real data applying the techniques of Data Mining, where variables were chosen (within the existing ones), students' entrance notes, performance in the first period of the course, final average in each discipline, it was possible, with an accuracy of 70% to show that it is feasible to predict student performance using the variables (student entry grade) and applied modeling techniques, allowing educators to take measures that prevent low student performance, or to improve pedagogical techniques until then.

Under the same point of view, França and Amaral (2013) presented results through the application of segmentation (a method where data share similar trends and patterns) on data collected from evaluations, to group students with similar learning difficulties in a Programming discipline. It was possible too to detect which content and at what level of learning (categorized in Bloom’s Taxonomy) was assimilated by the students and what could be the methods to be adopted to overcome the identified learning problem.

Within the studies by Almeida (2017) and Cortez and Silva (2008), the aim was the identification of the factors that influences the classification obtained by a student, in exams of Advanced Excel and Introduction to Excel, and the prediction of the results of the students with the identification of the factors that influence their educational success/failure in Mathematics and Portuguese classes, both applying DM techniques: MLPE, SVM, DT, NB and other techniques respectively. Therefore, it was possible to conclude that the examinations that had a very long MCQ enunciation are one of the main causes that can influence negatively the results obtained by these questions, either by the student’s interpretation or even misunderstanding of the objective of the question, the degree of difficulty and the topic of the subject, as well as, it is also possible to predict student outcomes, especially when associated with social and educational factors.

Prior Literature - Conclusions

Much has been discussed recently concerning the teaching and learning process, particularly the form in which educational objectives are defined and in which way these can guarantee the acquisition of knowledge and competence for a student; assessment is one of the most used strategies to measure theoretical and practical performance.

That being so, it is necessary to consider the existence of a plan, a guide or even a methodology that aids and guarantees integrity and concordance between all the stages involved in this learning process. Whether to analyze the educational process (acquisition of knowledge, competence or ability) of the student, or to create an assessment that is in accordance with defined objectives, Bloom's Taxonomy, composed of educational objectives (knowledge, understanding, application, analysis, synthesis and assessment), is used now.

Seeing that assessment is an object in all this process, it is important to mention that apart from the objectives included in the same, the format of the test also is part of the assessment method, where from the various existing formats this research centered only on multiple choice and open question tests. Bearing in mind the frequent transition of tests carried out on paper to electronic platforms, it is meanwhile necessary to understand the transition of assessments on paper (whether CR or MCQ) to electronic format also considered in some e-assessment literature, this latter being elaborated only in MCQ to deal with objective questions, thus giving the motive for the problem of research, to try to understand if some difference exists in the performance of the student between using CR on paper or MCQ using e-assessment, or, whether in the final analysis, there is no difference, employing the revision of works by authors who had already approached the matter and by methodology adopted by work using the techniques of Data Mining.

However, the literature shows that there is no consensus regarding the equality of questions in measuring the same things, or if both are distinct, the tendency being to favor MCQ to measure higher positive results, while there are those who defend an assessment system that includes both types of questions.

In the literature researched, there are several works on Data Mining, that approach the use of the technique for identifying student learning over a pedagogical approach, in a wider concept defined by Mining Educational Data, which includes the activities of estimating/foreseeing positive/negative results in a curricular unit, while there is a dearth of articles using Data Mining (since the majority used statistical methods) to verify the relation between the results of students obtained in CR and MCQ assessments (which revolve round the objective of this research).

Chapter 3 – Methodology

This chapter includes project analysis and design, from the problem and business understanding, data analysis and treatment, application of data mining predictive techniques to results interpretation, representing this dissertation topic [Constructed Response or Multiple-Choice for Evaluating Excel Questions? That is the Question], in which we can verify if the proposed objectives have been reached or not.

The project is developed following the CRISP-DM methodology, which the context, problem and description and treatment of the data are explained in this chapter.

3.1. Business Understanding

The first phase of the methodology refers to problem understanding and its context, as well as the proposed objectives. It is fundamental to understand all the details of the business at first step, from identifying the problem, determining the objectives, assessing the current situation, identifying each specific criterion and how the results are expected to be obtained, as the influence of each one can have on solving the problem.

The evaluation is important in educational sphere, especially for student in learning in the academic world.

Through the several test formats used for evaluation, this research addresses those with CR format and those of MCQ, in which the differences are with the presence of an open field for students to answer freely in the first type, and in the second with MCQ, there are 4 alternatives which only one of them is correct and others were considered as similar or merely distractors.

Moreover, this research illustrates the importance of evaluation in the teaching-learning process, and also verifies the similarity of the results from students in performing Excel exams consisting of CR questions and their MCQ equivalents/paired (see appendix 5 for examples).

This comparison requires the identification of relevant attributes, whether demographic, Bloom's Taxonomy categories (linked to processes of knowledge acquisition, aptitudes, attitudes, and those that composes teaching-learning process), and those that describe the student, the types of questions (CR or MCQ), the exam and the students' answers.

Firstly, the empirical experiments were based on Excel exams, in "Advanced Excel", performed at ISCTE-IUL, in the academic year 2016/2017, between the 1st and 2nd semester. These exams were performed using paper, which the content was Excel´s objective formulas (different from conventional questions with descriptions, interpretations and justifications, see appendix 1) and the basic structure of the exam consisted of two blocks, the first composed of 10 CR questions and the last by 10 paired MCQ respectively.

That is, previously the student answered a CR question by writing the Excel formula, and now with MCQ (which is elaborated similarly), the student must select one of the four options. Further, both formats had an image (table or completed worksheets) to help the student on answering.

Regarding CR score, the student gets one (1) point for each correct answer, zero (0) for incorrect and a grade on a scale of zero to one depending on what was expected. For MCQ, it was possible to identify three possible scoring cases, one for correct answer, 0 for unanswered (without choosing any of the four options) and 0.25 discount for each wrong answer.

The next phase is related to understanding the samples and the data collected and how they are relevant to the study since we described the problem and business understanding.

3.2. Data Understanding

This phase involves data collection and analysis, identifying data subsets, quality problems and so many other characteristics which defines their impact on obtaining results.

Firstly, the data collection, interpretation and storage were possible using the Microsoft Access, with relational models[5] design, organizing in tables and then load them to Excel Spreadsheets, to organize single answer's records and create new attributes to enrich the dataset.

[5] RM (relational model) is considered a way to manage data using a structure and consistent language, that helps the user in gets the best useful information from data.
[https://blogs.oracle.com/oraclemagazine/modeling-and-accessing-relational-data]

The initial dataset (since it will change in the 3rd phase of CRISP-DM - Data Preparation) includes a total of 2873 records corresponding to the students' responses in each question from the exam in Advanced Excel in ISCTE-IUL. Each observation of the dataset is defined by 54 attributes (see appendix 2) that will be described below, one by one or into categories for better understanding.

Considering the samples and data distribution, simple statistical methods were used, in which from the 283 tests performed, 240 were registered in the 1st semester, and the remaining 43 in the 2nd, in a proportion of approximately 86% of exams in the first semester and 14% in the last one (Fig. 7), however 36 students performed more than one exam. With the number of variants (identified by "ExamVariant" - created so that the classes did not repeat the same exam, however the structure of the exams still the same), "Laboratory of Languages and Transversal Competences (1B)", "Laboratory of Languages and Transversal Competences" (2A), "Laboratory of Languages and Transversal Competences (2B)", with approximately (65%), 14%, 11%, and two last to 5% each of the total samples (see Fig. 8).

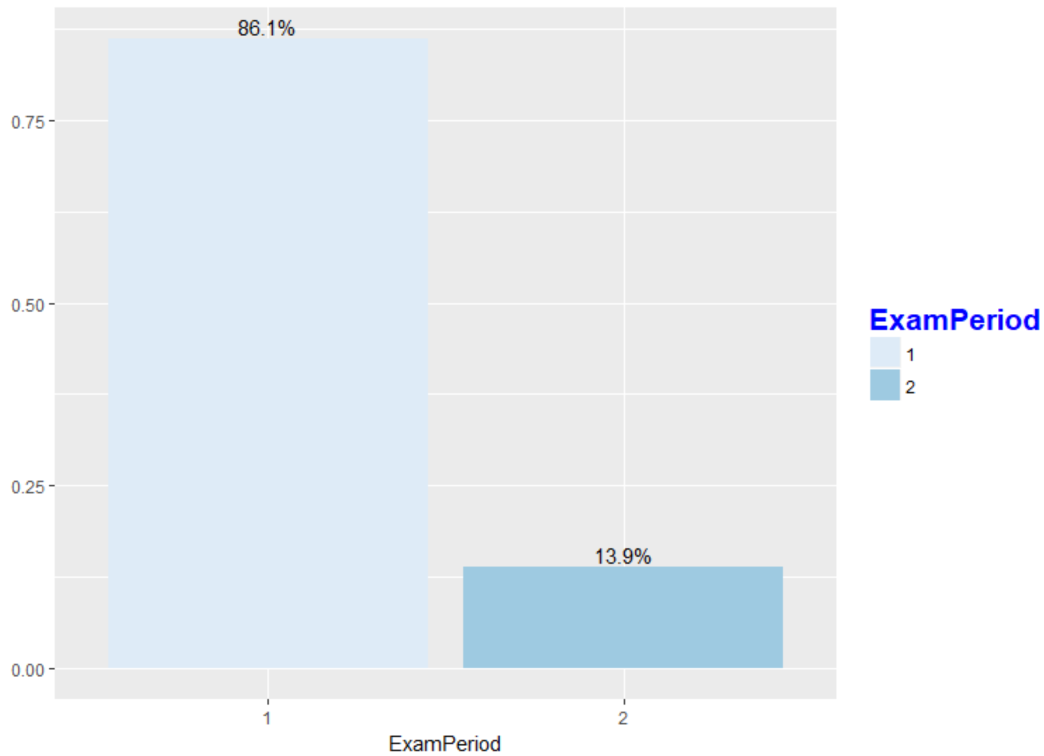

Figure 7: ExamPeriod Frequency in Percentage

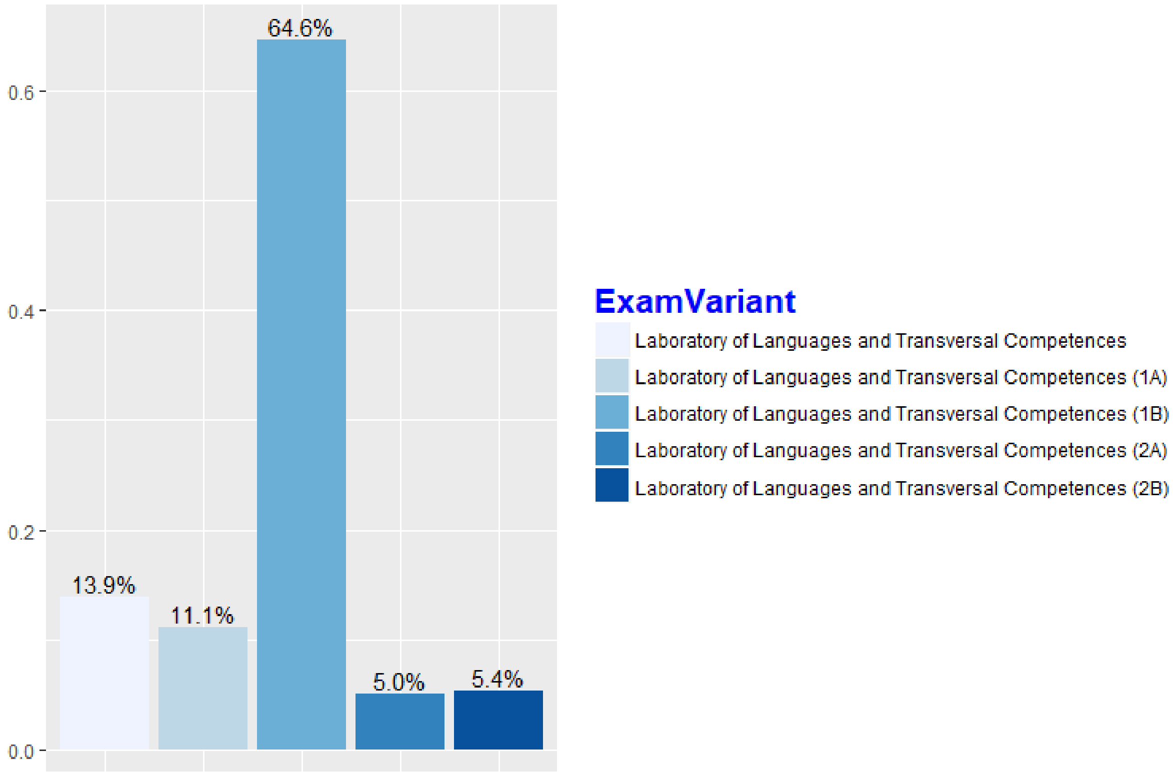

Figure 8: ExamVariant Frequency in Percentage

Finally, some attributes had the same value for all records, i.e., for the attribute "Year" the value "2016/2017", "" ExamYear "with" 2017 "," Semester " with "2", "Subject" with "Advanced Excel", "DurationTotal", "DurationFirstPart" and "DurationSecondPart" with "60min", "40min" and "20min" respectively, the attribute "NrChoices" with the value "4", "Image" registering "Yes", "IncomingMobility" with "No", "MobilityAgreement" to "No", and finally the "Degree" attribute with the "Licentiate".

Student Data

With student data, the "IdStudent" identifies the exam in which the student performed, “Gender” (like Sheaffer & Addo, 2012 used in their project) in which 247 students, 81% are males and the remaining 19% are female (see Fig. 9). About “Schedule”, 69% of the students learn in daytime and 31% in night classes (see Fig. 10), and only 1% of them learn “Advanced Excel” as an extracurricular unit. At the same time, the "Status" attribute divided into 6 categories, "Full-Time", "Worker", "Part-Time", "AEISCTE-IUL Athlete", "Special Educational Needs" with 87%, 6%, 3%, 2%, 1%, 1% (see Fig. 11).

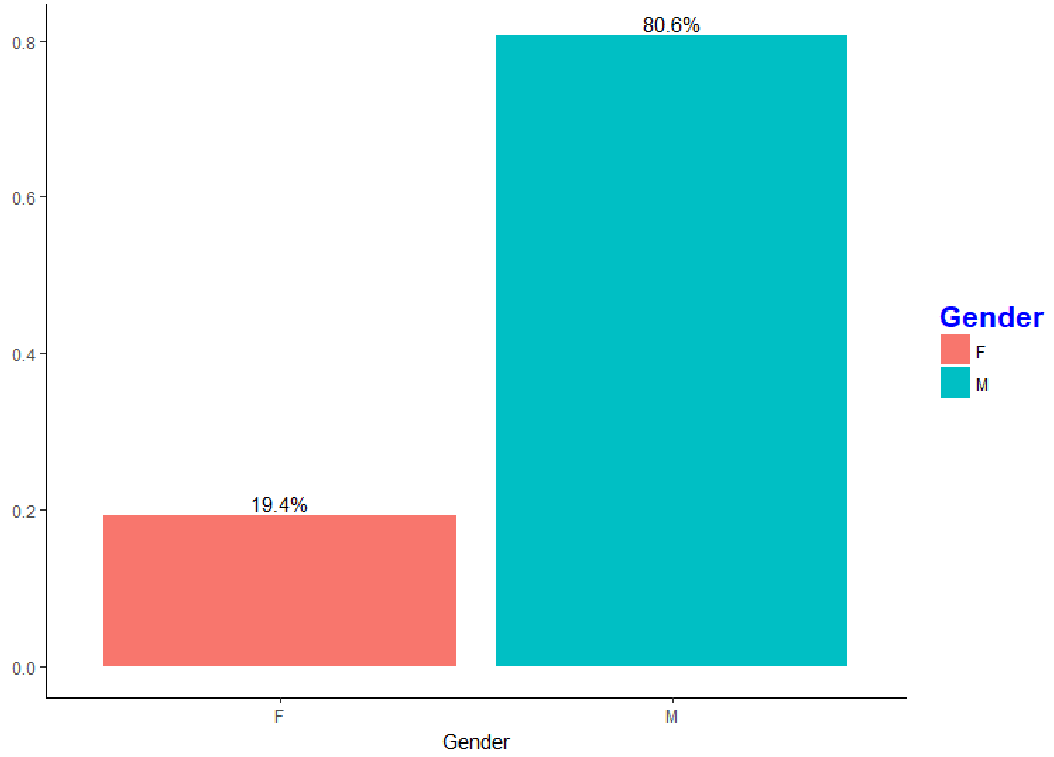

Figure 9: Gender Frequency in Percentage

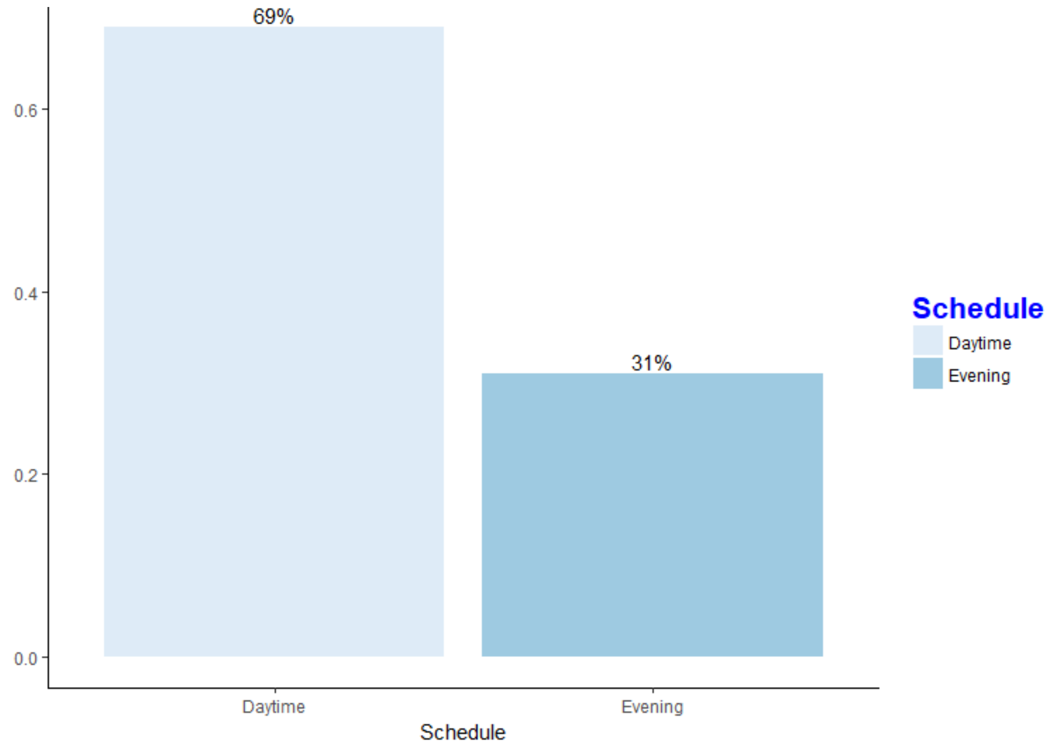

Figure 10: Schedule Frequency in Percentage

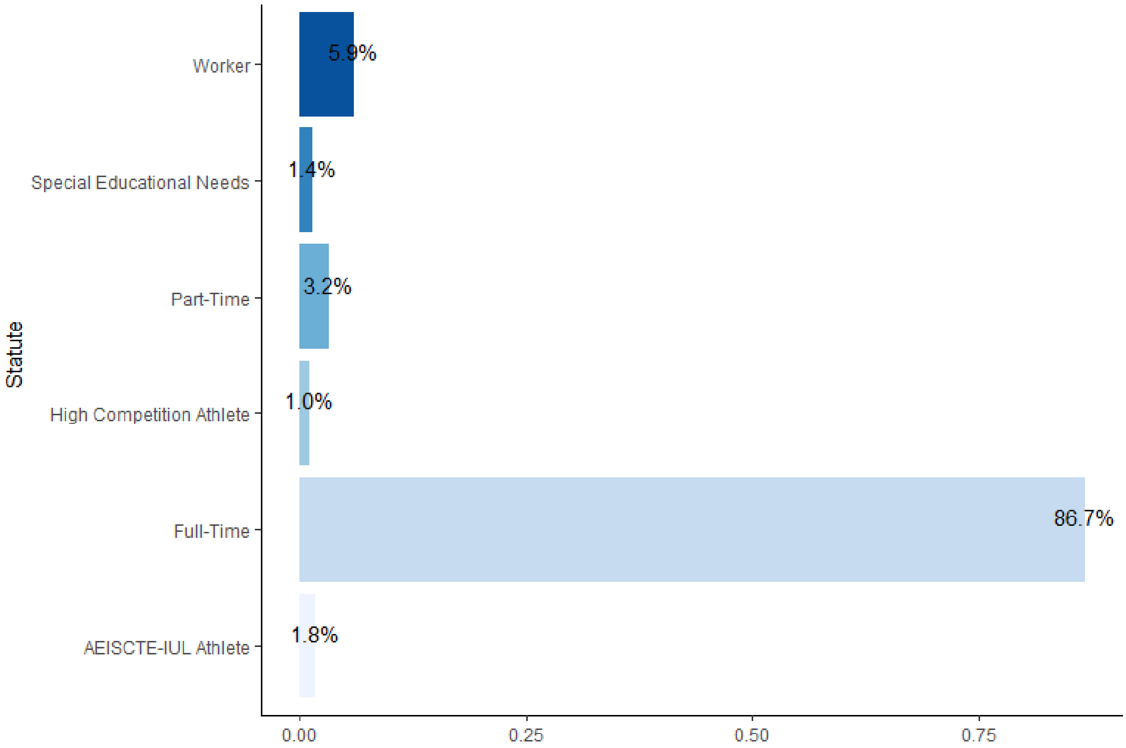

Figure 11: Status Frequency in Percentage

For "Course" attended by each student, it was possible to find 5 divisions (see appendix 3 for distribution of classes) such as "CE", "CSBM", "TCE", "IC" and "A" with percentages of 42.5%, 42.9%, 13.5%, 0.7% and 0.4% (since only 1 student attends the Anthropology course, illustrated in Fig. 12).

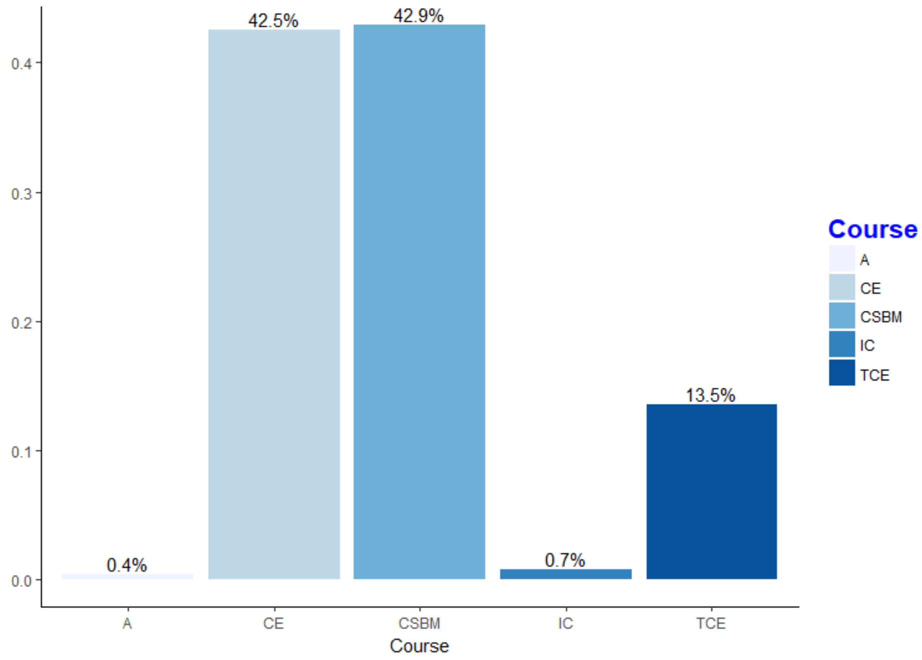

Figure 12: Course Frequency in Percentage

Finally, the attributes "GradeCR" and "GradeMCQ" that correspond to the total score obtained by the student in each block of questions.

Question Data

First with "Difficulty", most of the questions are "Medium", occupying 62%, followed by "Easy" with 26% and 12% for "Difficult", the identification of the question difficulty had as a criterion the composition of operations/formulas, in which for Easy those questions with simple operations (sum and difference), Medium to those composed by "reasonable" formulas and Difficult to questions with higher number of formulas (see Fig. 13, Almeida, 2017 and appendix 5).

Variable "Topic", as the topic of the question (CR or MCQ, since these are equivalent, the topic is the same): "Statistics", "Basic", "Date and Time", "Text", "Search", "Logical" and "Formula", occupying respectively 26%, 26%, 14%, 12%, 11% , 10% and 1% of the 200 questions as you can see in Fig. 14. In advance, the attribute "IDQuestion" identifies each question and associates a CR to its equivalent MCQ.

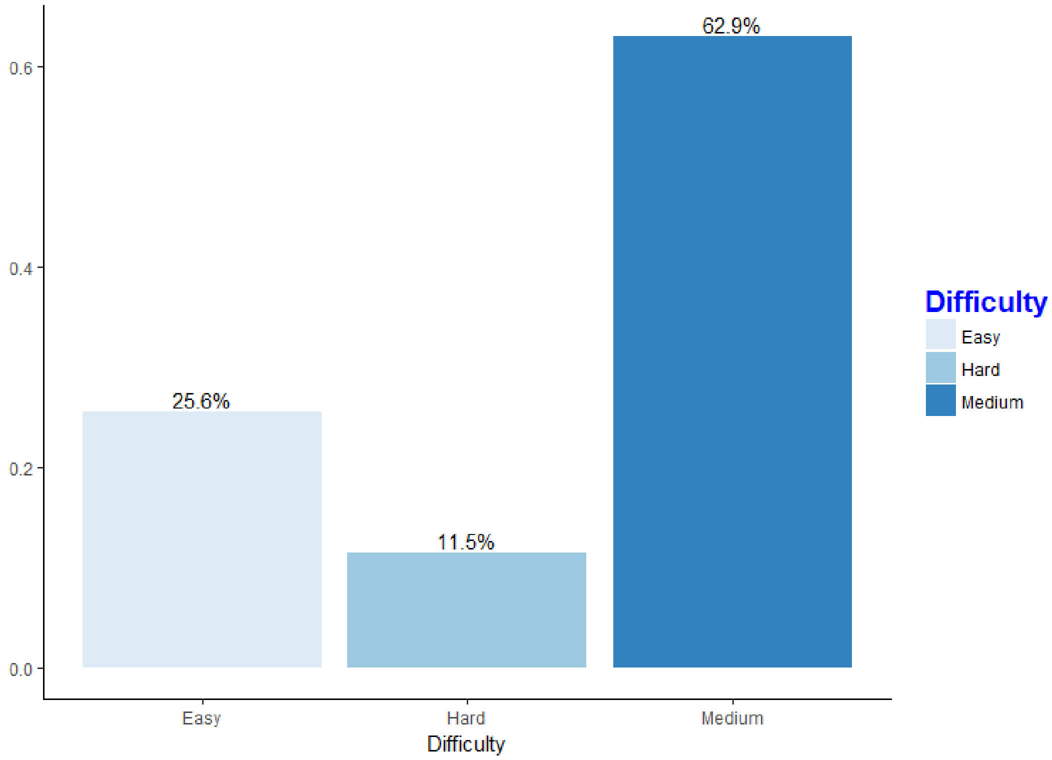

Figure 13: Difficulty Frequency in Percentage

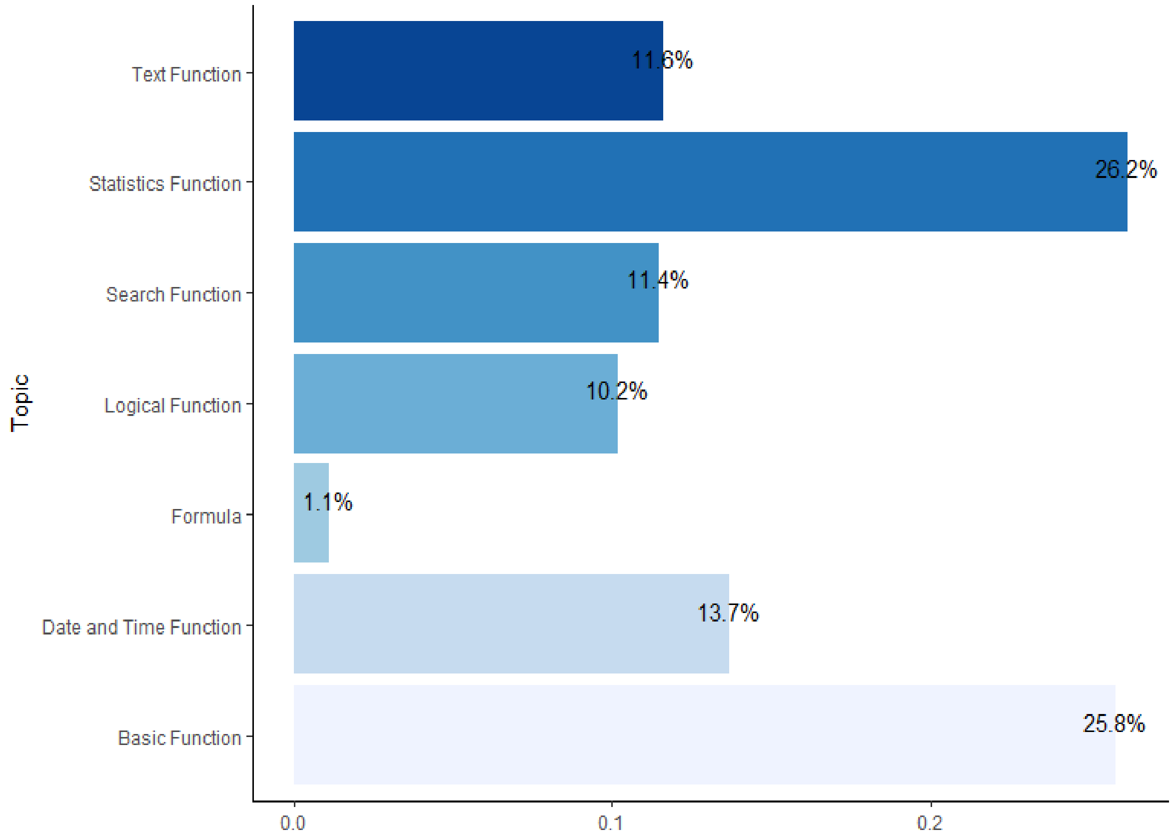

Figure 14: Topic Frequency in Percentage

Also, two new attributes were created in Excel, "NrSimilar" and "NrDistractor", indicating for each MCQ the quantities of similar responses to the correct answer and of distractors too, where 51% of MCQs contain only 1 similar option, 19% none, 17% with 2 similar options and 13% with 3 (see Fig. 15).

Regarding Bloom Taxonomy's learning level, the "Analyzing" level with 43.5% of the 200 questions (15% CR and 28.5% MCQ), "Remembering" at 26.5% (25% for CR and 1.5% for MCQ), "Applying" to 17.5% (7.5% CR and 10% MCQ), "Understanding" 10% (where all recorded in MCQ only), and last "Evaluating" to 2.5% (all registered in CR only).

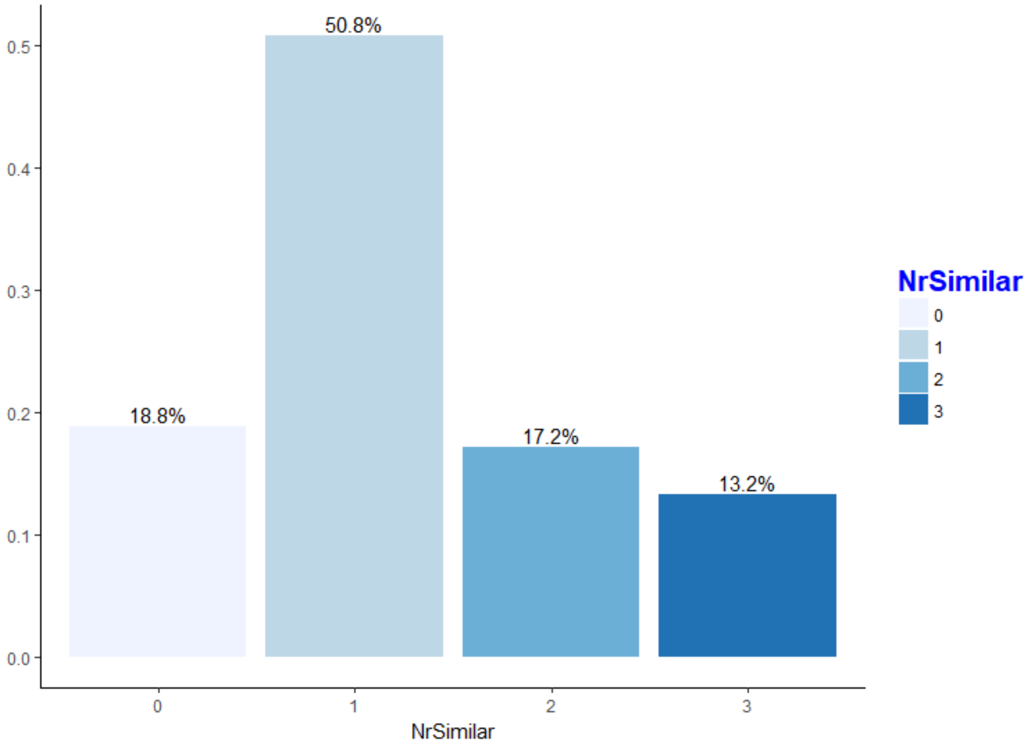

Figure 15: NrSimilar Frequency in Percentage

In addition, the attributes "TextChoice" and "TypeChoice" (from 1 to 4) correspond respectively to the text of each MCQ option and its type (correct, similar and distractor).

The text of the questions is contained in the attributes identified as "TextMCQ" and "TextCR". While text mining could be applied to it, this was considered out-of-scope of the present dissertation due to time constraints. We could use Text Mining techniques in which is a semiautomatic process of extracting interesting and non-trivial patterns of large amounts of unstructured textual data, to achieve a structured format (Miller, 2004 cited by Moreno, 2015), but since it is a different technique from those of Data Mining, and because it is not included in the objective of this research , the number of characters and number of words in each question where counted, creating attributes like "NrWordTextCR", "NrWordTextMCQ," and "NrCharacterTextMCQ", like Almeida (2017).

Answer Data

The attribute "IDAnswer" identifies a student's answer. Thus, it helps in pairing a CR question and MCQ. Next with "AnsweredCR" indicating if the student answered a CR question or not, where in 2787 registered, 91% of CR were answered and 9% not answered as Fig. 16 shows. Also, the attributes "MCQP" were created in Excel indicating the option that the student chose, "MCQCorrect" indicating the correct option and "MCQD" if the question was penalized or not.

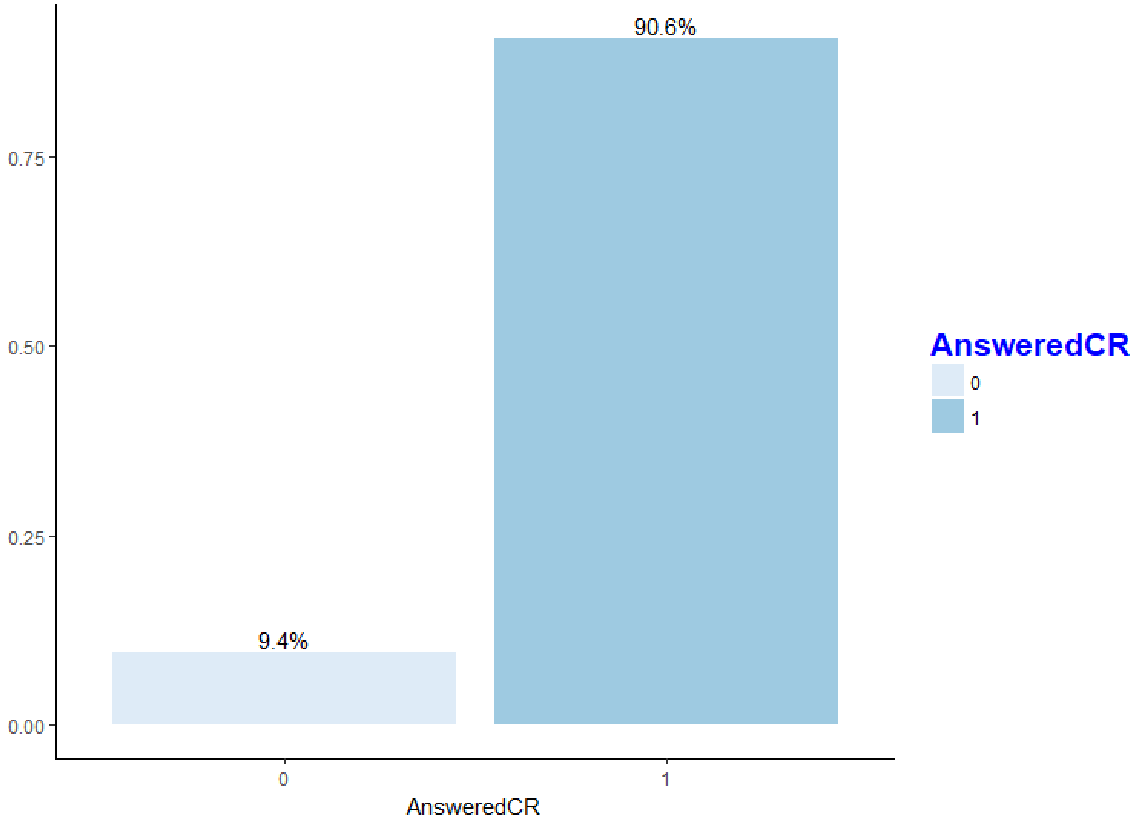

Figure 16: AnsweredCR Frequency in Percentage

3.3. Data Preparation

The data preparation consists in a set of activities with the purpose of constructing the dataset that will be used for creation and validation the model in the next phase (Quintela, 2005). Therefore, the column "Database" in Table 2 shows the attributes managed in Access, "DataExcel" containing attributes that will be used for creating the models, "Dataset (Type)" illustrates the data type of the attribute and "Include (Dataset)", those excluded, processed and created in the R environment.

Table 2: Attribute Analysis

Table	Name		Include	Type
	Database	**DataExcel**	**(DataSet)**	**DataSet**
Answer	IDAnswer	IDAnswer	✗	Numeric
	Answered	AnsweredCR	✓	Logical
	ScoreCR	ScoreCR	✗	Numeric
	ScoreMCQ	ScoreMCQ	✗	Numeric
	MCQP	MCQP	✗	Factor
	MCQD		✗	
Exam	IdExam	IdExam	✗	Numeric
	Year		✗	
	Semester		✗	
	ExamPeriod	ExamPeriod	✓	Factor (2)
	Subject		✗	
	DurationTotal		✗	
	DurationFirstPart		✗	
	DurationSecondPart		✗	
	ExamVariant	ExamVariant	✗	Character
Question	IDQuestion	IdQuestionMCQeq	✗	Numeric
	Text	TextMCQ	✗	Character
	Topic	Topic	✓	Factor (7)
	QuestionType		✗	
	Difficulty	Difficulty	✓	Factor (3)
	TextChoice1	TextChoice1	✗	Character
	TypeChoice1	TypeChoice1	✗	Factor
	TextChoice2	TextChoice2	✗	Character
	TypeChoice2	TypeChoice2	✗	Factor
	TextChoice3	TextChoice3	✗	Character
	TypeChoice3	TypeChoice3	✗	Factor
	TextChoice4	TextChoice4	✗	Character
	TypeChoice4	TypeChoice4	✗	Factor
	BloomLevel	BloomLevelMCQ	✓	Factor (4)
	NrChoices		✗	
	Image		✗	
Student	IdStudent	IDStudent	✗	Numeric
	NrStudent	NrStudent	✗	Numeric
	Name		✗	
	Gender	Gender	✓	Factor (2)
	Class	Class	✗	Factor
	Course	Course	✓	Factor (5)
	Subject		✗	
	Year		✗	
	ExamPeriod		✓	
	ExamYear		✗	
	GradeCR	GradeCR	✗	Numeric
	GradeMCQ	GradeMCQ	✗	Numeric
	GradeTotalExam		✗	
	Schedule	Schedule	✓	Factor (2)
	Status	Status	✗	Factor
	ExtraCurricular	ExtraCurricular	✗	Character
	Incoming Mobility		✗	
	Mobility Agreement		✗	
	Degree		✗	
		IdQuestionCR	✗	Numeric
		MCQCorrect	✗	Factor
		NrSimilar	✓	Factor (4)
		NrDistractors	✓	Factor (4)
		TextCR	✗	Character
		NrWordTextCR	✓	Numeric
		NrCharacterTextCR	✓	Numeric

		NrWordTextMCQ	✓	Numeric
		NrCharacterTextMCQ	✓	Numeric
		BloomLevelCR	✓	Factor (4)
		ScoreDifference	✓	Numeric

Among these attributes, "ScoreDifference" were selected as an input attribute (also called response) in obtaining a model with better performance.

The process started by importing the entire file "AnswerDF.csv" into RStudio, with 2787 rows (records) and 43 columns, opposed the previous 2873 described in data understanding, because at this current phase, only questions with their equivalent where taken into consideration when analyzing the dataset. Thus, each line of the file corresponds to a CR answer and its MCQ equivalent of each student in a specific exam.

The data frame was created with the name "exam_df", but the Excel sheet only contains some of the total identified attributes, since some of them would not contribute on obtaining a model with good performance (such as "Year" - "2016/2017", "Subject" - "Advanced Excel", "DurationTotal" - "60", "DurationFirstPart" - "40", "DurationSecondPart" - "20"), and for other reasons explained above in data understanding.

Subsequently, it was necessary to include the "Topic", because it would be interesting to verify the student's performance with the content type of the question, and even to identify in which topics the student achieve better results or not (see Hudson, 2012).

The inclusion of the attributes "BloomLevelCR" and "BloomLevelMCQ" to identify the different levels of skills and behaviors according to Bloom's Taxonomy objectives, like Scouller (1998). The "Gender" representing the samples categorized by "M" for male and "F" for female, to evaluate if the skills/scores in both are equal or not, as happens with Hudson (2012).

With "NrSimilar" and "NrDistractors" attributes (although complementary) is due to the fact of different experiments, in order to choose the attribute with the best performance avoiding redundancies.

The attribute "AnsweredCR" identifying the CR question, if the student answered or left it in blank, since in both cases the student may have a score of 0 (zero) when answering correctly or if chose not to respond. Unlike "ScoreCR" and "ScoreMCQ" attributes that were excluded, since they could induce the models in achieving the expected results (with almost 100% accuracy), however, they aided the creation of ScoreDifference attribute, the one that illustrates the difference between the scores in CR questions and their MCQ equivalent. Likewise, "GradeCR", "GradeMCQ" and "GradeTotalExam", excluded to avoid redundancy in the attributes, since if they are needed in the modeling phase, they can be obtained in the same way as the previous one.

Corresponding to the examination period by the "ExamPeriod" attribute (1st or 2nd period), and "Schedule" to the period that the student learns ("Daytime", "Evening") also included because they reinforce the justification of the results. As well as the inclusion of the attribute "Difficulty", representing for each pair of questions its difficulties.

On the other hand, the attributes excluded, "IDAnswer", "IDExam", "IDQuestionMCQeq", "IDQuestionCR", "IDStudent", "NrStudent", "Name", since they only identify the answer, the examination, the MCQ, the CR question, the student number and name.

“Class” attribute was also excluded because it was not considered enough for knowledge extraction considering the existence of several subcategories (53, see appendix 3), and because there is already a more consolidated and aggregated attribute in information (“Course”).

The "TextCR" and "TextMCQ" attributes were also excluded, as they did not provide useful information for the results, however they may perform better in Text Mining problems (see Moreno, 2015). However, attributes like "NrWordTextCR", "NrCharacterTextCR", "NrWordTextMCQ", "NrCharacterTextMCQ" where included because they are suitable for measuring if the result of the student can be influenced by the length of the question, like Almeida (2017).

The "MCQP" corresponds to the response the student chose and "MCQCorrect" to the correct option, both excluded by the possibility of bringing results biased to the model, for example, being able to influence the model in such a way that the target attribute would approach 0 (zero) and therefore to 100% satisfactory results, which would not have any benefits in relation to the objectives of the research, as in Almeida (2017).

The attributes "ExamVariant", "TextChoice" 1,2,3 and 4, "TypeChoice" 1,2,3 and 4 were excluded too, since they are unusable, as the first one was created only to identify the exams variants, the second containing texts of each MCQ option and the last one representing action that categorizes each TextChoice (correct, similar and distractor).

Two more attributes also not considered in the model, "Status" and "ExtraCurricular", the first one despite having aggregation and grouping did not bring any useful reference to the data model, as it happens with the last one, since only 2 of the total students are enrolled in the subject as an extracurricular unit, thus becoming dispensable to the model.

Thus, the exclusion of the attributes mentioned above as well as the remaining ones identified in Table 1 was due to the lack of utility in the model. The next phase includes several tests, seeking for important attributes and powerful results, and how they can be useful to knowledge extraction.

In addition, it is necessary to validate the data quality to the level required by the selected analysis techniques. This may involve cleaning data tasks, and apply techniques to estimate missing data (Chapman et al., 2000). In this case, we are dealing with outliers, since none missing value was detected in the data set as shown below with the code and his result.

```
> any(is.na(exame_df)) #looking for NA (missing values)
[1] FALSE
> which(is.na(exame_df)) #identify which gap contains missing value
integer (0)
```

The nonstandard values, as known as outliers, are records very distant from the others. The importance of detecting and removing them lies in the fact that their presence can lead to misleading results in data analysis (Gama et al. 2012). These authors also consider that, before taking any action to the observation of outliers, it is necessary to understand the causes that led to the emergence of these, either by measured errors, execution, and particularly to this research, by population samples variability. Thus, the detection was performed by graphical analysis (boxplot) only with the attribute "ScoreDifference", since all other attributes do not have observations with strong possibilities to be identified as outliers.

The Fig. 17 illustrates the presence of outliers, represented by the points above the line between 0.5 and 1.0, in fact using the code “boxplot.stats()” we can see which the limits on this distribution are, where all the samples higher than 0.75 are considered outliers.

> boxplot.stats(exame_df$ScoreDifference)

$stats

[1] -1.00 -0.50 0.00 0.00 0.75

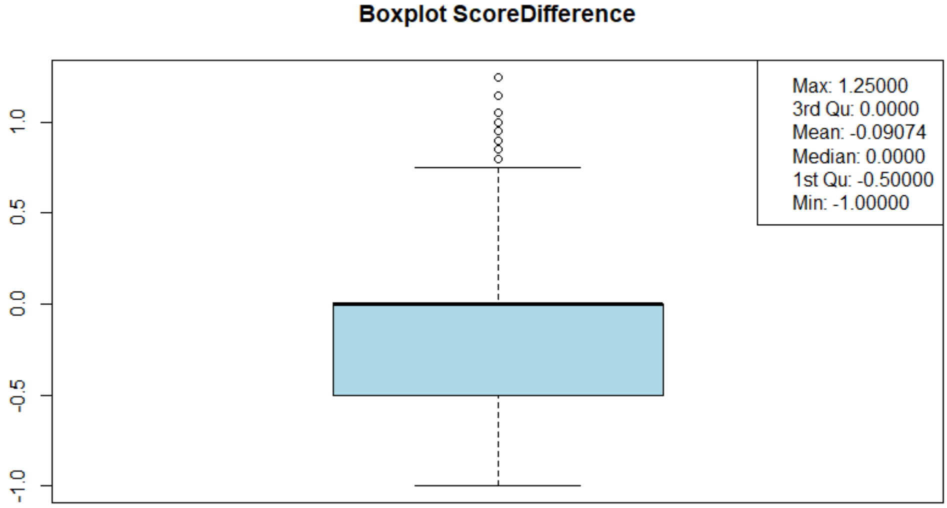

Figure 17: ScoreDifference Boxplot with Outliers

Therefore, we chose to eliminate the outliers (because they may change or modify the results of the model) and convert the attribute "ScoreDifference" to have better quality and finally without outliers.

> boxplot.stats(df3$ScoreDifference)

$stats

[1] -1.00 -0.50 0.00 0.00 0.75

$out

numeric(0)

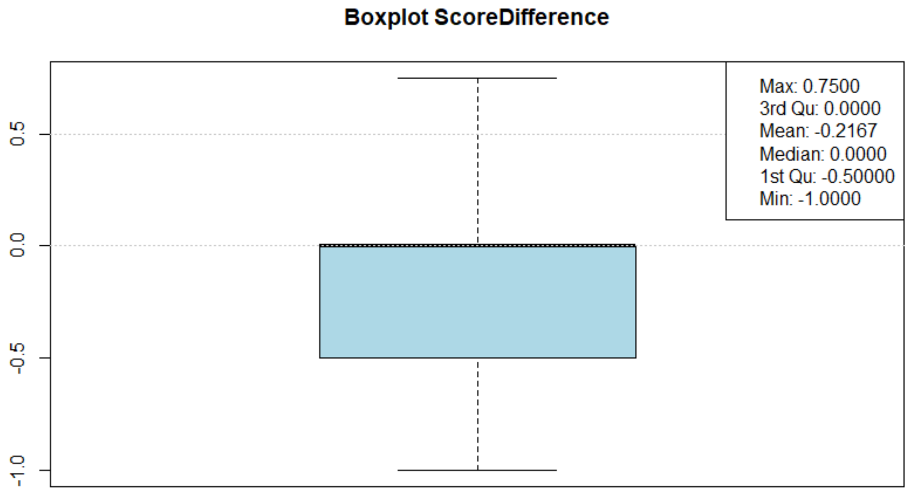

Figure 18: ScoreDifference Boxplot without Outliers

Finally, after completing the tasks of understanding and preparing data, the application of algorithms and techniques of Data Mining follows.

3.4. Modeling and Evaluation

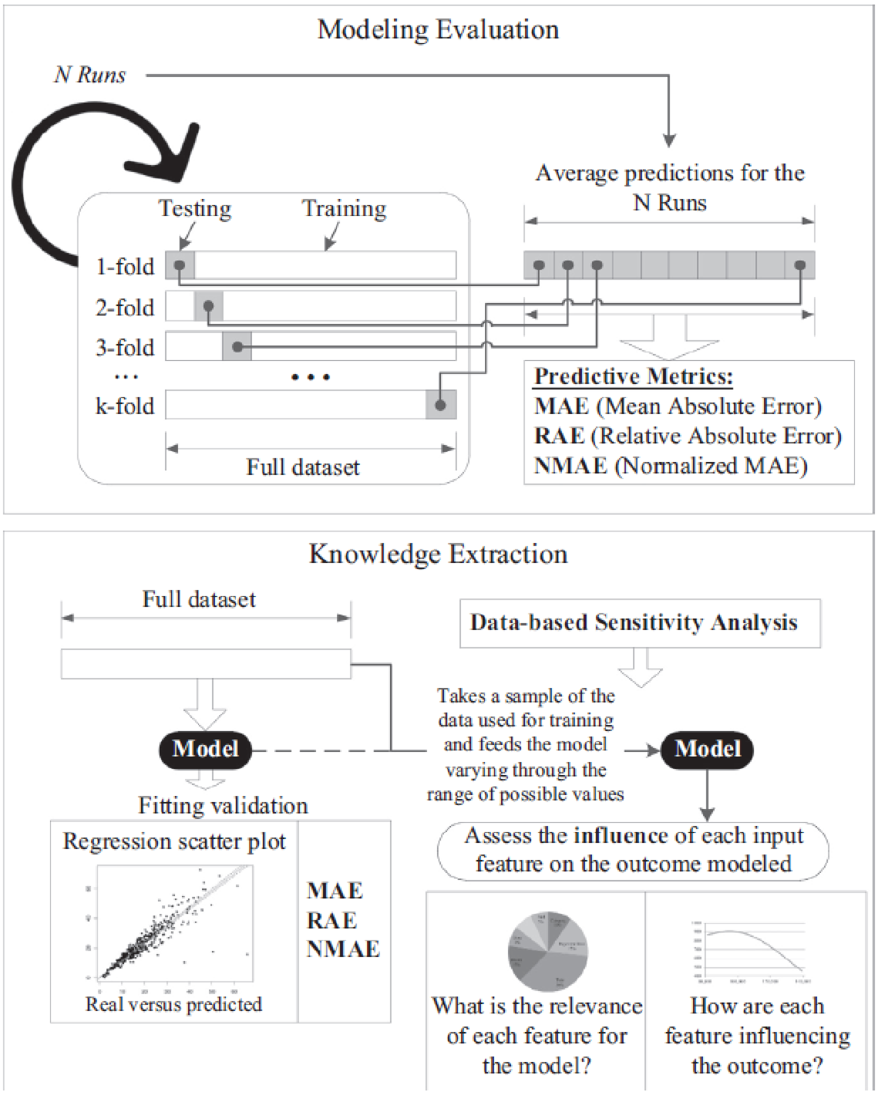

Figure 19: Modeling Evaluation Approach. [Source: Silva et al. (2018)]

The main activities performed in the Modeling phase and Evaluation are graphically illustrated in Fig. 19. From division of tests and training, application of prediction metrics, to the extraction of knowledge as result of DSA which helps decision making and business process.

Concerning a regression problem, data mining techniques previously described are selected and applied in this phase, as well as the activities of test design, applying mechanisms that will test the performance of the several models created, that is, firstly we separate the dataset into train and test sets, first for building and other to estimate its quality.

The dataset was divided into three parts, in a ratio of 2/3 for training and 1/3 for test for holdout validation in which it was performed only once (see Fig. 20 for example), distinct from k-fold validation, where the division consisted into K = 10 parts of 10 runs, each were used for both training and testing (9 for training and 1 for testing at each iteration) until the moment in which they all were tested on both sides, so the mean error can now be calculated to give a total average error value, as illustrated by the following code excerpt. It is important to emphasize that k-fold is the most robust method in relation to the previous one, besides prone to less modification because it uses the whole training set (Silva et al. 2018).

```
KSVM_ScoreDifference <- mining(ScoreDifference~., df3, Runs=10,
                        model="svm",
                        method=c("kfold",10))
HSVM_ScoreDifference <- mining(ScoreDifference~., df3, Runs=10,
                        model="svm",
                        method=c("holdout",ratio=2/3))
```

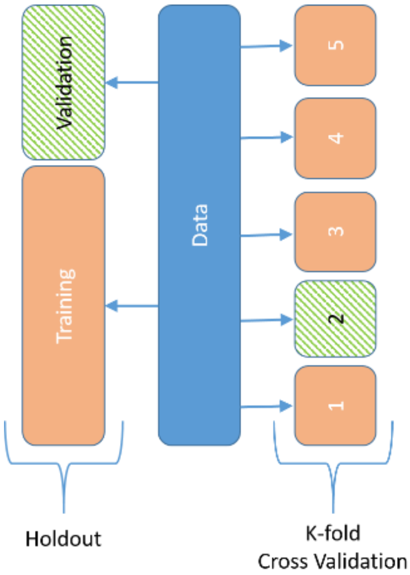

Figure 20: Holdout and K-fold Process[6]

During the methodology, 6 different predictive modeling algorithms were used to the 16 attributes (1 objective and 15 input), namely:

- DT
- RF
- K-NN
- SVM
- MLP
- MLPE

Conclusively, it is necessary to describe the results and possible convenience of the regression models generated, verifying also if the tasks and criteria were respected, comparing using error metrics, regression curves and sensitivity analysis, searching for the model with better performance which adjust the results with the research problem.

3.5. Project Development

The activities of data process were performed using Microsoft Access for collection, organization and data grouping, and Microsoft Excel to aggregate all fields into singular records for the answers, and finally the open source R statistical tool (with several

[6] Adapted in: https://bluewatersql.wordpress.com/2016/04/29/data-science-day-7-model-evaluation/

packages described below[7]), installed in the R[8] environment (RStudio[9] v 1.1.423), similar to Almeida (2017); Moro, Cortez, and Rita (2015) to build models for patterns trainings, explaining the influence of each attributes to initial objective and finally to extract knowledge.

As mentioned before, some packages were loaded into the environment, such as:

- caret: misc. functions for training and plotting classification and regression models.
- ggplot2: initializes a ggplot object. it can be used to declare the input data frame for a graphic and to specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden.
- ipred: improved predictive models by indirect classification and bagging for classification, regression and survival problems as well as resampling-based estimators of prediction error.
- maptree: functions with example data for graphing, pruning, and mapping models from hierarchical clustering, and classification and regression trees.
- nnet: fit single-hidden-layer neural network, possibly with skip-layer connections.
- rminer: facilitates the use of data mining algorithms in classification and regression (including time series forecasting) tasks by presenting a short and coherent set of functions.
- rpart: recursive partitioning and regression trees; fit a rpart model
- rpart.plot: plot a rpart model, automatically tailoring the plot for the model's response type.
- scales: generic plot scaling method

[7] Definitions based in https://cran.r-project.org/web/packages/
[8] R is a free Software environmental for statistical computing and graphics. (Available in: https://www.r-project.org/)
[9] RStudio é um IDE que torna o R mais fácil de usar e mais produtivo, combinando um conjunto de ferramentas de produtividade em um só ambiente: editor de código, depuração e visualização. (Available in: https://www.rstudio.com/)

Chapter 4 – Results and Discussion

4.1. Results and Evaluation

In this step, the degree of compliance of the final model with the business objectives is evaluated. Basically, the set of activities includes, understanding the results of data mining, interpretation, identification of which knowledge can be extracted and describe its usefulness, to verify the effects of the results on the initial objectives determined.

To validate and interpret of the models, we use variables which status was indicated by "✓ " in Table 1, in total 16 attributes (where 1 represents the target variable), with a sample of 2520 records that contain students' responses to each CR question and its MCQ equivalent.

The project included also two situations experienced in methodology stage, in which figures 21 and 22, show the distribution of the score difference in real values for the first figure and the second the distribution in absolute values, where it was necessary to compare them in order to choose the best format to the study.

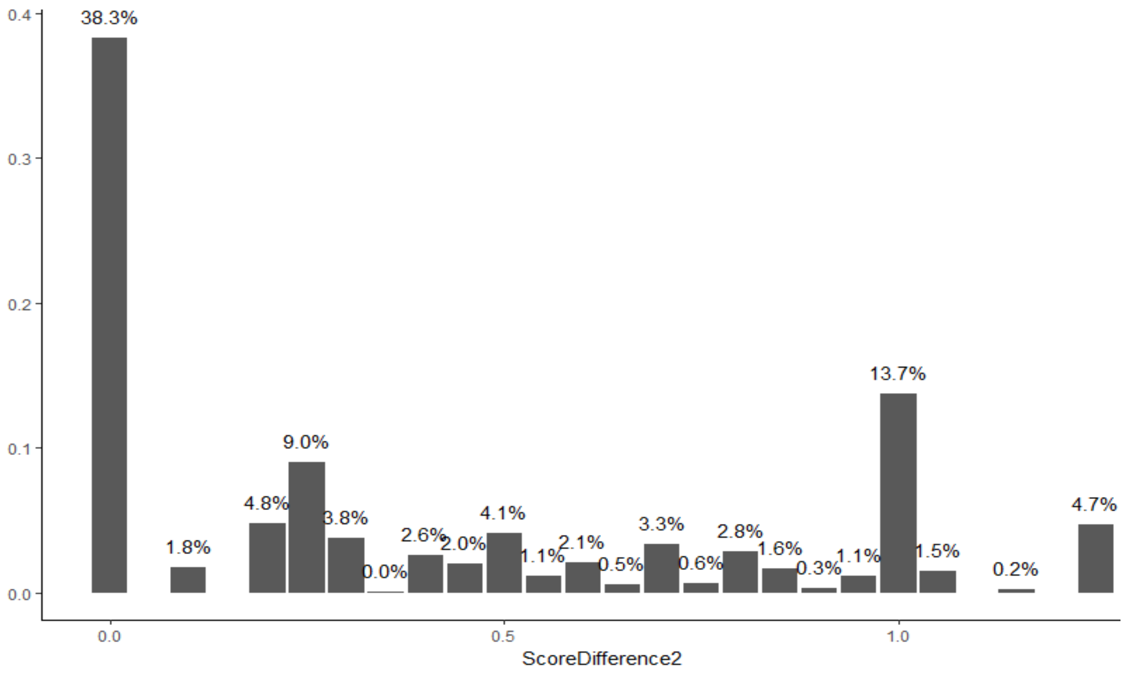

Figure 21: Score difference in absolute values

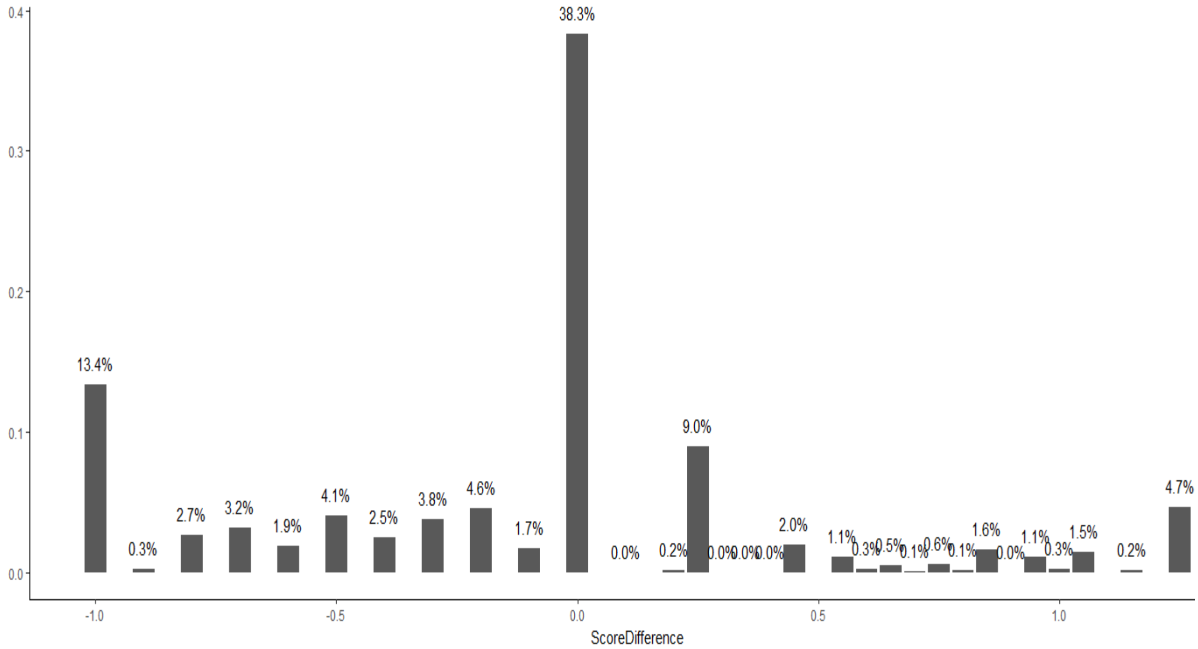

Figure 22: Score difference in real values

Thus, the cropped image (Fig. 23) of the dataset only illustrates clearly what would represent the difference of scores in real values (ScoreDifference) and absolute values (ScoreDifference2). We chose to use difference in real values not only for the reasons described above but also for the importance that the variable can bring with more details for the research.

ScoreCR	ScoreMCQ	ScoreDifference	ScoreDifference2	M
1	1	0	0.00	
1	1	0	0.00	
1	1	0	0.00	
1	1	0	0.00	
0.9	1	-0.1	0.10	
1	1	0	0.00	
1	1	0	0.00	
0.7	1	-0.3	0.30	
0.2	1	-0.8	0.80	
1	1	0	0.00	
1	1	0	0.00	
0.2	1	-0.8	0.80	
0.5	-0.25	0.75	0.75	
0.2	-0.25	0.45	0.45	
0.3	1	-0.7	0.70	
0.2	1	-0.8	0.80	
0	1	-1	1.00	
0	1	-1	1.00	
0.7	1	-0.3	0.30	
0	-0.25	0.25	0.25	
0.2	-0.25	0.45	0.45	
0	1	-1	1.00	

Figure 23: Data set excerpt illustrating both ScoreDifference

Attributes and Metrics Analysis

In fact, after building several regression models, it is necessary to apply some principles in which they can be evaluated and compared, and finally apply error measurement metrics. The same error metrics were applied to all models, so no error was inserted that would impair the comparison of the results.

Hence, from the Table 3, where the results of the metrics in each model are illustrated, by the k-fold validation, it is possible to verify that the best results are obtained by the SVM technique with the lowest value of MAE by approximately 0.31 and the NMAE by 18%, being the model with better accuracy (see REC Curves) and therefore with better performance, unlike MLP with high error values in relation to the others.

Table 3: Metrics Analysis

	MAE	**NMAE**
DT	0.35	19.79%
SVM	**0.31**	**17.68%**
RF	0.34	19.28%
K-NN	0.35	19.98%
MLP	0.35	20.12%
MLPE	0.34	19.52%

Opposed to the MLP model with higher values for error prediction with MAE = 0.35 and NMAE = 20,12%, the DT and K-NN models do not present poor results, relatively higher to RF and MLPE, but none of them less than SVM, leading to the conclusion that these models are not the best to apply in the research study.

However, for a clearer comparison between the models, the REC curve can be performed using *mgraph* function from the *rminer* library. For a clearer view, the graph of k-fold curves corresponding to the attribute “ScoreDifference” will be displayed.

In these curves, the higher the distance from the curve to the imaginary diagonal line, the better the model, proving once again, the SVM model represented by the grey line as the one with best performance in relation to the others.

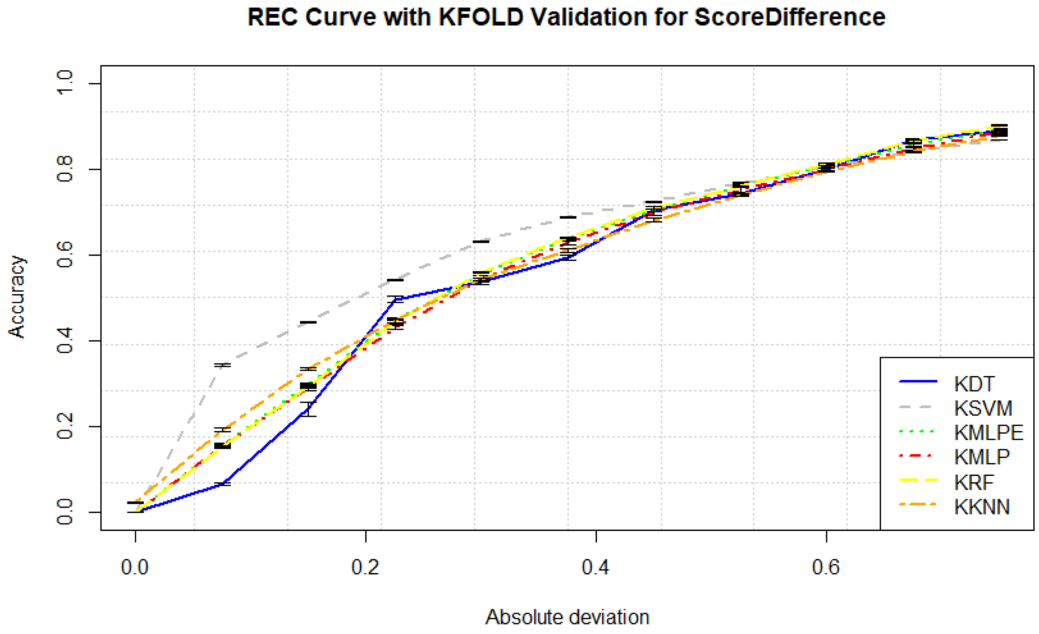

Figure 24: REC Curve for ScoreDifference

Therefore, the REC curve confirms a farther curve from the imaginary diagonal line to the SVM model, plotting the error tolerance curve against the accuracy of the function (i.e. the percentage of points predicted in that tolerance). However, it is important to confirm this graphical visualization with the values in Table 3, since in some cases it is possible to have two or more different models with "exactly" or approximately same curves, being preferred the one with lowest errors.

After considering that the method SVM obtains acceptable prediction results, the stage of knowledge extraction follows. In this way, we will look at the relevant attributes identified in Fig. 25 in descending order of percentage importance.

Knowledge Extraction

According to Silva et al. (2018), when dealing with black box models, it is often challenging to extract knowledge, consequently, methods such as extraction rules and sensitivity analysis (SA) have emerged to deal with this problem, where there is an assessment of importance of the input factors to a method (Saltelli et al. 2000 cited by Silva et al. 2018) and its effects on the results of the model. So, for this research, the DSA method was used, identifying as shown in Fig. 25 the relevant attributes to the model and their influence on the target "ScoreDifference".

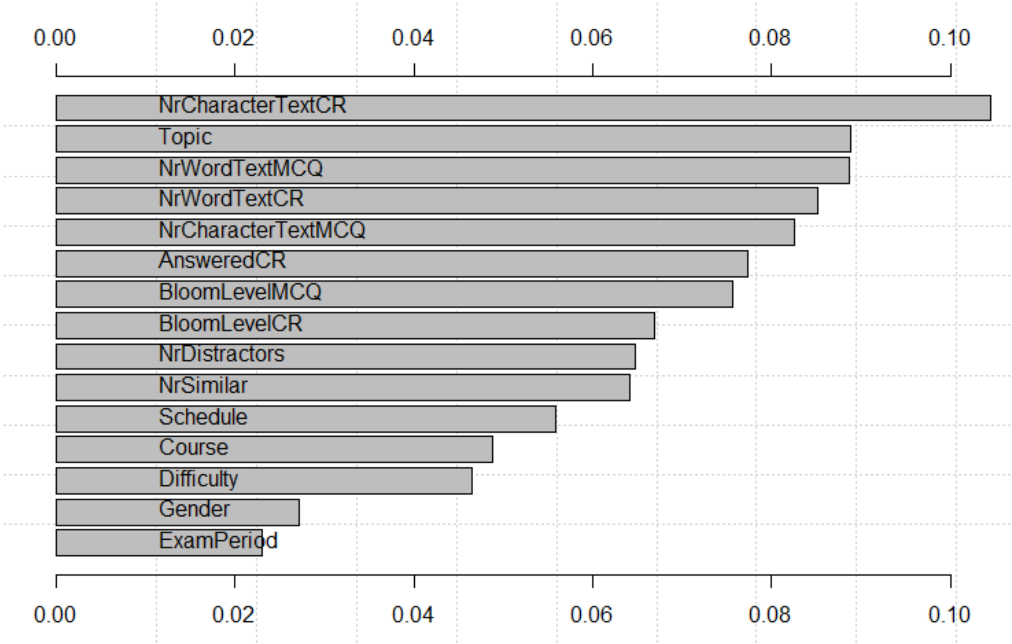

Figure 25: Attributes Relevance

The graphics showing the partial dependency of the attributes for the model are illustrated below, such as their influence on the values of ScoreDifference.

NrCharacterTextCR Attribute

The amount of number of characters refers to the text of the CR question, that is, this attribute with 11% importance, indicates how long the CR question text is. Thus, the Fig. 26 indicates that the higher the text of the CR question (the greater the amount of characters), the higher is the ScoreDifference value in negative grades, that is, the further the score from zero[10] and approaching more negative scores, it favors their MCQ equivalent as the one in which the student is most likely to succeed. For Santos et al. (2011) cited by Almeida (2017), the language and the dimension of the question text influence the ability of a good student perception.

[10] As previously mentioned, zero point indicates those tests performed by the student where the results obtained in CR questions and their equivalent MCQ are the same, meaning the formats are not different

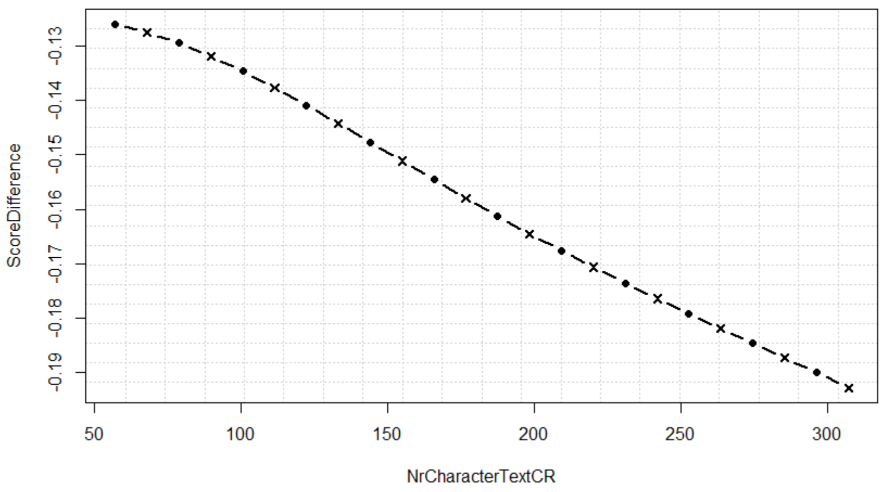

Figure 26: NrCharacterTextCR and ScoreDifference

Therefore, the probability that there is no difference in the scores obtained in both formats is very low, repeating, the higher the number of characters in CR questions, the higher the probability of the student to fail on this type of questions and to succeed on MCQ, and so the lower the probability of concluding that the two formats are equal on returning results.

Topic Attribute

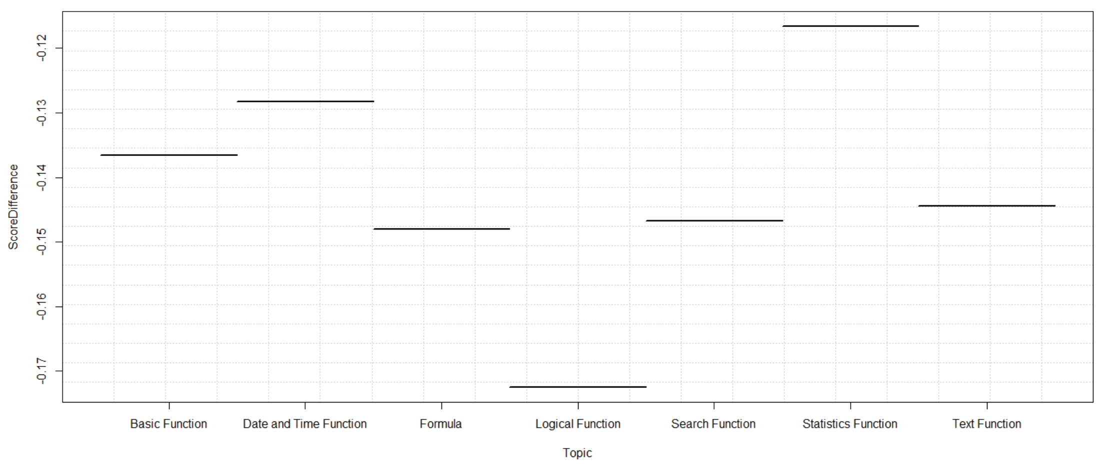

Figure 27: Topic and ScoreDifference

The second most relevant variable is Topic, with approximately 9% importance, it discriminates for each topic the influence they have on the values of the difference of the results in the two formats, so for Statistics Function this affects the variable target in approximating the scale of the values towards the zero point. That is, even assuming negative values it promotes better results in MCQ, and it can lead to a conclusion in which both exam formats are equal in terms of grades even though students have better results in CR. Unlike for example the Logical Function whose influence on ScoreDifference assumes a value of -0.17 further from the zero point and therefore without doubt, students are more likely to succeed in MCQ. According to Almeida (2017), the attribute Topic is a very important factor, since the teacher can obtain a sense of which topics of the subject the students have more difficulties, and therefore where they can be better applied, whether in CR exams or MCQ exams.

NrWordTextMCQ Attribute

By the interpretation of Fig. 28, we can consider that the larger the length of the MCQ text, the higher the probability of being misinterpreted by the student (see Almeida 2017), there is, the higher the number of words in the MCQ text, the higher the difference will be in both formats providing better achievements in CR questions.

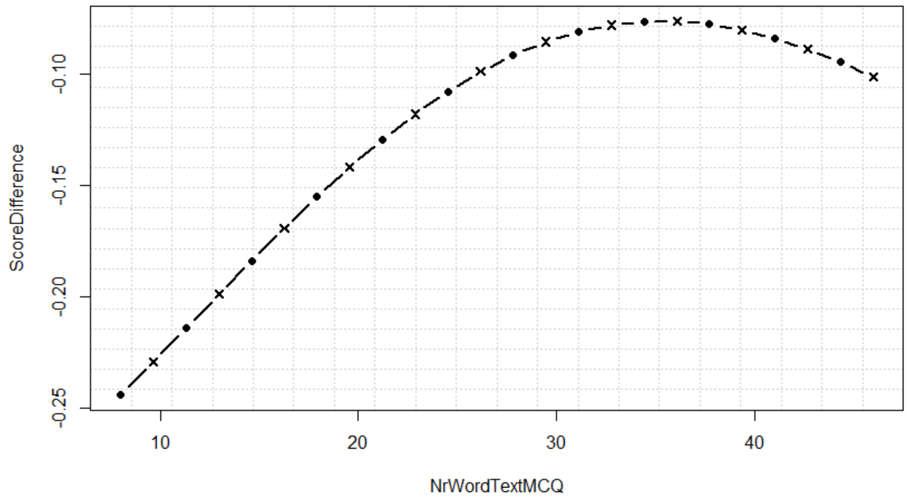

Figure 28: NrWordTextMCQ and ScoreDifference

This attribute is 9% important, even though it takes negative values for the whole sample in the graph, the curve tends to rise toward the zero point each time the number of words in the MCQ text grows.

In fact, Fig. 29 confirms the influence that the number of words in the MCQ text has on the target attribute, where the lowest probabilities (less than 4%) are distributed each time the number of words in the MCQ text increases.

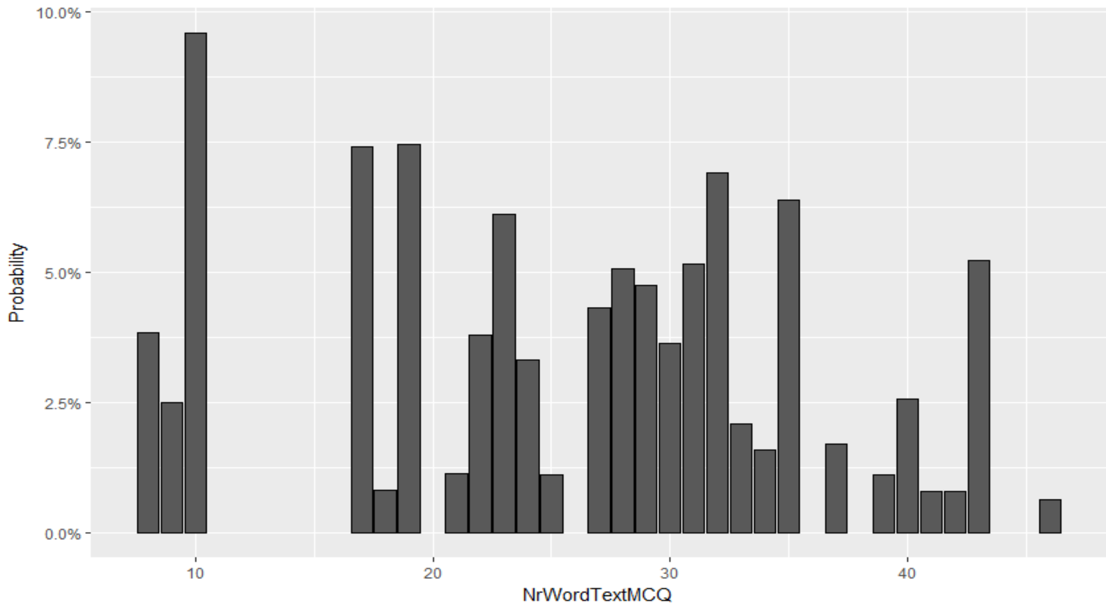

Figure 29: Probability of NrWordTextMCQ

NrWordTextCR Attribute

Following the analogy of the NrWordTextMCQ, it would be correct to state that the higher the number of CR words, the more difficulty the student would have to answer and therefore, better answers would be reached in MCQ, however, the Fig. 30 demonstrates an opposite scenario in which, the higher the number of words in the CR text, the higher the probability that the curve will reach the zero point and distance itself from negative values and may induce that there will be no difference in results in both formats or that students will be more successful in CR as the number of words in the CR text grows.

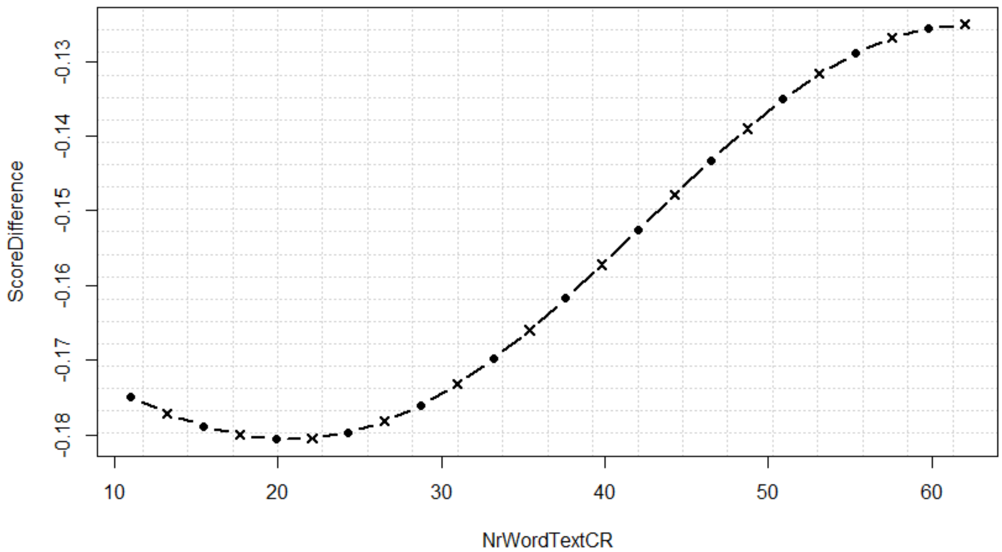

Figure 30: NrWordTextCR and ScoreDifference

NrCharacterTextMCQ Attribute

Since a word is completed by a set of characters, the same association in NrWordTextMCQ attribute can be applied to this attribute (NrCharacterTextMCQ with 8.3% importance), in other words, the higher the number of characters in the MCQ text, the larger the text of the question and therefore more likely to be misinterpreted and the student to fail.

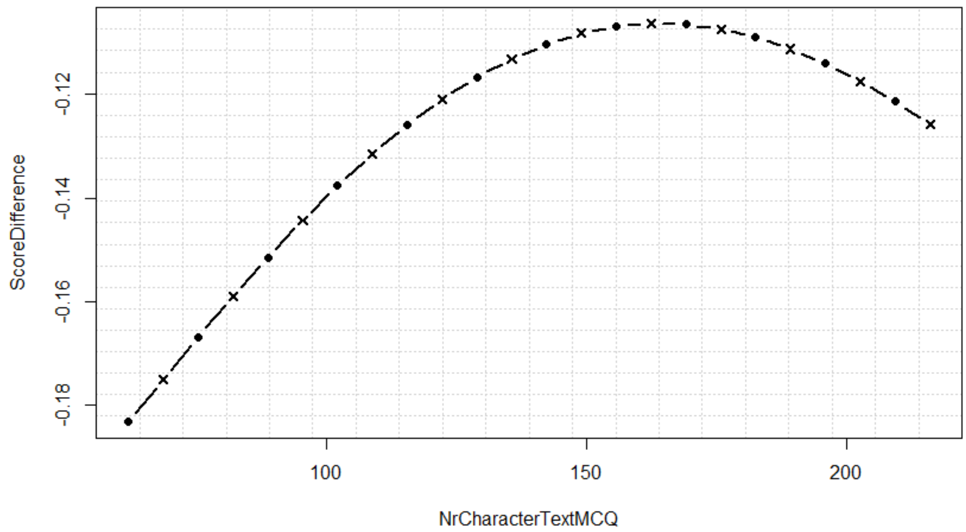

Figure 31: NrCharacterTextMCQ and ScoreDifference

By Fig. 31, the curve represents the variation of ScoreDifference values, and the higher the number of characters in the MCQ text, the closer the values will be to zero, although students are more likely to have better results by performing MCQ. As stated by Dubins et al. (2016) cited by Almeida (2017), the misinterpretation factor of the text can be considered as a factor for the student not to answer the question, since it may be associated with an incorrect reading or difficulty in interpreting a poorly elaborated question.

AnsweredCR Attribute

As mentioned in Data Understanding, this attribute indicates whether the CR question was answered or not, where, a zero score indicates that the student incorrectly answered the question, or if he chose to leave the question blank. Therefore, as shown in Fig. 32, this attribute can influence ScoreDifference (with importance of 7.8%) to the point where, higher probabilities are detected when the student chose to answer the CR question than the moment he chose not to.

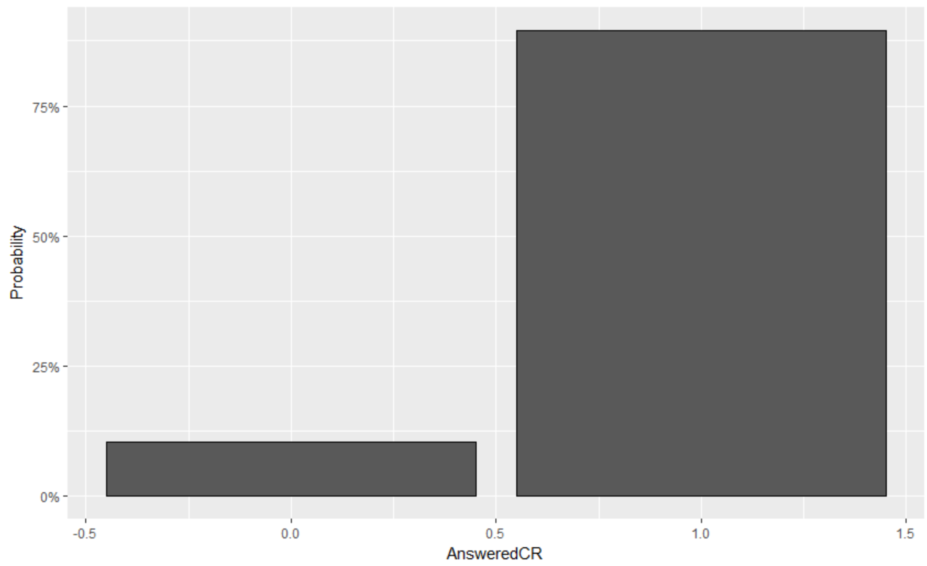

Figure 32: AnsweredCR Probability

It was also possible to make a comparison with the number of words in the question text CR, which in Fig. 33, the distribution of questions not answered (panel represented by zero) was relatively low regarding questions answered, where the smaller the number of CR words, the higher the probability that the student will respond to the CR question.

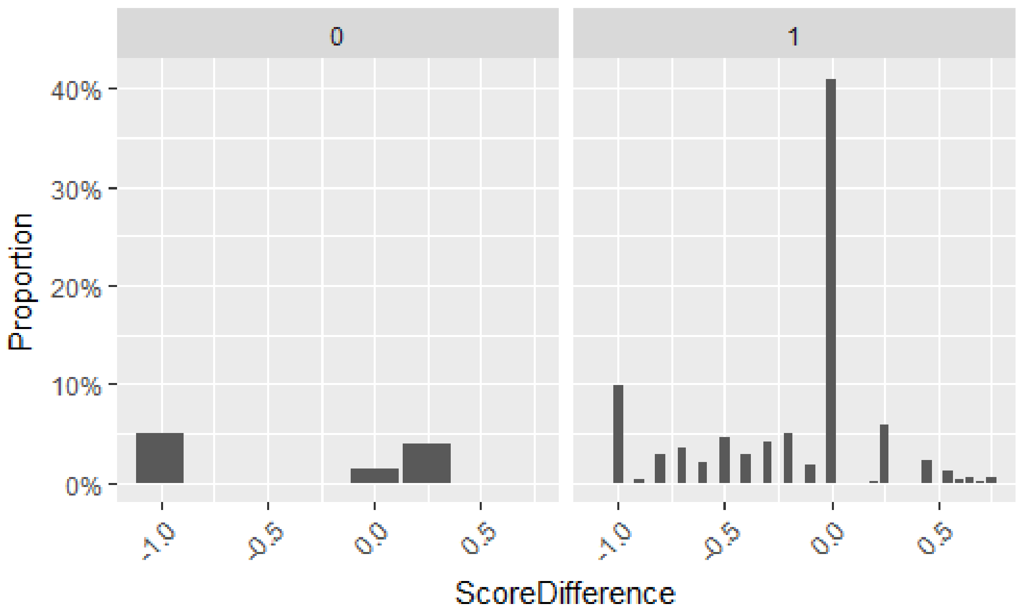

Figure 33: AnsweredCR proportion by ScoreDifference

BloomLevelMCQ and BloomLevelCR Attributes

These attributes with importance of 7.4% and 6.4% respectively, indicate the student's level of learning regarding to educational objectives of the Bloom Taxonomy. Firstly, with the BloomLevelMCQ attribute as the seventh most influential variable representing educational levels for MCQs, it can be verified that for Analysis, Applying and Remembering levels, these can influence the ScoreDifference to values very close to zero, although for these levels the student is more likely to be successful in MCQ, which is not the same with Understanding level, becoming clear from Fig. 34 the distance from zero, thus suggesting that at this level students have a chance of succeed more in MCQ.

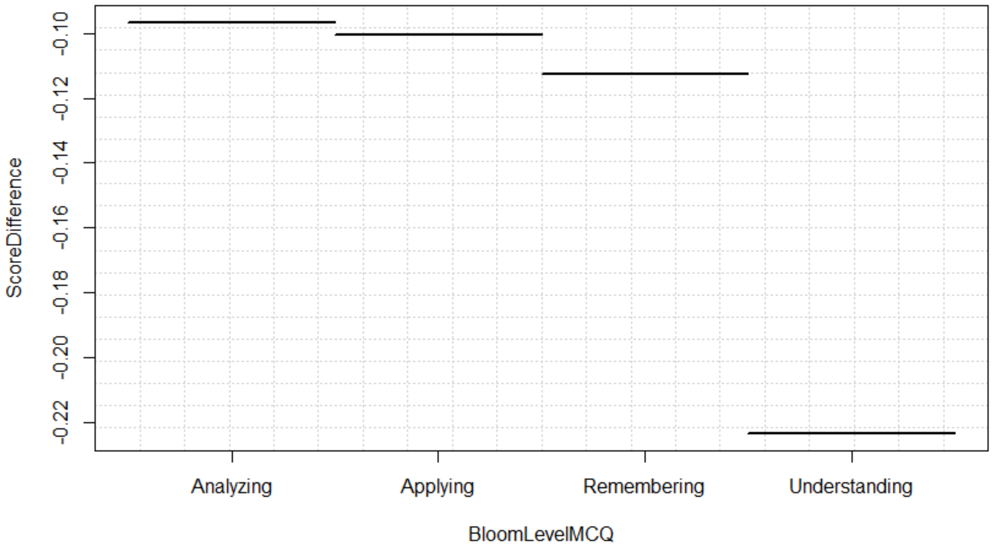

Figure 34: BloomLevelMCQ and ScoreDifference

As for BloomLevelCR, the Analyzing level emerge with higher negative values in the CR formats, which also happens with Applying which relatively higher negative results has, so in these two levels the student has a better chance of succeed in MCQ and not in CR, opposite to Evaluating that approaches values close to zero suggesting either there is no difference in both formats or registering higher results in CR questions performed.

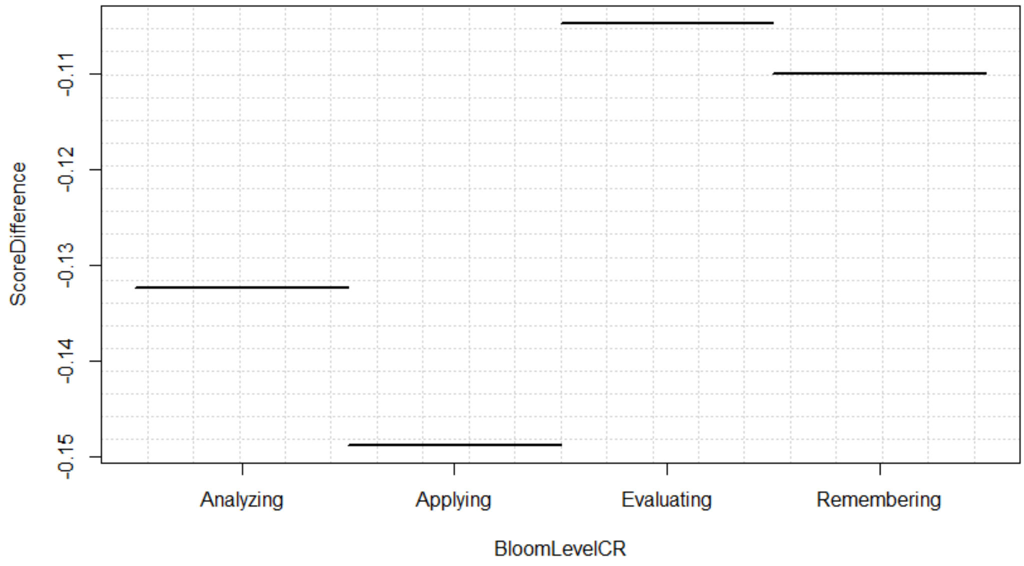

Figure 35: BloomLevelCR and ScoreDifference

As stated by Gronlund, 1998 cited by Salume, et al. (2012), the last two Bloom's objectives would be better evaluated in CR questions, and the first four could be applied in MCQ format. Note that for Remembering it remains the same in both formats, that is, for both MCQ and CR, this level remains at an average score of -0.11 for ScoreDifference, still promoting MCQ as the format in which students return better results.

NrDistractors and NrSimilar Attributes

Because these attributes were complementary, it was decided to analyze them simultaneously, the first with 6.3% of importance indicating how many distractors have to the correct option and the last with 6.2% of importance, the amount of similar options to the correct one.

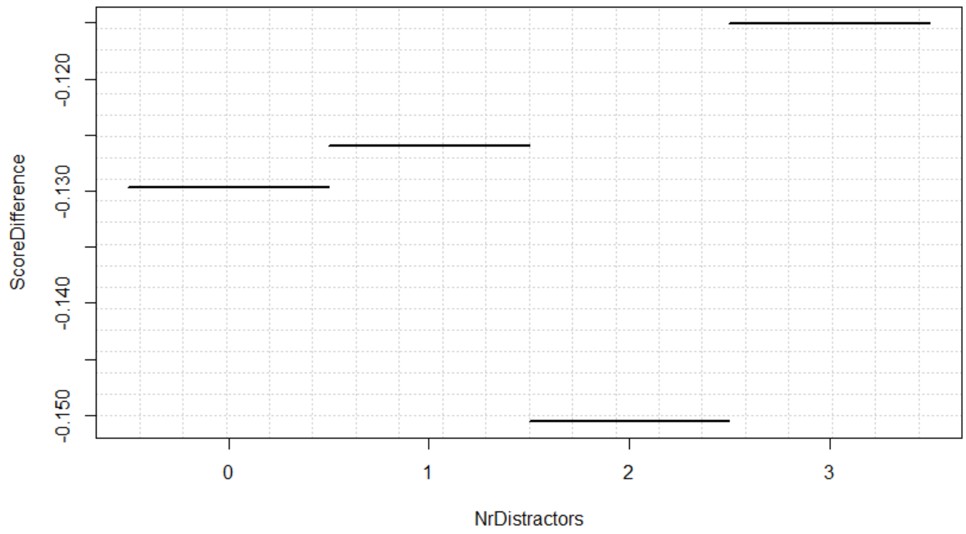

Figure 36: NrDistractors and ScoreDifference

Firstly, questions with two distractors result in a high score difference benefiting the MCQ format, contradictory for example when questions with three distractors tend to approximate the model to the zero point.

In fact, questions that have up to two distractors are more likely to influence the target variable.

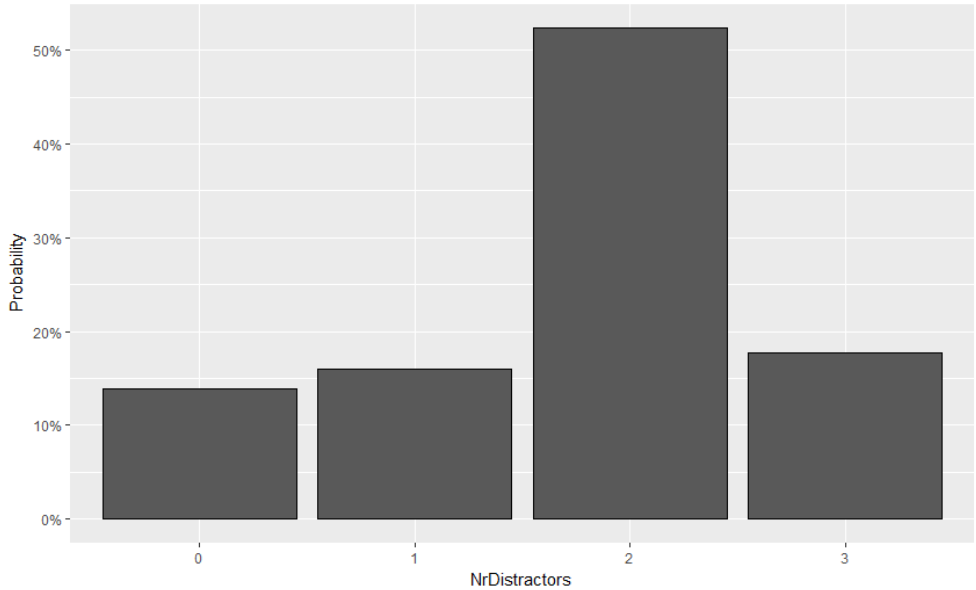

Figure 37: NrDistractors Probability

The same analogy to the NrSimilar attribute, since the scenario proposes only one correct option, as previously said when the student faces at least two distractors is more likely to succeed, then MCQ with only one similar becomes more relevant to justify the difference, where in MCQ the students have more chances to succeed.

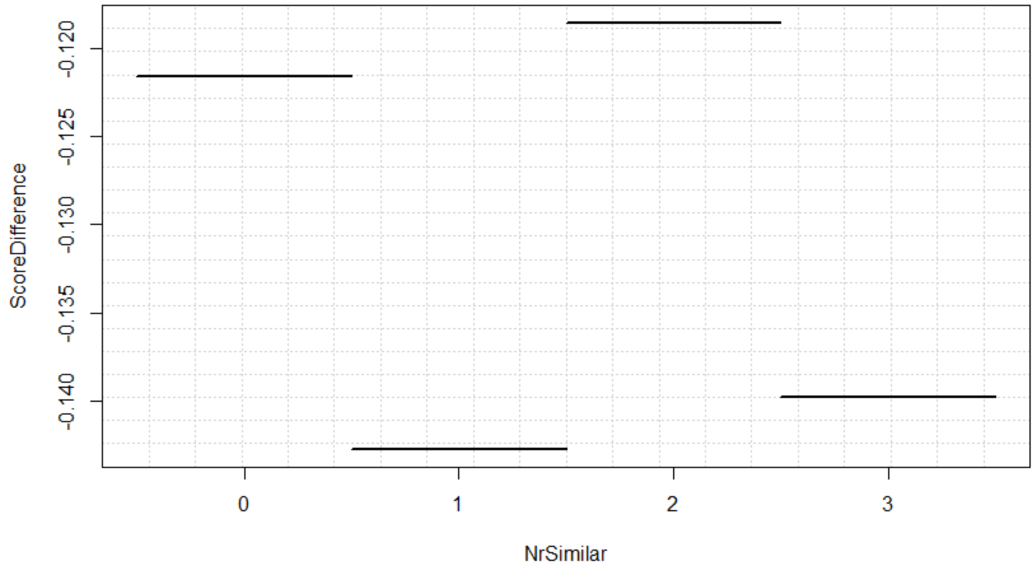

Figure 38: NrSimilar and ScoreDifference

According to Dubins et. al.,(2016) cited by Almeida (2017), the more similar options to the correct answer, the higher the probability of success by guessing, and the less options close to the correct, the higher the probability that student settle the question with less hesitation in "risking".

Conclusively, situations in which MCQ have only one similar option, there is a higher probability that the student will score in MCQ and thus the ScoreDifference will diverge towards the negative values, which it is the contrary when MCQ has two similars, covering two possibilities: having no difference in both formats regarding the number of similar options or even more the student fails in MCQ and benefit CR as the questions with better results.

Schedule Attribute

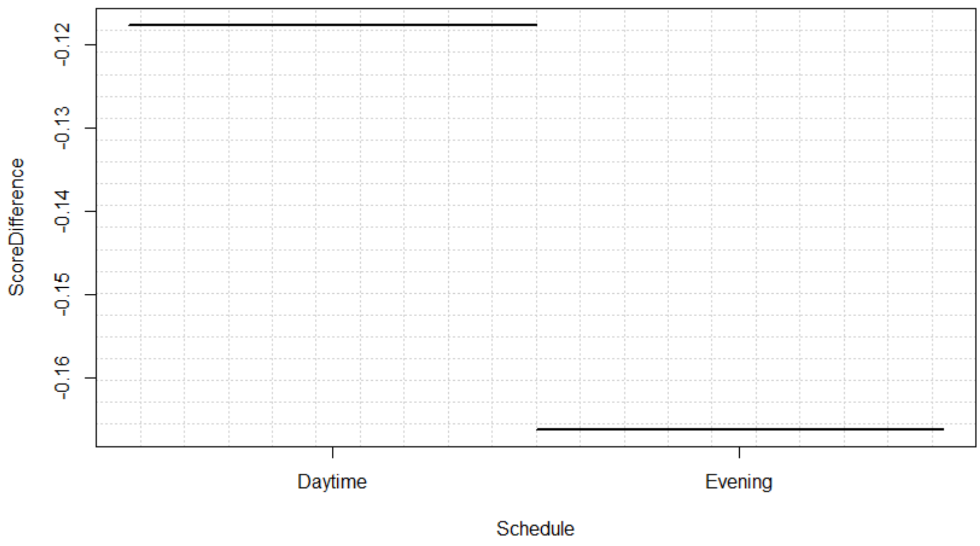

Figure 39: Schedule and ScoreDifference

The Schedule attribute corresponds to the period in which the student learn, being an attribute with importance of approximately 5.6%, it shows that the difference of the results in both exam formats performed by students in Daytime was very close to the zero point, although still indicate better results in MCQ, which is not the same in the Evening period, where the ScoreDifference average is relatively far from zero therefore promoting MCQ as the format with more probability that the student succeed.

Course Attribute

In relation to Course attribute, with 5% of importance, students who attend Institutional Course (IC) had higher probability of succeed in MCQ in an average of -0.18 quite distant from the zero point, followed by CSBM (Computer Science and Business Management), TCE (Telecommunications and Computer Engineering) and lastly Anthropology represented by A.

On the other hand, in CE (Computer Engineering) the average of the results by the difference of the scores in both exam formats approached the zero point, nevertheless, with an average of -0.11, this model suggests that that students attending CE course are more likely to succeed in MCQ.

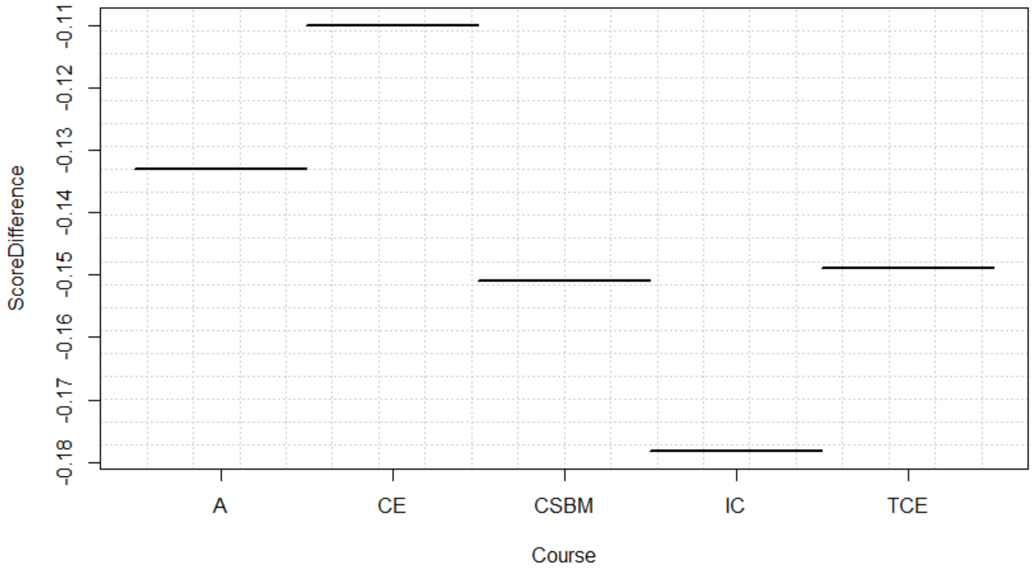

Figure 40: Course and ScoreDifference

It was compared too, the difference of the scores in both formats considering the difficulty of the questions for each course. Therefore, the students of the CE course were more likely to score in MCQ when the difficulty of the questions was Medium, although there were records for correct answers also on Easy questions, as the same to Hard level, where it can be observed that the bars of the graph tend to vary to negative values and with very low records close to zero.

The same analogy can be applied to the following courses, CSBM and TCE, note that the fact some bars are not displayed for Easy and Hard levels in A and IC courses, probably because most of the questions are identified as Medium.

Figure 41: ScoreDifference proportion per Course and Difficulty

Difficulty Attribute

About Difficulty (with 4.5% importance), it can be considered *a priori* that the student find it complex in answering a difficult question and to succeed in it, in this case it remains to verify how the degree of difficulty of the question is relevant to the ScoreDifference, where by the Fig. 42 it is possible to verify that the Hard level is described as being the one that is distant from zero than Easy and Medium, being the level in which the student has higher probability to have correct answers in MCQ, contrary of Easy with an average of approximately -0.11. Note that, the difference of the results in both formats tends to approach zero, assuming there is no notable difference in the student's results, nevertheless, having more correct answers in MCQ.

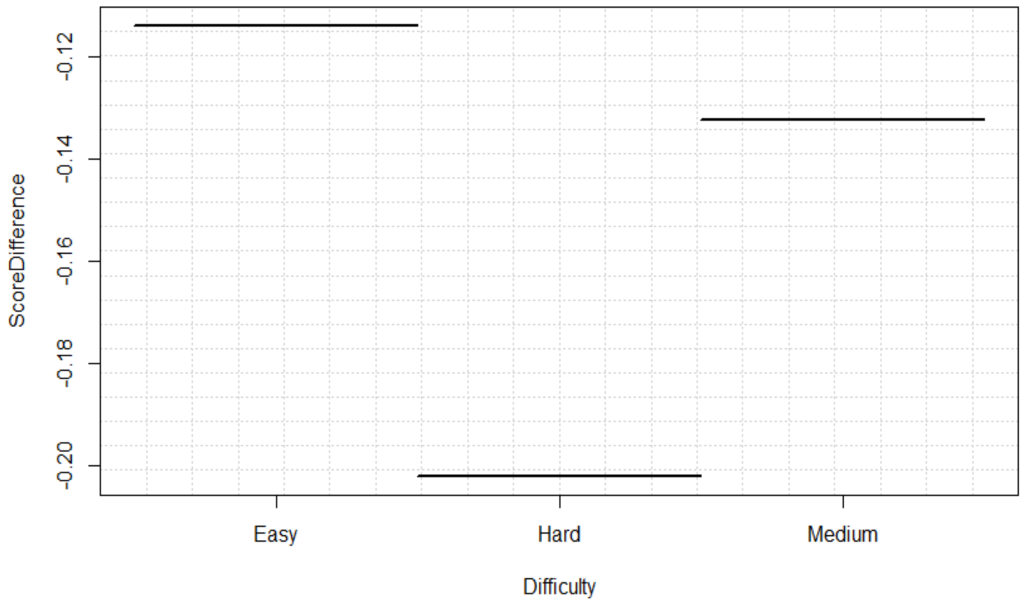

Figure 42: Difficulty and ScoreDifference

Among the 3 levels of difficulty, the higher number of CR questions answered were registered at Medium level followed by Easy and finally Hard where students were more likely to succeed in MCQ, but the decision of the student not to answer CR (for all levels of difficulty), by visualizing the charts may lead to the conclusion that the student has better results in CR or that there is no difference in the question's formats, nevertheless, greater chances of success were registered in MCQ.

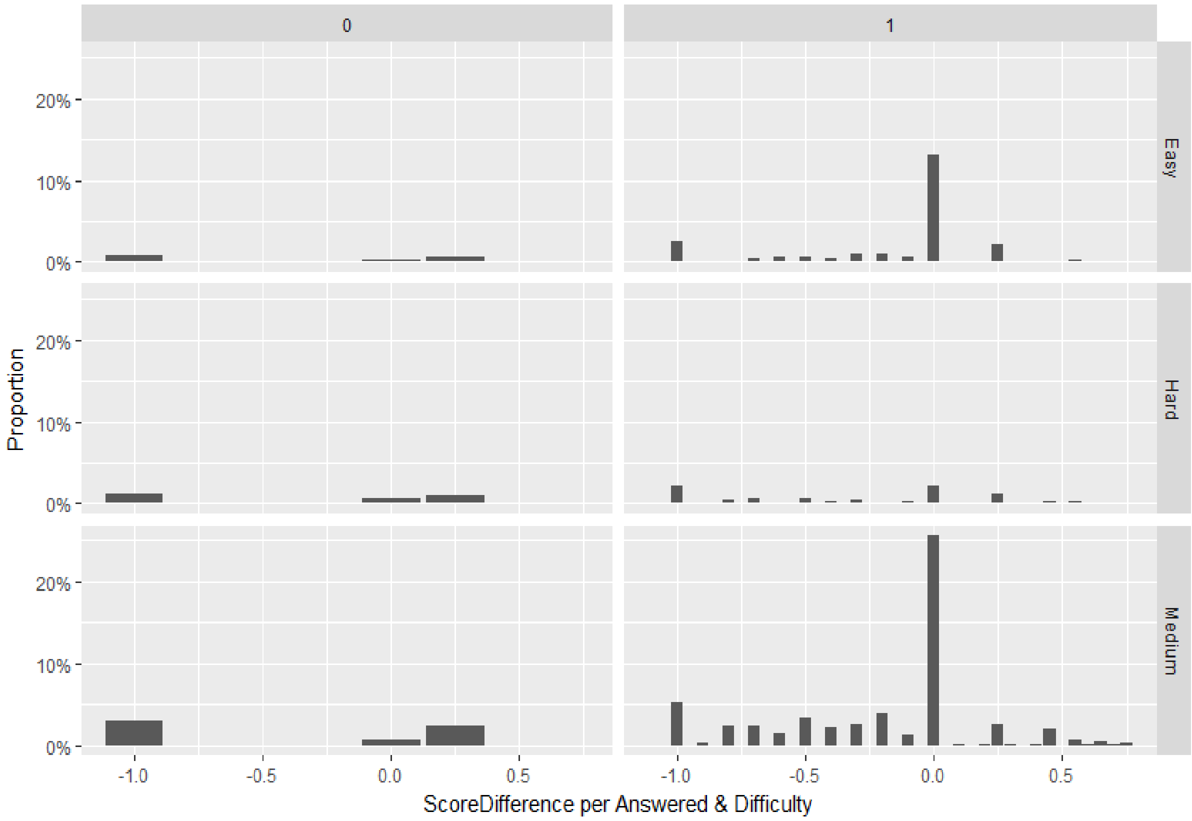

Figure 43: ScoreDifference proportion per AnsweredCR and Difficulty

Gender Attribute

This attribute, with 2.5% importance, as its name suggests, indicates the gender of the student who performed the exam, and it was verified that, although both Female and Male have a higher probability of correct answers in MCQ, the ScoreDifference by students of the sex F tends to approach the zero point, totally opposite to what was registered with M, where these students have more possibilities to succeed in MCQ.

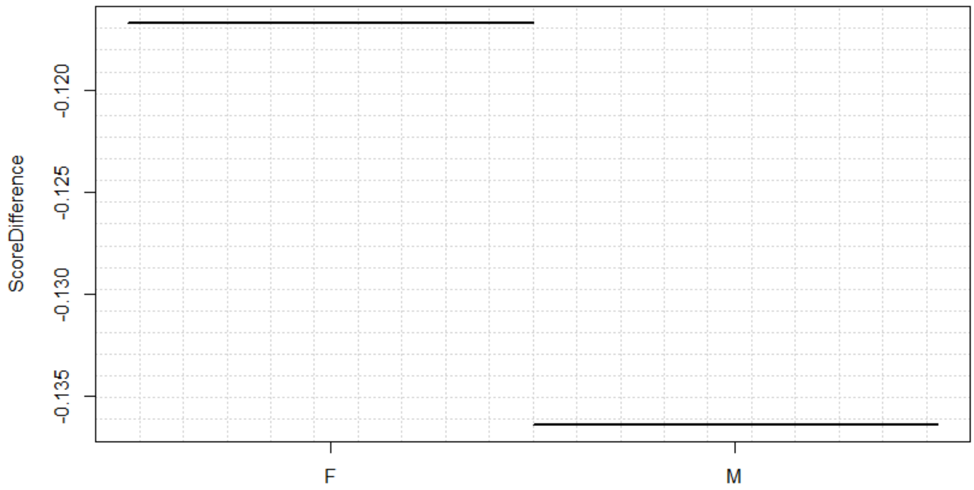

Figure 44: Gender and ScoreDifference

ExamPeriod Attribute

For this attribute (2.5% of importance), the students who performed the exam in the first period were more likely to answer correctly MCQ, although the same happens with those in the second period, the average of the ScoreDifference in the second period approaches towards the zero point, suggesting a scenario in which we assume both formats are equal, but in both periods, the students have more possibilities to succeed in MCQ, in which the students from the first period achieve higher scores in MCQ than those of the second.

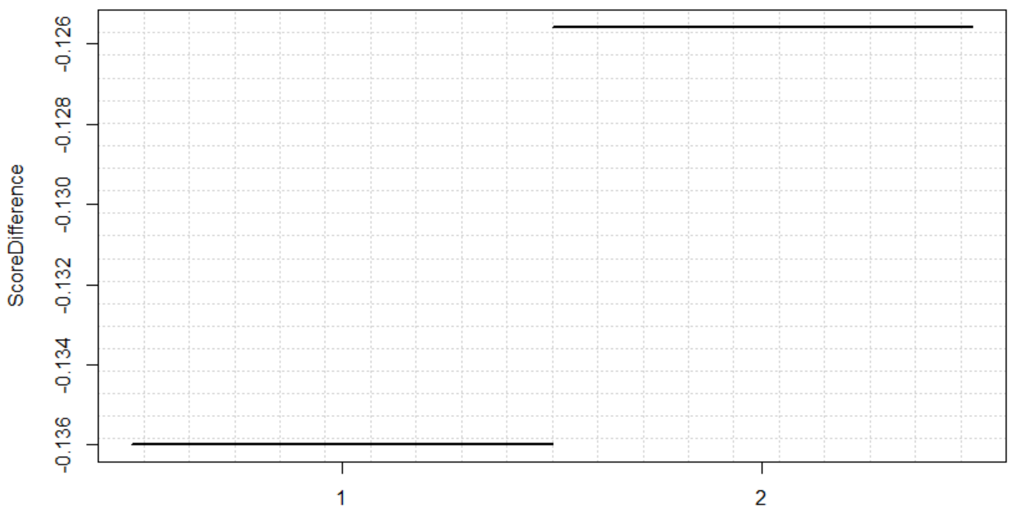

Figure 45: ExamPeriod and ScoreDifference

4.2. Discussion

From several models developed to describe the target ScoreDifference attribute, using a set of input attributes, it was possible to verify by table 3, except for model MLP, the models obtained good results. Thus, the metrics MAE and NMAE calculated were approximately 0.31 and 18% for SVM, a model that presented smaller average of errors and therefore chosen as the one that obtained better performance.

In analyzing the results, the aim was to investigate the relationship between ScoreDifference and other input attributes. The results were analyzed by regression metrics, DSA, graphical illustrations and finally knowledge extraction. It should be noted, Sensibility Analysis (DSA) was applied to verify the relevance of each attribute when influencing the ScoreDifference attribute.

Therefore, the attributes were relatively close to one another, so they were all considered to knowledge extraction. All of them were identified as good attributes to explain the assumption that students have a better chance of succeeding in Multiple Choice Questions (MCQ), but even if they have better success in MCQ, for the attributes NrWordTextCR, BloomLevelMCQ, BloomLevelCR, NrSimilar, NrDistractor, Difficulty, Gender and ExamPeriod, they presented to some cases, values that approached the zero point (having no difference in the exams format in relation to the student's performance).

One of the difficulties experienced was interpreting the results of the NrCharacterTextCR attribute where it contradict the NrWordTextCR attribute, which the first one suggests that the higher the number of characters in the text, the more likely the student score in MCQ, while the second one proposes that the higher the number of words the higher the student success in CR, an unexpected situation since a word is a combination of characters, so the interpretation of both graphs should be in compliance. Unless we consider that in CR questions where the text has a greater amount of excel functions / formulas, the students consider it complicated and fail, whereas, in CR with smaller excel functions, they are more likely to success (see appendix 6 for example).

Moreover, the knowledge extracted from the attribute NrWordTextCR contradicts Santos et al. (2011) cited by Almeida (2017), in which, the language and dimension of the text influences the ability of the student to understand the question, that is, the higher the number of words, the higher is the difficulty of understanding, therefore, the higher the number of words of the CR, the higher the probability the student fail and succeed in MCQ, a scenario that unfortunately is not illustrated by the graph 30.

Chapter 5 – Conclusions

The problem of this project is connected to the question "Constructed Response or Multiple-Choice for Evaluating Excel Questions? That is the Question", in which we considered 283 exams from the academic year 2016/2017, performed on the curricular unit of Advanced Excel of ISCTE-IUL. Then, the activities performed were always focused on achieving the objectives of the work:

- To measure the existing discrepancy between the results obtained by students from CR and MCQ tests;
- To explain the possible reasons that may have influenced the existing discrepancy.

Therefore, implementing Data Mining algorithms for regression problems, the SVM model had better performance, through the calculations in the training and test models (k-fold), and it was possible to verify that both exams are not equal, where the students have better chance of succeed in exams with Multiple Choice Questions (MCQ) format.

In this sequence, the main factors that influenced this discrepancy between the two question formats were identified, always benefiting the MCQs as those in which the student has a better chance of success.

Firstly, a question with long text, determined a difference between the formats, and it was verified that the longer the question text, the higher the probability of the student to fail and to succeed in the equivalent format, except for CR, since it was shown that the higher the CR question text the higher the chance of both formats being equal or the student being more likely to success in this type. Haladyna et al. (2012) cited by Almeida (2017) stated that the elaboration of questions requires efforts, and the teachers must elaborate questions as clearly as possible, avoiding long texts.

On the other hand, the question topic, and difficulty level of the exam were tested and measured, returning results that helps the teacher in identifying which subjects the student is most likely to respond easily, as well as the complexity of the question. The difficulty of the question can not only be measured by the number of formulas/calculations but also considering distractors or similar to the correct question.

For Bloom's Taxonomy Objectives, Simkin and Kuechler (2005) found that results tend to be positive in MCQ if these are developed in Understanding level of taxonomy, and CR respectively in Application, confirming the results of the research, thus, the teacher will know which procedures can help in constructing this type of questions, as the requirements of each level and how can the student reach them successfully, finally reaching high levels of performance for both teacher and student in the learning teaching process.

Since several researchers determined that the use of ICT is an added value in the evaluation process with the application of MCQ (e.g. Azevedo, 2017; Scouller, 1998), then it is clear that the results of this research confirm this premise, since in most cases the student is more likely to succeed in MCQ, however if the teacher prefers to evaluate in CR, the e-assessment may not return satisfactory results, requiring other research approaches or proposing the implementation of exams with both questions formats.

Chapter 6 – Limitation and Future Research

From the results obtained, it was possible to verify the difference in the exams formats in relation to the student performance, through the application of Data Mining techniques and tools for the project development.

An explanatory model was created to help the teachers in the elaboration of questions and which format they must consider in association to the objectives and results that are expected from the student.

As mentioned, only Advanced Excel exams were considered, thus limiting other teachers from different curricular units that requires this model to elaborate their exams. Another limitation, the model was able to identify only the average of difference of scores between the two question formats, which factors influence the probabilities of the students to be succeed in MCQ, where would be interesting to understand also in which values these factors would approach zero or identify CR questions as those in which students have the probability to achieve better results.

However, elements like number of similar, distractor, question text characters, Bloom's levels of learning, can be useful to other curricular units, since they are common attributes in conceiving questions for evaluation.

Still in Bloom's taxonomy levels, during the elaboration of the questions the teacher may have a different objective in relation to the objectives of each level, using a verb from a cognitive level that would not be appropriate to evaluate the student at that moment. Thus, in this research only the verb in the question was considered to classify the taxonomy level.

For future work, it is possible to expand from this research other methods for testing and analysis, as well as to expand to other curricular units, or to apply more rigorously the students' levels of learning, so that they must develop diversified cognitive abilities and those that are asked to demonstrate.

This study will attend a work in which the block of exams with CR questions can be verified (since one of the examples of the MCQ block was studied by Almeida ,2017) to complement the area of investigation and to return greater performance in the elaboration of more suitable evaluation to the students, and increasing teaching-learning process.

References

Abreu, J. M. (2016). *Modelos de Data Mining para Suporte a Avaliações Sensoriais do Vinho.* Instituto Superior de Engenharia do Porto, Porto.

Almeida, D. P. (2017). *Fatores de sucesso na avaliação de questões de escolha múltipla: o caso de exames de Excel.* Dissertação, ISCTE-IUL, Departamento de Ciências e Tecnologias da Informação.

Azevedo, J. M. (2017). *Avaliação sumativa em matemática no Ensino Superior com recurso a questões de escolha-múltipla: uma abordagem utilizando a metodologia investigação-ação.* Tese para obtenção do Grau de Doutor em Educação, UNIVERSIDADE DA BEIRA INTERIOR, Ciências Sociais e Humanas, Covilhã.

Barbosa, A. S., & Rolim, L. L. (2017). *Previsão das notas dos filmes no Internet Movie Database (IMDb).* Retrieved from Ensinando Máquinas - Inteligência Computacional para Mineração de Dados: https://ensinandomaquinasblog.wordpress.com/category/sem-categoria/page/2/

Brito, D. M., Júnior, I. A., Queiroga, E. V., & Rêgo, T. G. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. *BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - SBIE),* 882-890. doi:http://dx.doi.org/10.5753/cbie.sbie.2014.882

Buchweitz, B. (1975). TESTES DE MÚLTIPLA ESCOHA E DE RESPOSTA LIVRE EM FÍSICA GERAL. *XXVI Reunião Anual da SBPC* (pp. 3-6). Recife: Cadernos de Pesquisa/16.

Camilo, C. O., & Silva, J. C. (2009). *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.* Technical. Retrieved from http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf

Camilo, H., & Silva, J. A. (2008). *OS TESTES DE ESCOLHA MÚLTIPLA (TEM).* Universidade de Coimbra, Departmento de Educação Médica da Faculdade de Medicina. Essências EDUcare.

Čandrlić, S., Katić, M. A., & Dlab, M. H. (2014). Online vs. Paper-Based Testing: A Comparison of Test Results. *MIPRO*, pp. 775-780.

Chan, N., & Kennedy, P. E. (2002). Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and "Equivalent" Constructed-Response Exam Questions. *Southern Economic Journal, 68*(4), 957-971. doi:10.2307/1061503

Chapman, P., Clinton, J., Kerber, R., T., K., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0. In *SPSS* (pp. 1-74). Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Cherinda, N. (2016). ATITUDES DOS ESTUDANTES DO ENSINO À DISTÂNCIA EM RELAÇÃO AO USO DO EXCEL NA APRENDIZAGEM DA

ESTATÍSTICA NA PLATAFORMA AULANET. *LYALOSHO: Revista Científica de EAD, 1*.

Cortez, P., & Silva, A. (2008). *USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE*. Universidade do Minho, Information Systems/ Algoritmi R&D Centre, Guimarães.

Cymrot, R. (2006). A UTILIZAÇÃO DO PROGRAMA EXCEL NO APRENDIZADO DE TÉCNICAS ESTATÍSTICAS UTILIZADAS NA METODOLOGIA SEIS SIGMA. *World Congress on Computer Science, Engineering and Technology Education*, (pp. 607-610). São Paulo. Retrieved from http://meusite.mackenzie.com.br/raquelc/WCCSETE2006Excel.pdf

Denari, G. B., Saciloto, T. R., & Cavalheiro, É. T. (2016). AVALIAÇÃO DO USO DE PLANILHAS COMPUTACIONAIS COMO UMA FERRAMENTA DIDÁTICA EM QUÍMICA ANALÍTICA QUALITATIVA. *Educação, 39*(3), 371-375. Retrieved from http://dx.doi.org/10.5935/0100/4042.20160029

Ferraz, A. P., & Belhot, R. V. (2010). Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. *SciELO*, 421-431.

França, R. S., & Amaral, H. J. (2013). Mineração de Dados na Identificação de Grupos de Estudantes com Dificuldades de Aprendizagemno Ensino de Programação. *RENOTE - Revista Novas Tecnologias na Educação, 11*(1). doi:http://dx.doi.org/10.22456/1679-1916.41634

Gama, J., Carvalho, A. P., Faceli, K., Lorena, A. C., & Oliveira, M. (2012). *Extração de Conhecimento de Dados - Data Mining* (1 ed.). Lisboa: Edições Sílabo, Lda.

Hickson, S., & Reed, W. R. (2009). *DO CONSTRUCTED-RESPONSE AND MULTIPLE-CHOICE QUESTIONS MEASURE THE SAME THING?* Working Paper No. 08/2009, University of Catenbury, Department of Economics and Finance, Christchurch, New Zealand.

Hudson, R. D. (2012). Is there a relationship between chemistry performance and question type, question content and gender? *Science Education International, 23*(1), pp. 56-83.

Janert, P. K. (2010). *Data Analysis with Open Source Tools.* O'Reilly Media, Inc.

Kastner, M., & Stangl, B. (2011). Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter? *Procedia - Social and Behavioral Sciences, 12*, pp. 263-273. Retrieved from www.sciencedirect.com

Kuechler, W., & Simkin, M. G. (2003). How Well Do Multiple Choice Tests Evaluate Student Understanding in Computer Programming Classes? *Journal of Information Systems Education, 14*(4), 389-399.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. *Journal of Educational Measurement, 31*(3), 234-250.

Maresca, M. (2016). The Spreadsheet Space: Eliminating the Boundaries of Data Cross-Referencing. *IEEE Xplore Digital Library, 49*(9), 78-85. doi:10.1109/MC.2016.272

Miranda, P. H. (2015). *AVALIAÇÃO DA APRENDIZAGEM: MÚLTIPLA ESCOLHA VERSUS QUESTÕES ABERTAS EM COMPUTADOR VERSUS PAPEL.* Dissertação, Pontifícia Universidade Católica de Goiás, Escola de Ciências Sociais e Saúde: Curso de Psicologia, Goiânia.

Moreno, A. C. (2015). *Análise de Sentimentos na Classificação de Comentários Online Aplicando Técnicas de Text Mining.* ISCTE - IUL, Departamento de Ciências e Tecnologias de Informação, Lisboa.

Moro, S., Cortez, P., & Rita, P. (2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications, 26*(1), pp. 131-139.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011* (pp. 117-121). Eurosis.

Mullen, K., & Schultz, M. (2012). Short Answer Versus Multiple Choice Examination Questions for First Year Chemistry. *International Journal of Innovation in Science and Mathematics Education*, 1-18.

Nossa, V., & Chagas, J. F. (1997). Usando programação linear na contabilidade decisorial. *Revista Brasileira de Contabilidade (RBC)* .

Ohri, A. (2013). *Visual Guides to CRISP-DM ,KDD and SEMMA*. Retrieved from Decision Stats: https://decisionstats.com/2013/04/10/visual-guides-to-crisp-dm-kdd-and-semma/

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education*.

Pinto, A. C. (2001). FACTORES RELEVANTES NA AVALIAÇÃO ESCOLAR POR PERGUNTAS DE ESCOLHA MÚLTIPLA. *V Jornadas Psicopedagógicas de Gaia* (pp. 23-44). Porto: Psicologia, Educação e Cultura, 5.

Quintela, H. (2005). *Sistemas de Conhecimento Baseados em Data Mining: Aplicação à Análise de Estabilidade de Estruturas Metálicas.* masterThesis, Universidade do Minho, Escola de Engenharia. Retrieved from http://hdl.handle.net/1822/6255

Salume, P. K., Braga, J. O., Latini, F. H., Chequer, K. M., Gonçalves, R. F., & Ribeiro, D. Q. (2012). O ENADE Avalia o Administrador de Acordo com o Perfil Exigido pelo MEC? Uma Análise sob a Perspectiva da Taxionomia de Bloom. *XXXVI Encontro da ANPAD.* Rio de Janeiro.

Scouller, K. (1998.). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*, pp. 453 – 472.

Sheaffer, E. A., & Addo, R. T. (2012). Pharmacy Student Performance on Constructed-Response Versus Selected-Response Calculations Questions. *AMERICAN JOURNAL OF PHARMACEUTICAL EDUCATION, 77*(1). doi:10.5688/ajpe7716

Silva, A. T., Moro, S., Rita, P., & Cortez, P. (2018). Unveiling the features of successful eBay smartphone sellers. *Journal of Retailing and Consumer Services, 43*, 311-324. Retrieved from https://doi.org/10.1016/j.jretconser.2018.05.001

Silva, E. A. (2009). O ENSINO DE ESTATÍSTICA NO CURSO DE PEDAGOGIA USANDO O EXCEL COMO INSTRUMENTO FACILITADOR DA APRENDIZAGEM. *Revista Eletrônica Interdisciplinar, 2*.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-Choice Tests and Student Understanding: What is the Connection? *Decision Science Journal of Innovative Education, 3*(1), 73-98.

Thomé, A. (2017). Redes Neurais - Uma ferramenta para KDD e Data Mining. *apostila_kdd_mbi*. DocGo. Retrieved from https://docgo.net/apostila-kdd-mbi

Xu, X., Kauer, S., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology, 2*(2), 147-158.

Annex and Appendices

Appendix 1: Excel Functions and Formulas[11]

Topic	Functions Examples	Total Questions		
Basic Functions	SUMIF; EXP; PRODUCT; SUMPRODUCT	1462	**CR**	743
			MCQ	719
Date and Time Function	DATE; DATEDIF; DAY; HOUR; TODAY	763	**CR**	382
			MCQ	381
Formula	[Mix of Formulas and Functions]	105	**CR**	31
			MCQ	74
Logical Function	IF; AND; FALSE; IFS; NOT; OR; TRUE	565	**CR**	282
			MCQ	283
Search Function	VLOOKUP; ADDRESS; AREA; SELECT; COL; PROCH; INDEX; MATCH	679	**CR**	360
			MCQ	319
Statistics Functions	COUNTIF; AVG; COUNTVAL; VAR; MAX/MIN	1438	**CR**	708
			MCQ	730
Text Functions	RIGHT; CONCATENATE; CODIGO; CONCAT; EXACT; SEG.TEXT	648	**CR**	324
			MCQ	324
	Total	**5660**	**CR**	**2830**
			MCQ	**2830**

[11] Available in: https://support.office.com/en-us/article/excel-functions-by-category-5f91f4e9-7b42-46d2-9bd1-63f26a86c0eb

Appendix 2: Features Description

Table	Name		Include (DataSet)	Type		Description
	DataBase	**DataExcel**		**DataBase**	**DataSet**	
Answer	IdAnswer	IDAnswer	✘	AutoNumber	Numeric	Answer Identification
	Answered	AnsweredCR	✓	Yes/No	Logical	If the student answered the CR question or not
	ScoreCR	ScoreCR	✘	Number	Numeric	Score achieved by the student on his/her CR Answer
	ScoreMCQ	ScoreMCQ	✘	Number	Numeric	Score achieved by the student on his/her MCQ Answer
	MCQP	MCQP	✘	Number	Factor	MCQ Option picked by the student
	MCQD		✘	Yes/No		Penalty if the student missed the correct option
Exam	IdExam	IdExam	✘	AutoNumber	Numeric	Exam Identification
	Year		✘	Short Text		Year, Semester and Period when the exam was made
	Semester		✘	Short Text		
	ExamPeriod	ExamPeriod	✓	Number	Factor (2)	
	Subject		✘	Short Text		{Advanced Excel}
	DurationTotal		✘	Number		Duration of each group, and total of exam in minutes
	DurationFirstPart		✘	Number		
	DurationSecondPart		✘	Number		
	ExamVariant	ExamVariant	✘	Short Text	Character	Variant of the exams
Question	IdQuestion	IdQuestion MCQeq	✘	AutoNumber	Numeric	Multiple Choice Question Identification
	Text	TextMCQ	✘	Long Text	Character	Text of the Multiple Choice Question (see Table for details)

	Topic	Topic	✓	Short Text	Factor (7)	Discipline Topic (7) based in Excel Formulas (see Table for details)
	QuestionType		✘	Short Text		CR or MCQ (see Table for details)
	Difficulty	Difficulty	✓	Short Text	Factor (3)	{Easy, Medium, Hard}
	TextChoice 1	TextChoice 1	✘	Short Text	Character	MCQ´s first option text
	TypeChoice1	TypeChoice1	✘	Short Text	Factor	{Correct, Similar, Distractor}
	TextChoice 2	TextChoice 2	✘	Short Text	Character	MCQ´s second option text
	TypeChoice2	TypeChoice2	✘	Short Text	Factor	{Correct, Similar, Distractor}
	TextChoice 3	TextChoice 3	✘	Short Text	Character	MCQ´s third option text
	TypeChoice3	TypeChoice3	✘	Short Text	Factor	{Correct, Similar, Distractor}
	TextChoice 4	TextChoice 4	✘	Short Text	Character	MCQ´s fourth option text
	TypeChoice4	TypeChoice4	✘	Short Text	Factor	{Correct, Similar, Distractor}
	BloomLevel	BloomLevelMCQ	✓	Short Text	Factor (4)	Bloom´s Taxonomy Levels (see Table for details)
	NrChoices		✘	Number		Number of choices to the MCQ
	Image		✘	Yes/No		If the question was supported or not by an image or table
Student	IdStudent	IDStudent	✘	AutoNumber	Numeric	Student Identification
	NrStudent	NrStudent	✘	Number	Numeric	Student Number in the University
	Name		✘	Long Text		Student Name
	Gender	Gender	✓	Short Text	Factor (2)	{M/F}
	Class	Class	✘	Short Text	Factor	{EIA2, IA1, …} (see Table for details)
	Course	Course	✓	Short Text	Factor (5)	{CE, CSBM, IC, …} (see Table for details)

Subject		✘	Short Text		{Advanced Excel}
Year		✘	Short Text		Year, Semester and Period when the student made the exam
ExamPeriod		✓	Number		
ExamYear		✘	Short Text		
GradeCR	GradeCR	✘	Number	Numeric	Grade achieved by the student on his/her CR Group (in a 10 grade scale)
GradeMCQ	GradeMCQ	✘	Number	Numeric	Grade achieved by the student on his/her MCQ Group (in a 10 grade scale)
GradeTotal Exam		✘	Number		Total grade in exam (both group)
Schedule	Schedule	✓	Short Text	Factor (2)	{Daytime/ Evening}
Status	Status	✘	Short Text	Factor	{Full-Time, Worker, Part-Time, …}
ExtraCurricular	ExtraCurricular	✘	Yes/ No	Character	If the subject of the exam is an extracurricular
Incoming Mobility		✘	Yes/ No		If the student comes from another University
Mobility Agreement		✘	Yes/ No		If the student has a discount or agreement on his/her mobility
Degree		✘	Short Text		{Licentiate}
	IdQuestion CR	✘		Numeric	Constructed Response Question Identification
	MCQCorrect	✘		Factor	MCQ correct option
	NrSimilar	✓		Factor (4)	Number of choices similar to the correct answer
	NrDistractors	✓		Factor (4)	Number of choices considered distractors to the correct answer

TextCR	✗	Character	Text of the CR Question (see Table for details)
NrWordTextCR	✓	Numeric	Number of words and characters of the question
NrCharacterTextCR	✓	Numeric	
NrWordTextMCQ	✓	Numeric	
NrCharacterTextMCQ	✓	Numeric	
BloomLevelCR	✓	Factor (4)	Bloom´s Taxonomy Levels (see Table for details)
ScoreDifference	✓	Numeric	Difference between CR and MCQ marks

Appendix 3: Class and Course Attributes

Class	Course	Description Course
LEI-PL	CE	Computer Engineering
LEIA1/CI-CT02	CE	Computer Engineering
LEI2/CI-CT02	CE	Computer Engineering
LAA1	A	Anthropology
I-PLA2/CI-CT08	CSBM	Computer Science and Business Management
IPLA2	CSBM	Computer Science and Business Management
I-PLA1/CI-CT08	CSBM	Computer Science and Business Management
I-PLA1/CI-CT02	CSBM	Computer Science and Business Management
I-PLA1	CSBM	Computer Science and Business Management
IB4/CI-CT08	CSBM	Computer Science and Business Management
IB2	CSBM	Computer Science and Business Management
IA-PLA2/CI-CT02	CSBM	Computer Science and Business Management
IA4	CSBM	Computer Science and Business Management
IA3	CSBM	Computer Science and Business Management
IA2/CI-CT08	CSBM	Computer Science and Business Management
IA2/CI-CT02	CSBM	Computer Science and Business Management
IA2	CSBM	Computer Science and Business Management
IA1	CSBM	Computer Science and Business Management
ETC1PL/CI-CT08	TCE	Telecommunications and Computer Engineering
ETB1	TCE	Telecommunications and Computer Engineering
ETA4/CI-CT03	IC	Institutional Course
ETA4/CI-CT02	TCE	Telecommunications and Computer Engineering
ETA4	TCE	Telecommunications and Computer Engineering
ETA3/CI-CT08	TCE	Telecommunications and Computer Engineering
ETA3	TCE	Telecommunications and Computer Engineering
ETA2	TCE	Telecommunications and Computer Engineering
ETA1/CI-CT09	TCE	Telecommunications and Computer Engineering
ETA1	TCE	Telecommunications and Computer Engineering
ETA	TCE	Telecommunications and Computer Engineering
EIC2/CI-CT08	CE	Computer Engineering
EIB2PL	CE	Computer Engineering
EIB2	CE	Computer Engineering
EIAPL2	CE	Computer Engineering
EIA4/CI-CT08	CE	Computer Engineering
EIA4/CI-CT04	CE	Computer Engineering
EIA4/CI-CT02	CE	Computer Engineering
EIA4	CE	Computer Engineering
EIA3/CI-CT08	CE	Computer Engineering
EIA3	CE	Computer Engineering
EIA2PL/CI-CT09	CE	Computer Engineering
EIA2PL/CI-CT08	CE	Computer Engineering
EIA2PL/CI-CT02	CE	Computer Engineering

EIA2PL	CE	Computer Engineering
EIA2/CI-CT09	CE	Computer Engineering
EIA2/CI-CT03	CE	Computer Engineering
EIA2/CI-CT02	CE	Computer Engineering
EIA2	CE	Computer Engineering
EIA1PL/CI-CT08	CE	Computer Engineering
EIA1PL	CE	Computer Engineering
EIA1/CI-CT02	CE	Computer Engineering
EIA1	CE	Computer Engineering
CI-CT03/CI-CT08	IC	Institutional Course

Appendix 4: Bloom´s Taxonomy Categories and Verbs[12]

Old Version	New Version	Description	Verbs
Creating	Evaluating	• Builds a structure or pattern from diverse elements. Put parts together to form a whole, with emphasis on creating a new meaning or structure. • Can the student create new product or point of view?	categorizes, combines, composes, creates, devises, designs, explains, generates, modifies, organizes, plans, rearranges, reconstructs, relates, reorganizes, revises, rewrites, summarizes, tells, writes
Evaluating	Synthesis	• Make judgments about the value of ideas or materials. • Can the student justify a stand or decision?	appraises, compares, concludes, contrasts, criticizes, critiques, defends, describes, discriminates, evaluates, explains, interprets, justifies, relates, summarizes,
Analyzing	Analysis	• Separates material or concepts into component parts so that its organizational structure may be understood. Distinguishes between facts and inferences. • Can the student distinguish between the different parts?	analyzes, compares, contrasts, diagrams, deconstructs, differentiates, discriminates, distinguishes, identifies,

[12] Adapted from: http://www.nwlink.com/~donclark/hrd/bloom.html and https://www.pdc.edu/wp-content/uploads/2012/01/Bloom-Taxonomy_2012.pdf

			illustrates, infers, outlines, relates, selects, separates
Applying	Application	• Use a concept in a new situation or unprompted use of an abstraction. • Can the student use the information in a new way?	applies, changes, computes, constructs, demonstrates, discovers, manipulates, modifies, operates, predicts, prepares, produces, relates, shows, solves
Understanding	Comprehension	• Comprehending the meaning, translation, interpolation, and interpretation of instructions and problems. State a problem in one's own words. • Can the student explain ideas or concepts?	comprehends, converts, defends, distinguishes, estimates, explains, extends, generalizes, gives an example, interprets, paraphrases, predicts, rewrites, translates
Remembering	Knowledge	• Recall or retrieve previous learned information. • Can the student recall or remember the information?	defines, describes, identifies, knows, labels, lists, matches, names, outlines, recalls, recognizes, reproduces, selects, states

Appendix 5: CR and equivalent MCQ examples/ Question Difficulty

Example of type of question, equivalent question and question difficulty

1. Knowing that bonuses are given to sellers who bill customers for a day over € 400, indicate in H3 the formula (copied to the remaining lines) that identifies the winning sellers with "X". *[CR Question]*

2. [Easy] Knowing that bonuses are given to sellers who bill more than € 400, indicate which of the following formulas (same to the remaining lines) allows you to put "X" in H3 if the seller has reached this value *[MCQ Question]*

=IF(G3>400;”X”;””)

=IF(OR(G3>400;G3<400);””;”X”)

=SUMIF (H3:H13;”X”; G3:G13)

=IF(G3>400;””;”X”)

3. [Medium] Applications in the Alentejo area or under € 10,000 will be analyzed by 1 panel of judges, while the remaining will be by two panels. Indicate which formula to insert in K4 (same to the remaining lines) that generates a "1 panel" or "2 panels" message, as applicable. *[MCQ Question]*

=IF(AND(D4<10000;C4=”Alentejo”);”1 painel”;”2painéis”)

=IFS ((D4<10000; C4=” Alentejo”);”1 painel”;”2painéis”)

=IF (OR (D4<10000; C4=” Alentejo”);”1 painel”;”2painéis”)

=IF (OR (D4<10000; C4=Alentejo);1 painel;2painéis)

4. [Hard] Indicate which formula should be entered in Q14 (same to the remaining lines) to allow the average readers' opinions to be checked against products with positive opinion (> = 6) and unit price of € 500. *[MCQ Question]*

=AVERAGEIF(J4:J22;$J$4:$J$22;O12;$D$4:$D$22;>500)

=AVERAGEIF($J$4:$J$22;$J$4:$J$22;O12;$D$4:$D$22;”>500”)

=AVERAGE($J$4:$J$22;$J$4:$J$22;O12;$D$4:$D$22;”>500”)

=AVERAGEIFS($J$4:$J$22;$J$4:$J$22;O12;$D$4:$D$22;”>500”)

Appendix 6: Amount of number character and number of words

1. Descreva o que irá obter com a fórmula =INDEX(C3:C13;MATCH(MAX(G3:G13);G3:G13;0))

Summary:

- Words: 10
- Characters (no space): 75

2. A fórmula que permite determinar em N11 a receita total gerada com categoria "RM", é dada por

Summary:

- Words: 18
- Characters (no space): 79