U.PORTO

FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

LEARNING ANALYTICS OF K-MEANS CLUSTERING: A PILOT STUDY

Tânia Sofia Monteiro Duarte

Dissertation

Master in Data Analytics

Supervised by

Pedro José Ramos Moreira de Campos

September 2018

Biographical note

Tânia Sofia Monteiro Duarte was born on the 22nd of November of 1995 in Guimarães, Portugal.

In July of 2016, Tânia concluded his bachelor degree in Economics at the School of Economics and Management of the University of Porto. In September of the same year, motivated by the desired of learning more about data analysis, Tânia started another academic challenge by enrolling in the master's degree in Data Analytics of School of Economics and Management of the University of Porto.

In 2017, Tânia joined an internship program at Sonae MC, where she is currently working as Business Analyst in the Supply Chain department in Porto.

Acknowledgements

I should like to thank all those who contributed to this dissertation.

To which the FEP represented, a thank you to all the great teachers with whom I had the opportunity to learn. More specifically, to Professor Pedro Campos for the support, confidence and guidance throughout the elaboration of this work.

To the friends that I met at FEP, who shared with me all this adventure and never stopped walking with me. To my childhood friends who have always been patient and understanding with me during this year.

Lastly and most importantly, I want to thank to my family, who have always helped me in my frustrations and allowed me to get here. Specially, to my mother, my sister and my stepfather for the opportunities they gave me, for the advices and, above all, for all the support that has never been exhausted.

Abstract

Nowadays, the use of virtual learning systems to gather educational information is part of the routine of almost every student and teacher. For studying or, even for searching information about certain exam date or deadlines of assignments, students can easily found that information in the virtual learning system of their schools, like Moodle for example. Professors can put any useful type of information in that systems for support students, or, professors can take any doubt of students in existents discussion forums.

The search for supporting students in a better and easy way has led to the emergence of two communities, the Educational Data Mining community and the Learning Analytics and Knowledge community.

The goal of this study is to analyze, from an educational perspective, the impact that an improved environment for education has on the students' performance and how it is possible to better learn some data mining algorithms.

This research will be based on the analysis and later, classification of the data set obtained through an opinion questionnaire conducted to students of different courses of the University of Porto, on an innovative teaching approach of the k-means clustering model built in the NetLogo software. The analysis around the set of educational data obtained allows us to investigate the impact that an improvement can bring to the field of Learning Analytics teaching.

Keywords: educational data mining; learning analytics; online/virtual learning systems: moodle; educational techniques.

Resumo

Nos dias de hoje, o uso de sistemas virtuais de aprendizagem para agrupar informações educacionais faz parte da rotina de quase todos os alunos e professores. Para estudar ou, mesmo para pesquisar informações sobre determinada data de um exame ou sobre prazos para entregas de trabalhos, os alunos podem facilmente encontrar essa informação no sistema de aprendizagem virtual de suas escolas, como o Moodle, por exemplo. Os professores podem colocar qualquer tipo de informação útil nesses sistemas para suporte aos estudantes, ou podem tirar qualquer dúvida aos alunos em fóruns de discussão existentes aí.

A procura de melhores e mais fáceis formas de apoio aos alunos levou ao surgimento de duas comunidades, a comunidade *Educational Data Mining* e a comunidade *Learning Analytics and Knowledge*.

O objetivo deste estudo é analisar, de uma perspetiva educacional, o impacto que um ambiente melhorado para a educação teria na performance dos alunos e como é possível ensinar com abordagens inovadoras algoritmos de Data Mining.

Esta investigação será baseada na análise e posteriormente, classificação do conjunto de dados obtido através de um questionário de opinião realizado a alunos de diferentes cursos da Universidade do Porto, sobre uma abordagem inovadora de ensino do modelo *k-means clustering* construído no *software* NetLogo. A análise em torno do conjunto de dados educacionais obtido permite-nos investigar qual o impacto que uma melhoria poderia trazer no campo do ensino de *Learning Analytics*.

Palavras-chave: *educational data mining; learning analytics*; sistemas de aprendizagem virtuais/online: moodle; técnicas educacionais.

Table of Contents

List of Tables

List of Figures

1. Introduction

The increased attention to analytics today is mainly due to the increased availability and access to computation. The advances in computation are also motivated by this growth attention to data analytics.

The combination of different types of data has a lot of forms and it can reveal useful insights. New Media Consortium (2015) identified Learning Analytics as part of "midterm horizon" for higher education. The deployment of data mining techniques in education can be transformative, altering not only the old methods of teaching and learning but also the ones of administration and academic work. There are a lot of contributions that an improvement of data analytics techniques can take to the education world.

In the education world, the need for new learning and teaching methods have been developed through, mainly, new extensive educational media and advances in computation. Through this emergence, two distinct research communities, Educational Data Mining and Learning Analytics and Knowledge, have developed in response (Siemens, G. & Baker, R., 2012). Educational Data Mining is about developing and applying methods which detect patterns in large educational datasets, being, otherwise almost impossible to analyze mostly due to the enormous volume of the datasets (Scheuer& McLaren, 2012). Learning analytics has a relatively greater focus on the human interpretation of data and visualization (Baker & Inventado, 2014).

Given that this is a recent topic and therefore, a subject of immense research, this dissertation also comes with this purpose: the purpose of understanding the impact of new learning methods on students' performance and their ease of learning.

In summary, and given the scarcity of research in this field, this study follows a methodological approach to answer the following research questions:

Is it possible to implement improvements in the practice of teaching clustering techniques (more specifically, k-means clustering), based on more visually presentations?

Which is the impact of a new learning method for students' performance and which are the students' opinion about that? Is it easier to learn? Is it more motivating to learn this way?

1.1. Motivation

This study is mainly due to the big opportunities that an improvement of teaching data mining techniques can take to the education world in any possible field. Data mining can be applied to any course, any field and any subject, and that may be the trigger that education needs to be more advanced.

This search for improvements in educational techniques needs an open and transparent research environment. As related, but dissimilar, educational data mining and learning analytics and knowledge can offer a robust force for quality in research in this area.

The motivation for the development of this dissertation arose from the desire to search for more knowledge in learning analytics field, which is a recent topic and it can bring a lot of benefits in the education field. The passion for learning analytics theme comes from a previous assignment of the author aiming at improving a k-means Clustering model in NetLogo. This improved model will be developed during this dissertation.

This research work will be based on the analysis of a dataset from a questionnaire that was applied during the thesis evolution, which the main objective was to do an intervention with the NetLogo improved model and test students' performance. The goal of this dissertation is, through this experimentation, to conclude if there are better ways of teaching and learning clustering techniques and what improvements can be done to achieve that goal.

2. Learning analytics, educational data mining and the effectiveness of teaching: a literature review

This chapter provides an overview of the state-of-the-art of educational data mining and learning analytics' main concepts. Therefore, it starts firstly with a summary of the educational data mining and learning analytics approaches and its procedures.

2.1. Educational Data Mining

The International Educational Data Mining Society defines educational data mining as follows: "Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in." Educational data mining, as we can see in Figure 1, can be seen as a mixture of three key concepts: computational sciences, education and statistics (Romero and Ventura, 2013).

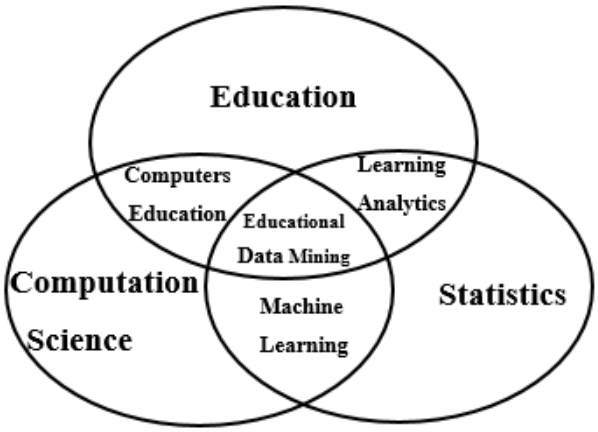

Figure 1- Educational data mining: Combination of three areas ([adapted from Romero and Ventura,2013])

This study proves how the analysis of educational data – educational data mining – can tell us a lot about the performances of the students and predict, or even prevent, students' grades, for example. The goal of educational data mining is, essentially, to improve education by creating models capable of identifying the main problems associated with poor learning. After, the results are used as strategies to do an improvement in teaching techniques, for better learning as a future goal. Educational data mining and learning

analytics share the goal of improving the excellence of analysis of educational data, to sustenance not only basic research but also practice in education.

In this chapter, some key concepts there will be highlighted, as well as some studies carried out in the scope of educational data mining.

2.1.1. Main Researches

For the perceptibility of this work, there are some key concepts about educational data mining that should be accentuated. As previously said, educational data mining pursues innovative ideas in data and search for new procedures and new reproductions.

There are some researches on this topic. The first educational data mining review was presented by Romero and Ventura, in 2007, in the Expert Systems with Applications Journal, with the title: Educational Data Mining: a survey from 1995 to 2005 (Romero and Ventura, 2007). Following this, in 2013, it was lanced the more complete review of educational data mining in the IEEE Transactions on Systems, Man and Cybernetics Journal with the title: Educational Data Mining: a review of the state of the art, by the same authors (Romero and Ventura, 2013). After that, a lot of prestigious international magazines publish works about educational data mining as the Journal of Educational and Behavioral Statistics, the International Journal of Data Mining & Knowledge Management Process and also some podcasts as the Leading Learning podcast, among many other resources.

According to Romero and Ventura (2013), the application' process of data mining for learning systems can be interpreted in different perspectives. Firstly, in an educational point of view, it can be seen as an iterative cycle of hypothesis and tests formation, as we can see in Figure 2. In this process, the goal is not just to transform data in knowledge, but also to filter the extracted knowledge to decision-making about how to change the educational environment and how to improve the learning process.

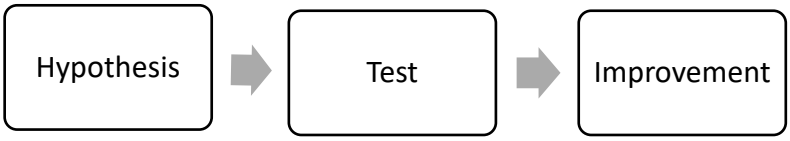

Figure 2 - Process of educational data mining (adapted from Romero and Ventura, 2013)

2.1.2. Educational Data Mining Methods

There are several educational data mining methods that come, directly, from data analysis of students' contact with virtual learning systems. Each method depends on the goal of the educational data mining. Some of these methods are universally known, but others are specially used in educational data mining field. In Table 1, the educational data mining methods described by Baker (2010) are presented. The last five methods presented were added by Romero and Ventura (2013).

Educational data mining method	**Goal**	**Educational data mining application**
Prediction: • Classification • Regression • Density estimation	Technique used for forecasting the future	Predict students' behavior and performance
Relationship Mining: • Association rule mining • Sequential pattern mining • Correlation mining • Causal data mining	This method identifies relationships between variables and then creates rules	To identify students' behavior standards and then to diagnostic which are their difficulties and more frequent mistakes
Clustering	Technique used for creating groups based on the similarities of the dataset's observations	To promote collaborative learning and to group the students to give them differentiated tasks according to their capabilities
Discovery with models	This method uses an existent model obtained from prediction, for example, and then this model is used as an element in another prediction technique.	To support relationships identification of students and their characteristics.

Distillation of data for human judgment	This method aims to represent the data in a more visual and legible way to improve the human comprehension and then to support important decisions based on data.	To help students and teachers to analyze students' course activities and the use of information.
Text Mining	Method that obtains useful and rich information from a text dataset.	To analyze the discussion forums, chats and documents context.
SNA – Social Networks Analysis	Method that measures relationships between entities in networks. SNA seeks social networks through connections.	SNA can be used for interpreting the structure of relationships in cooperative tasks and interactions with communication tools.
Outlier Detection	This method aims to discover data points very different from the others	To detect students with learning difficulties and irregularities in learning processes.
Knowledge Tracing	This method is used to estimate the student capacity in some knowledge areas	To follow students' performance over time
Process Mining	Method that aims to extract related knowledge with the process through events registration in information system	It can be used to reflect the behavior of students with respect to their evolution and performance over the course of their academic path.

Table 1 - Educational data mining methods according to Romero and Ventura, 2013

Frequently, educational data mining techniques come from data mining field (Baker, 2012). In the most of times, it is necessary to adapt these techniques due to the existents particularities of educational environments and its data. According to the same author, the main educational data mining applications are: domain and student modelling, scientific investigation and pedagogic support.

2.1.3. Metrics

Researchers have been using, frequently, performance metrics of the models' in order to do a benchmark. This happens because there is not, nowadays, a standard metric for these models.

According to Pelánek (2015), metrics for evaluating student models depends of the type of the model. The most often used type of student models are models of student skills (Desmarais and Baker, 2012).

Next, an overview of metrics for evaluation student models, will be presented, adapted from Pelánek (2015).

Metrics Classification	**Formula**	**Definition**
Probabilistic understanding of errors	Mean Absolute Error	$\frac{1}{n}\sum_{i=1}^{n}\lvert o_i - p_i\rvert$
	Root Mean Square Error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2}$
	Log-likelihood	$\sum_{i=1}^{n} o_i \log(p_i) + (1 - o_i)\log(1 - p_i)$
Qualitative understanding of errors	Accuracy	$(TP + TN)/n$
	Precision	$TP/(TP + TN)$
	Recall	$TP/(TP + FN)$
Assessing ranking of examples	Receiver operating curve (ROC)	ROC curve reviews, through all thresholds, the qualitative error of the prediction model
	Area under curve (AUC)	AUC is the area under ROC curve and it gives, through all thresholds, the probability of a random positive observation has a higher predicted score than a random negative observation.

Table 2 - Metrics ([adapted from Pelánek, 2015])

2.2. Learning Analytics

According to the definition of learning analytics set in the 1st international conference on Learning Analytics and Knowledge (LAK 2011), "Learning analytics is the measure, analysis and communication of data about students and their contexts for understanding and optimizing the learning process and its environments".

In 2011, The Society for Learning Analytics Research (SoLAR) was formed to develop and advance a research agenda in learning analytics and to input the use of analytics in learning. While learning analytics and educational data mining share many attributes, and have similar goals and interests, they have distinct technological, ideological, and methodological orientations (Siemens and Baker, 2010).

"Learning analytics need not to simply focus on student's performance. It might be used as well to assess curricula, programs, and institutions. It could contribute to existing assessment efforts on a campus, helping provide a deeper analysis, or it might be used to transform pedagogy in a more radical manner." (Johnson et al, 2011, p. 28).

Learning analytics is fundamental to change the installed fog around higher education. For students, it is important to receive information about their performance when compared to their colleagues for having more motivation. For professors, it is significant to have available information about students' performance, more precisely about at-risk students, for planning correctly the teaching activities. For these reasons, learning analytics can take doubts not only in how to distribute resources and to develop competitive advantages, but also can improve life's quality and students' experience of learning (Siemens and Long, 2011).

2.2.1. Learning Analytics Process

There are some related works about how academic analytics proceed. Learning analytics come from academic analytics; it is a mix of academic analytics processes. These frameworks are presented in previous Table 3.

Academic Analytics Frameworks	Author	How it works?
Knowledge continuum	Baker, 2007	Predictive analytics corresponded with the renovation of knowledge to wisdom

Web analytics objectives	Rogers, McEwen and Pond, 2008; Hendricks, Plantz and Pritchard, 2008;	Stakeholders use essential metrics to identify what types of outcomes they want from users (p.233)
The five steps of analytics	Campbell and Oblinger, 2008	Authors consider five steps: capture, report, predict, act and refine. Similar to knowledge continuum, but with the addition of refine that recognizes analytics as "self-improvement project".
Collective application model	Dron and Anderson, 2009	Model with five layers (select, capture, aggregate, process and display) divided into three cyclical phases (information gathering, information processing and information presentation).

Table 3 - Learning Analytics Processes [adapted from Elias, 2011]

Learning analytics process results brings important results to teachers, mainly to reflect on their teaching methods and how it impacts students' behavior or even to understand if their teaching goals were achieved.

On the contrary of educational data mining, learning analytics applies known models to response questions related to learning environments and it does not create models. One of the most important applications of learning analytics is to predict students' performance to help, mainly, students identified as students at risk of failure.

Educational data mining and learning analytics have many similarities, but there are some differences. According to Baker (2010), it is essential to use human judgment in learning analytics using automatized tools. In educational data mining, it is important to discover new ways of automation and the human judgment is just a secondary tool.

2.3. The human domain in the teaching process

As previously mentioned, the main goal of this dissertation is to evaluate the results of a pilot study conducted between students and understand the impact of a new approach of teaching an algorithm differently from the normal way. For that reason, some concepts within human domain as the self-perception of students on some subject and the effectiveness of a new teaching method must be developed to the results obtained in the real world have a perceptible review.

2.3.1. Level of expertise

Intelligence is an important variable in the pursuit to comprehend academic achievement (Valentini & Laros, 2014). With the intelligence comes the students' auto perception of their intelligence and how they see themselves in the world.

According to Vaz Serra (1988), self-perception is related to the self-concept that the human being has of himself. The author defines self-concept as being "Self-concept can be defined in a simple way, such as the individual's perception of himself and the concept that, due to this, form of self, and assume that for the construction of this definition there are four types of influences in human self-perception: (i) the way other people observe an individual, (ii) the individual's notion of his or her performance in specific situations, i.e., he or she can judge whether or not is competent or incompetent, (iii) confrontation of the conduct of the person with that of the social peers with whom it is identified, (iv) the evaluation of a specific behavior according to values conveyed by normative groups.

Historically, education goals have tended to fluctuate from emphasis solely on cognitive outcomes to major concern with social and affective ones (Shavelson et al, 1976). The number of studies on self-concept reflects the emphasis on noncognitive outcomes of education. For that reason, the improvement of a student's self-concept seems to be valued as an educational outcome in its own right.

According to Shavelson (1976), self-concept is inferred from a person's responses to situations. The situations and the responses may be physical or symbolic. In most educational examinations of self-concept, a distinction is made between self-concept and inferred self-concept. Self-concept is restricted to a person's report of. Inferred self-concept is another's attribution of a person's self-concept. The same author considers three facets of self-concept: (i) physical, (ii) social, (iii) academic, thus, academic may be more related to achievement than the others. Another study of this author considers self-concept

as hierarchical on a dimension of generality. The different facets of self-concept should form a hierarchy *bottom-up* – starting with individual experiences in particular situations on the bottom and ending with a general self-concept at the top, as presented in Figure 3.

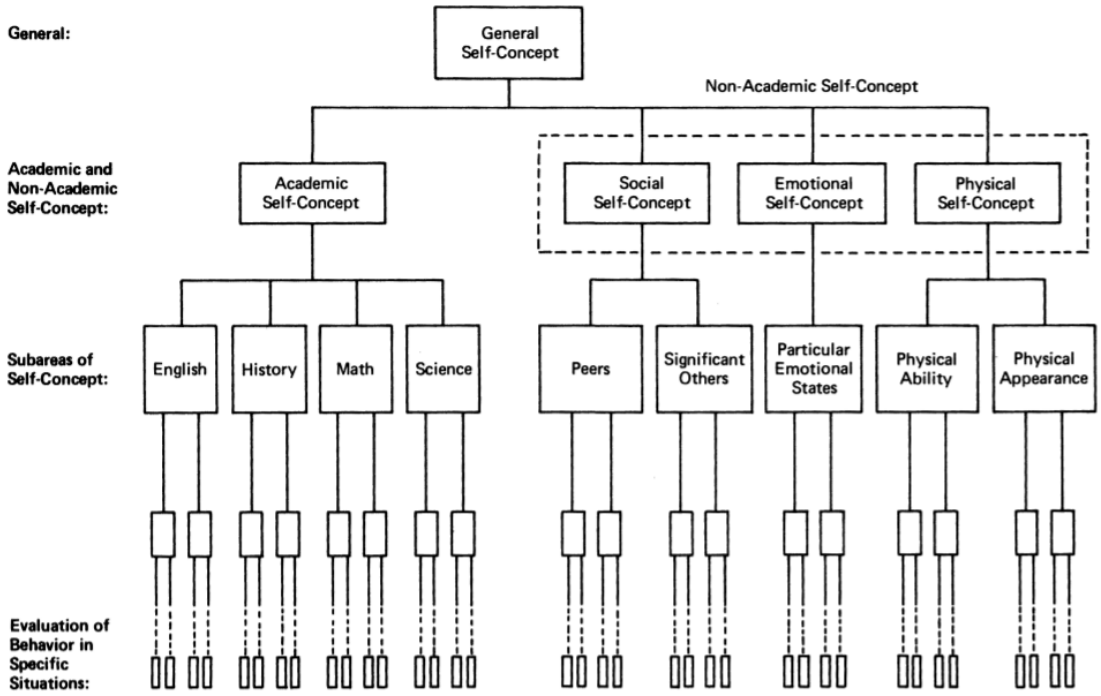

Figure 3 - Self-concept hierarchy ([from Shavelson et al,1976])

Self-concept surges on top and it is divided into academic and non-academic self-concept. Academic self-concept is divided into various academic areas and, then, specific subjects. Non-academic self-concept is divided into social, physical and emotional self-concept. The specification of the division level of this hierarchical structure is increasing until the bottom of the hierarchy. Focusing on the side of academic self-concept of the hierarchy, Shavelson (1976) points for two hypotheses: a) self-concept of mental ability should be more closely related to academic achievement than to ability in social and physical situations; b) self-concept of academic ability in science should be more closely related to achievement in science than to achievement in overall grade-point-average. After testing some tests and correlations on its methodologies, Shavelson (1976) suggests that academic self-concept should cluster together and it should be distinct from a cluster of items on emotional self-concept, for example.

Another important part of self-concept is the self-esteem (Bausmeister, 2014) because is a part of the formation of the self-concept and it derives from evaluation

processes that a student does of himself (Vaz Serra, 1988). Thus, self-esteem and the learning process are directly linked since that the learning difficulties can cause a low self-esteem on the student and the low self-esteem can cause a low learning process of the student (Cavalcanti, 2003). The process of teaching-learning may take into account the self-esteem of the student because this will be an important part of the self-concept and, then, self-perception of the student. A positive self-perception of the student will lead him to a higher performance on any subject.

An important students' feature for their self-concept is their age. According to a study conducted by Bloom (2006), older persons incorporate certain of the positive stereotypes of ageing into their self-concept. The findings of the author's research work indicate that similarities in self-perception with age outweigh differences. It was found a curvilinear relationship between self-acceptance and chronological age supporting theory of life stages and time perspective.

For the pilot study of this dissertation, it will be analyzed the self-perception of the student on the clustering theme and it is important to note that students' self-perception influences the teaching-learning process.

2.3.2. The effectiveness of teaching process

According to Shulman (1987), a teacher can transform understanding, performance skills or desired attitudes or values unto pedagogical representations and actions, and he can do this talking, showing, enacting or, otherwise representing ideas so that the unknowing can come to know, those without understanding can comprehend, and the unskilled can become adept. Thus, the teaching process necessarily begins with a teacher's understanding of some subject. Today, more than ever, it is not so important for the student to have lessons: it is important, rather, that he has access to effective means and tools of learning, whatever they may be. It is not important to have a teacher who exposes the subject, but before there are occasions of interaction that facilitate the understanding and intelligent integration of the contents (Trindade, 2005). An efficient teaching process is synonymous with a good teaching process.

E-learning courses help students achieve the desired results on learning processes. The effectiveness of e-learning courses is being measured by assessing the impact of a certain training course on the learners (CommLab, 2010).

To measure the impact of the effectiveness of the teaching process, the most influential measure of performance of teachers is students' ratings. It can be noted that students' ratings are the single most valid source of data on teaching effectiveness and it seems to be agreement among the experts on faculty experiments that students' ratings provide an excellent source of evidence for formative and summative decisions (Berk, 2005).

In this section, a literature review containing the key concepts around educational data mining, learning analytics and the effectiveness of teaching-learning process was presented. In the next section, the methodological approach of this dissertation is presented through a description of some key concepts about k-means clustering and the improved model developed for this dissertation.

3. K-means clustering algorithm: a literature review

In 1967, James MacQueen uses for the first time the concept of k-means. The author explores a process in which the main goal is the partition of a simple sample into k sets, for having an efficient within-class variance (MacQueen, 1967). Further, he proves k-means method as the best application to problems with clustering or similarity groups, allowing to any investigator in obtaining a qualitative understanding of large amounts of data by providing him with reasonably good similarity groups.

This study proves how the k-means process can help us to analyze almost all of the datasets of today's, which are defined as having a greater dimension.

In this chapter, there are presented some key concepts in the scope of the k-means clustering method.

3.1. Key concepts of the algorithm

For the content of this research to be perceptible, there are some concepts about clustering techniques and the k-means clustering process that should be highlighted.

3.1.1. Clustering algorithms overview

Nowadays, a large amount of data is founded in every kind of areas. These data can be used to understand a new phenomenon through the research of features that can describe it and, further compare it with other known objects based on the similarity or dissimilarity, according to some standards. Thus, the main mean of dealing with these data is to classify or group them into a set of clusters (Xu & Wunschll, 2005). These authors distinguish classification systems as supervised or unsupervised, depending on the kind of input generated – if it has discrete supervised categories, which come from labeled data, or unsupervised categories, which come from unlabeled data, respectively. According to Xu and Wunschll (2005), in clustering, which is an unsupervised classification, no labelled data are available and the goal of it is to separate this kind of data into a discrete set of structured data with categories or clusters. The main goal of cluster analysis is to split a group – cluster - of objects into homogeneous subgroups through *a priori* chosen a measure of similarity, such that the similarity within each object of each subgroup is higher when compared to the similarity of each object of other subgroups (Backer & Jain, 1981).

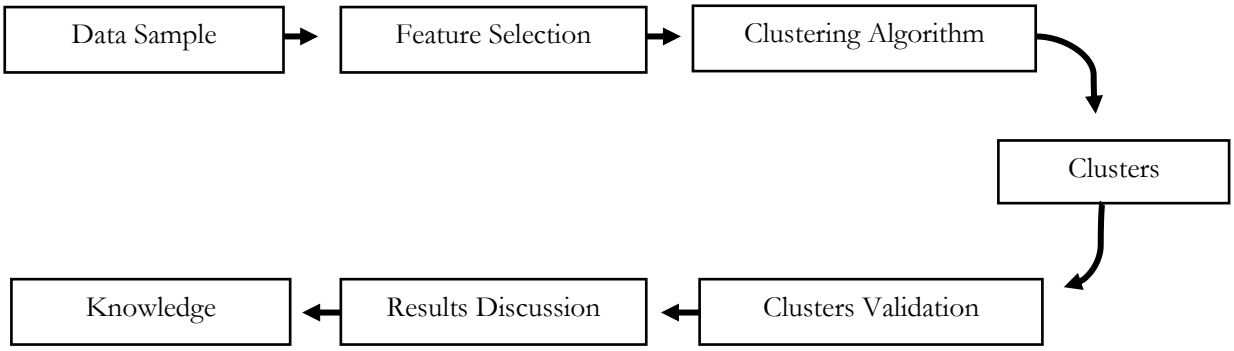

Figure 4 - Typical cluster analysis process ([adapetd from Xu & Wunschll, 2005)]

As it can be seen in Figure 4, cluster analysis process is simple, but due to its nature in a lot of circumstances, it is a process that needs to be repeated more than once. In the process, the clustering algorithm to be used need to be chosen. It can be seen in Table 4 the different clustering algorithms.

Clustering Algorithms	**Some Examples**
Distance and similarity measures	Euclidean distance; City-block distance; Pearson correlation; Mahalanobis distance
Squared error-based	K-means; genetic K-means; Partitioning around medoids
Graph theory-based	Delaunay triangulation graph; Chameleon; CAST (cluster affinity search technique)
Estimation via mixture densities	GMDD (Gaussian mixture density composition)
Hierarchical (Agglomerative and Divisive)	Single linkage; Complete linkage; Centroid linkage; divisive analysis
Fuzzy	FCM (Fuzzy c-means); FCS (Fuzzy c-shells)
Neural-networks based	LVQ (Learning vector quantization)
Sequential data	Sequence similarity; Statistical sequence clustering;
Combinatorial search techniques-based	TS clustering; GGA (Genetically guided algorithm)
Kernel-based	Kernel K-means; SVC (support vector clustering)
High-dimensional data	PCA (Principal component analysis); Isomap
Large-scale datasets	WaveCluster; ART

Table 4 - Different clustering algorithms and some examples ([adapted from Xu & Wunschll, 2005)]

The clustering algorithm targeting this research is the squared error-based type, more specifically the k-means clustering. It is a very simple algorithm and it can be implemented in solving problems, and for that reason, it can be taught in a new way in order to lead the students in learning better and easier.

3.1.2. The goal of k-means clustering

As previously mentioned, the k-means is one of the methods of clustering techniques and is a type of unsupervised learning, which is used when it exists unlabeled data. Clustering techniques consist of iterating a dataset automatically by its within similarity degree (Steinbach & Kumar, 2000). The similarity degree depends on the problem definition and on the algorithm used. The most popular clustering algorithms are the partitional and the hierarchical ones. The simplest form of clustering is partitional clustering, which aims at partitioning a given dataset into disjoint subsets (clusters) so that specific clustering criterion are optimized (Likas & Verbeek, 2004). The k-means is a partitional algorithm and it minimizes the clustering error.

The k-means procedure subdivides data points of a certain set into clusters based on nearest means values and for determining the optimal division of these data points into clusters, the distance between points must be minimized. This algorithm's goal is to minimize an objective function, in this case a squared error function. The objective function is defined as

$$M(P, C) = \sum_{k=1}^{K} \sum_{i \in P_k} ||x_i - c_k||^2 \qquad \textbf{(1)}$$

Where P is a k-cluster partition of the object set represented by vectors x_i ($i \in I$) in the N-dimensional feature space, consisting of non-empty, non-overlapping clusters M_k, each with a centroid c_k (k=1,2,…K) (Kodinariya & Makwana, 2013).

To assign each data point to one of k clusters based on feature similarity of the dataset, the k-means algorithm works iteratively and the final results are the centroids of the k clusters (which can be used to label new data). There is an example of this in Figure 5. Investigating the centroid feature weights can be used to qualitatively understand what kind of group each cluster represents (Hamerly & Elkan, 2004).

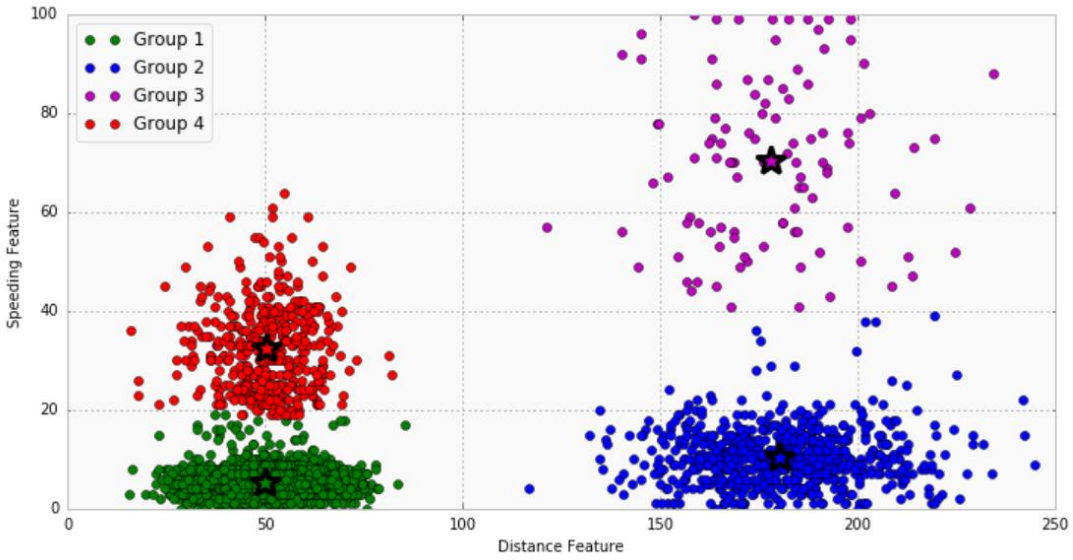

Figure 5 – Example of final centroids formed from k-means iteration ([adapted from Hamerly & Elkan, 2004])

To achieve the final result, the k-means clustering algorithm uses an iterative refinement (Bradley & Fayyad, 1998). As mentioned previously, the algorithm inputs are only the number of clusters (k) and the dataset, which is a collection of features of each data point. First, the algorithm starts with an estimation for k centroids, which can be randomly generated and then iterates in two steps:

1. Data assignment: Based on the squared Euclidean distance – "the distance is computed by finding the square of the distance between each score, summing the squares and finding the square root of the sum" (Oyelade & Obagbuwa, 2010) – each data point is assigned to its nearest centroid, and each centroid describes one of the clusters. Formally, if c_i is the group of centroids in the set C, each data point x is assigned to a cluster based on:

$$argmin_{c_i \in C}\ dist(c_i - x)^2 \qquad (2)$$

   Where $dist(c_i - x)^2$ represents the Euclidean distance, and being the set of data points assignments of a i cluster a D_i.

2. Centroids update: Reassignment of the centroids formed in the previous step. This reassignment is achieved by recalculating the average of all data points assigned to the cluster of this centroid, based on:

$$c_i = \frac{1}{|D_i|}\sum_{x_i \in D_i} x_i \qquad (3)$$

As mentioned in the previous section in Table 4, there are more distance measures that can be used in clustering algorithms, the distance is, frequently, chosen according to the type of data. K-means works assigning data points to the closest centroid using Euclidean distance from data points to the centroid. This algorithm is implicitly based on Euclidean distance, since the sum of the centroid squared deviations is equal to the sum of the paired Euclidean quadratic distances divided by the number of data points. The term "centroid" itself comes from Euclidean geometry (Oyelade & Obagbuwa, 2010).

The two steps work iteratively until a criterion is reached, which means that the sum of the distance is minimized and the objective function of k-means (1) is achieved. The result of algorithm iteration is guaranteed and, normally, is a local optimum – an optimal solution within a neighboring set of candidate solutions (Selim & Ismail, 1984).

The k-means algorithm described discovers the clusters for *a priori* chosen k. There is not a specific method to determine the value of k, but there are different techniques that can be used. For this research work, one of the most common methods was used to find the value of k, as it can be seen in the sections 3.1.3 and 3.1.4.

3.1.3. The optimal number of clusters

An important note at the beginning of k-means process is the needed to give the input of quantity of k, which represents the numbers of clusters in the data. This input will be imperative in the quality of the clusters, mostly when datasets have more than three variables.

There are some methods that validate the numbers of clusters. For this research, it was applied the Elbow method, which is a visual method (Kodinariya & Makwana, 2013). The Elbow method exists upon the idea that one should choose a number of clusters so that adding another cluster does not give much better modelling of the data (Bholowalia & Kumar, 2014).

3.1.4. The convergence criterion of Elbow method

The percentage of variance explained by the clusters is plotted against the number of clusters. The first clusters will add much information but at some point, the marginal gain will drop dramatically and gives an angle in the graph. The correct "k", i.e., the number of clusters is chosen at this point, hence the "Elbow criterion" (Kumar, 2014).

It starts with 2 clusters (k=2) and keeps increasing it in each iteration by 1, calculating the clusters and the training cost. (Kodinariya & Makwana, 2013). At some point, the cost

of finding the number of clusters (k) will drop aggressively and, at this point, it reaches the desired value of k. This means that, after this point, the increase in the number of clusters takes the new cluster to a very near of an existing one. Then, this point is called the stabilization point because is it the point where the convergence criterion is reached. It is possible to observe an example in Figure 6.

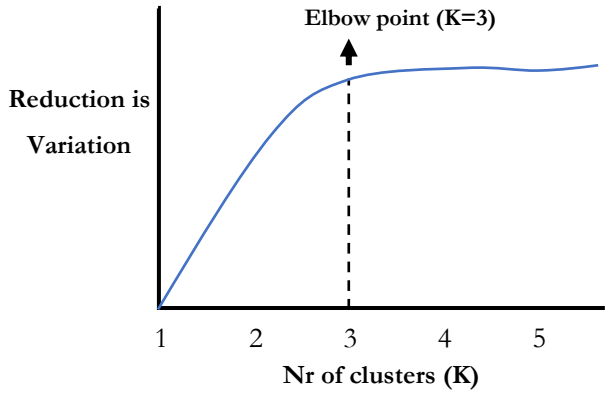

Figure 6 - Identification of elbow ([adapted from Kodinariya and Makwana, 2013])

The increase on the number of clusters will reduce the distance to data points, then increasing k will decrease the elbow metric until the extreme of reaching zero when the value of k is equal to the number of data points. For this reason, the function of k is plotted as the mean distance to the centroid and the elbow point (stabilization point) is used to determine the value of k.

4. A k-means clustering model in NetLogo

4.1. Introduction and Purpose

In this section, the original and a new improved k-means clustering model in NetLogo (Wilensky, 1999) software are presented. We also briefly present Netlogo, and the paradigm of Agent-Based Modeling. As this model fits on multi-agent programmable modelling environment, it was decided to use the Overview, Design Concepts, Details (ODD) protocol as the methodology of this work. According to Grimm (2010), ODD improves the rigorous formulation of models and helps make the theoretical foundations of large models more visible.

The original model developed in Netlogo was designed to find the k-clusters, with the k defined by the user, in a certain subset of unlabeled data (Hjorth & Wilensky, 2014). The model aims to minimize the dissimilarity within groups of the unlabeled data, making it easier to the user to use it. This model has an innovation character because it is very visual and interactive, so it brings a different way of teaching this clustering technique. Then, we show the improved model (developed by the author of this dissertation) that brings the advantage to the user of not having to choose the value of k to find the clusters between a certain number of data points defined, because of the introduction of the Elbow method.

4.2. NetLogo

The original and improved models have been developed in NetLogo software, as previously mentioned. Created by Wilensky in 1999, NetLogo is a multi-agent programming language and modelling environment for simulating complex phenomena and it was design for both research and education, being used across a wide range of education levels (Tisue & Wilensky, 2004). The decision about using this model was, mainly, due to its rapid prototyping and initial testing of multi-agent systems, particularly suited to systems with agents situated and operating in a restricted space, as well as an excellent animation tool of the modelled system and it also proved to be an excellent educational platform for teaching artificial intelligence (Sakellariou & Stamatopoulou, 2008).

.

4.3. Agent-based modelling

In agent-based modeling (ABM), a system is modeled as a collection of autonomous decision-making entities called agents and it can be interpreted as a mindset, more than a technology (Bonabeau, 2002). An ABM is described as a set of models for simulating interactions between agents for measuring their effects on the modeled system.

The main feature of ABM are the competitive interactions between agents that are repetitive, representing the power of it to explore the dynamic of the system. For that, each agent may execute different behavior in each interaction in order to assess the appropriate one in the system it represents. In case of this dissertation, the agents are the data points and they will interact until they find the appropriate, or optimum cluster that they do part of.

According to Bonabeau (2002), ABM has benefits over other modelling techniques and they can be apprehended in three statements:

i. ABM is very flexible.
ii. ABM offers a natural description of a system.
iii. The most important one – ABM has the ability to capture emergent phenomena.

An agent-based model is defined as a set of differential equations, each describing the dynamics of one of the system's constituent units and it enables to deals with complex individual behavior, including learning and adaptation (Bonabeau, 2002).

According to Epstein (1996), "ABM may change the way we think about explanation in social sciences. What constitutes an explanation of an observed social phenomenon? Perhaps people will interpret the question, 'Can you explain it?' as asking 'Can you grow it?'". ABM community is concerned to promote a new way of approaching phenomena from a perspective of redefining a scientific process entirely and not from a traditional modelling perspective.

4.3.2. Entities, state variable and scales

The improved model includes two main entities: data points and clusters. The data points are set by the user of the simulation and the number of optimal clusters (k) is defined by the method introduced.

4.3.2. Process Overview and Schedule

The simulation of the improved model will run in two separate steps:

1. When the user tries to find the optimal number.
2. When the user knows which is the optimal number, according to the Elbow method.

In step one, the user defines manually the number of data points that he wants to cluster. In this step, in the improved model the position of data points was set to be assigned randomly, to avoid possible bias, on the contrary of the original model. After defining the number of data points, the user runs the model. For ending the step one, in the improved model it was created a plot with the Elbow method, which finds the optimal number of centroids taking into account the number of data points chosen in the beginning. In this way, in step two the user knows the optimal number of centroids through this plot, so, he can run the model again and get the best possible clusters for the generated data points.

4.3.3. Design Concepts

i. Basic principles

The basic principle addressed by the improved model is the random assignment of the data points and the introduction of the Elbow method to find, automatically, the optimal number of clusters.

ii. Emergence

In the presented model, the number of data points defined by the user in the beginning of the simulation lead to the optimal number of clusters founded by the Elbow method, which minimizes the squared error.

iii. Adaptation

The model has a completely adaptive behavior. It just depends on the user decision of how many data points he wants to generate for forming the optimal number of clusters.

iv. Objectives

The main objective of the model is to minimize the distance between data points. The Elbow method was introduced through this goal, it increases k (number of centroids) until the extreme of the number of data points to found the elbow point – stabilization point - where the convergence criterion of this method is achieved and the optimal number of clusters too.

v. Learning

As at each simulation, the model does a reset on their ticks – improvement done in the environment of the improved model, for the model do not memorize the past number of centroids – the data points do not have learning abilities to find the optimal number of clusters, but the Elbow method has.

vi. Prediction

Elbow method can predict the optimal number of clusters if the user indicates the number of data points to be generated.

vii. Sensing

Sensing is important in the improved model - the Elbow method is assumed able to identify which is the optimal number of clusters, regarding a chosen number of data points, minimizing the distance between them.

viii. Interaction

The improved model assumes interaction between data points with the objective of finding the nearest mean centroid of them. For helping data points to do this more efficiently, the Elbow method was introduced.

ix. Stochasticity

Stochasticity was implemented in the improved model, as previously mentioned, in order to avoid a possible bias when the data points are assigned to be generated.

x. Collectives

There is no consideration for collectives of data points in the improved model or in the original model, because each unlabeled dataset only permits a generation of a number of data points.

xi. Observation

To allow the observation of the optimal number of clusters, the model runs the data points until the elbow point is founded, for starting again the setup of the model with the correct number of clusters.

4.3.4. Initialization

At the beginning of the simulation, the data points set are randomly assigned. The optimal number of clusters is found through the elbow plot, and finally, the model is set up with the optimal number of clusters to minimize the distance between data points.

4.3.5. Input Data

The model has no input data because the model environment is normally constant. The model just needs the indication of the number of data points to be generated.

4.3.6. Submodels

At each step of the model, data points that are assigned randomly in the environment perform movements in the space until the optimal number of clusters is founded. After this number is achieved by the elbow point, which minimizes the distance between data points, the model runs again and found the best local solution with the best number of clusters to the chosen number of data points.

4.4. The model

The original k-means clustering model of NetLogo was designed to find the clusters, with the users guessing the number of clusters to search for until the users find one that best characterizes their data. The result of the original model is a set of centroids and each of them is located at the average position of a corresponding cluster. This means that, in the original model, the k is the guess of the users at the many centroids to search for. Thus, the original model works in two steps:

1. After the user chooses the number of clusters, data points and centroids to create, the data points assign themselves to their closest centroid taking on its color.
2. Then, all centroids are moved to the average position of the data points assigned to it. It can be seen as an example of an output of the original model in Figure 7.

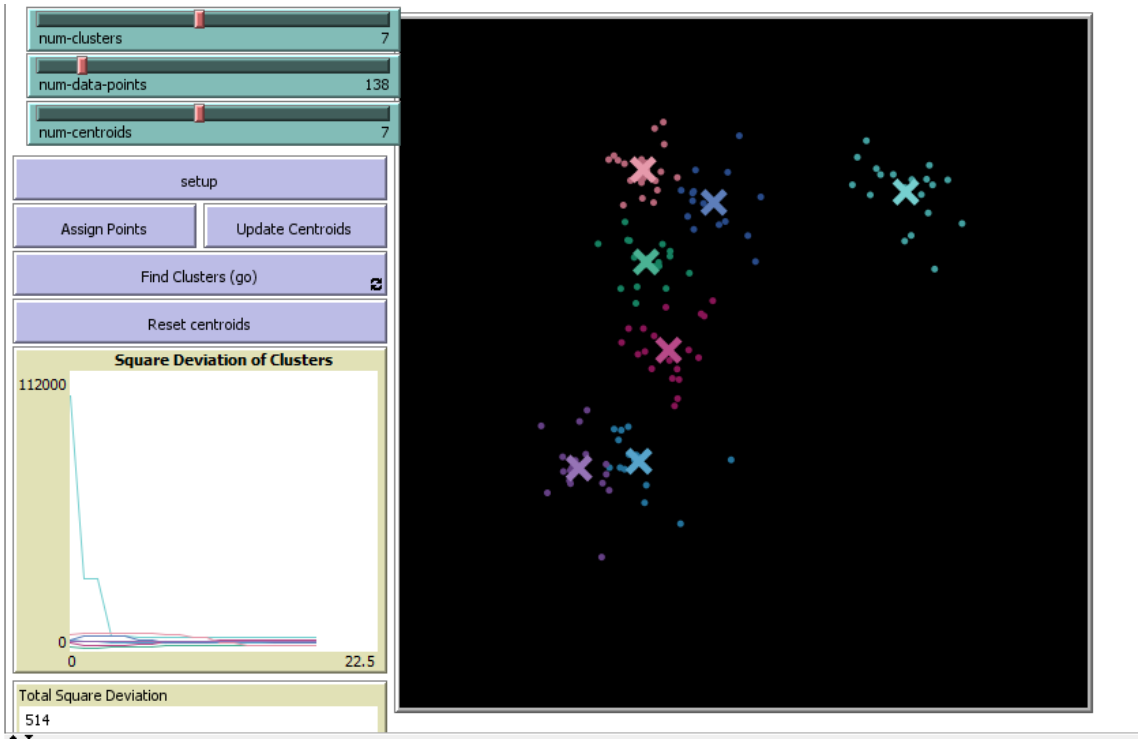

Figure 7 - Interface of original k-means clustering model in NetLogo

The original model performs these steps until the data points assign to the centroids do not change, which means that in this phase the centroids have converged. It is perceptible that this model does not find the best solution and, therefore, it is a very premature model. The improved model appears in this context, in order to make it easier for the user to not have to randomly choose the number of clusters.

As the k-means algorithm intends to minimize an objective function, more precisely a squared error function, and for having a more precise number of clusters, the proposal to improve the model was to insert a method that would automatically find the optimal number of clusters. With this proposal in mind, the Elbow method was chosen because it is a method that should choose a number of clusters so that adding another cluster does not give much better modeling of the data. The interface of the improved model is presented in Figure 8.

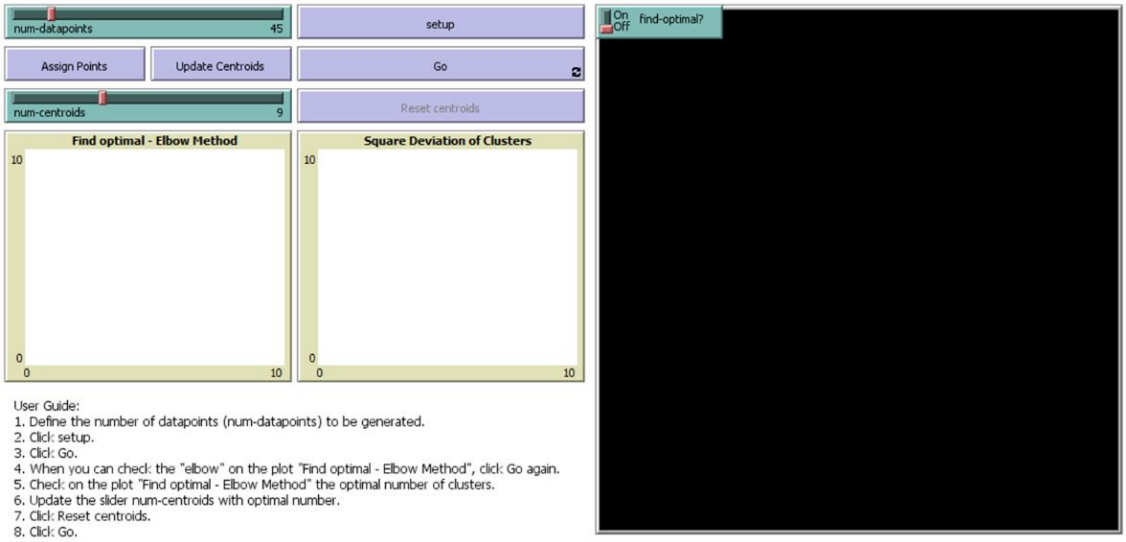

Figure 8 - Interface of the improved k-means clustering model in NetLogo

The improved model run in two separate steps:

1. When we try to find the optimal number of clusters, i.e., when the button "find-optimal?" is on.
2. When we know the optimal number of clusters, i.e., when the button "find-optimal?" is off. We do not need to change the switch manually, it updates automatically when no more centroids are found.

To find out the optimum number of clusters, the improved model has undergone some changes:

- In order to avoid a possible bias, we set the position of data points to be assigned randomly.
- The ticks (in NetLogo models, time passes in discrete steps called "ticks") are related with the number of centroids in the original model, so we added reset-ticks in the second setup, i.e., in the second step when the optimal number of the cluster were found. This command resets the tick counter to zero and generates the data points.
- When checking the optimal number, we set the first number of clusters to be tested as 1.
- Each tick on the simulation will correspond to a number of clusters and will show its distribution. We added the "+1" to ticks because, as we are talking about centroids, it makes sense to start the process with 1 instead of 0 clusters.

- We created a plot with the Elbow method, which finds the optimal number of centroids taking into account the number of data points chosen by the user in the beginning. Thus, after verifying the optimal number of centroids from this plot, we can put that number in the "num-centroids" slider and get the best clusters for the generated data points.
- The second part of the simulation is similar to the original model. Commands go, assign-clusters and update-clusters were not changed, but we change the code of command reset-centroids because we do not want to clear our plots – elbow plot. The only difference on reset-centroids is that we do not want to clear our plots (elbow Plot), and we rewrite the code of the square deviation plot to take into account the elbow method added to the model.

With these changes to the original model, we were able to find a way for the algorithm to automatically calculate the number of clusters of a given number of data points initially defined by the user. As already mentioned, it was decided to use the Elbow method because it is the simplest to use in the context in question.

4.5. Performance metric

As the main goal of this dissertation is to measure the impact that the simulation in NetLogo had on learning the k-means clustering algorithm on the part of the student, we decided to create a metric based on the answer of the two questions that follow the simulation during the questionnaire. That is, after students simulate the k-means clustering algorithm in NetLogo, two questions are asked to realize whether the simulation really helped in their learning or not. These questions are related to the optimal number of clusters and the convergence criterion of the algorithm, more specifically it is question 7 and question 8 of the questionnaire, presented in Appendix 2.

First of all, it should be noted that during the data preparation of the questionnaires a scaling of these two nominal variables was performed for binary variables to make it easier to analyze the dataset. Thus, these two questions were converted to 1 if the answer is correct and 0 if the answer is wrong. Since there are only two questions where we evaluate the performance of the students after the simulation and their answers are inferred through the simulation, that is, it is perceptible during the simulation, it was decided to give a weight

of 50% to each of the answers to the questions to evaluate the student's performance. Thus, the performance metric we built was as follows:

$$\begin{aligned} &\boldsymbol{Students' performance\ after\ simulation} \\ &= (\mathbf{50\%}\ \text{x}\ \textit{Answer of the question related to the optimal number of clusters}) \\ &+(\mathbf{50\%}\ \textit{x Answer of the question related to the converge criterion}) \end{aligned} \quad \textbf{(4)}$$

With this metric, it is possible in the data analysis section (5.4) to evaluate the impact of simulation on students' learning through their performance in the mentioned questions and, to compare these values with the rating to which the students evaluated the simulation, defined in section 5.2. This performance metric was created in order to understand the impact of simulation on student learning.

5. Pilot Study

As the main goal of this dissertation is to recognize that the improved k-means clustering, mentioned on the previous section, brings a better way to teach the algorithm, it was developed a questionnaire addressed to students of the University of Porto in order to evaluate the impact of that method in their learning about k-means clustering. In the investigation of this work, this model will be used in the target group of questionnaires, for trying to understand if this way of learning clustering techniques is better than the ones that are more traditional nowadays.

The technique of data collection used, was the survey by questionnaire, which is a non-documentary data collection technique without direct observation. This technique has advantages such as the possibility of collecting data on a large sample, allows the comparison of participants' responses and it is possible to generalize the results of the sample to the entire population. However, formulating the questions in a questionnaire requires rigor. According to Afonso (2005), the application of a questionnaire survey makes it possible to convert the information obtained from the participants into pre-format, facilitating access to a large number of subjects and contexts differentiated.

In this chapter, we presented the students' performance analysis, through the answers of the questionnaire made, with a universe of objective and attribute variables to determine the improvement in the practice of learning clustering techniques based on some students' features. It will be used learning analytics tools to analyze the dataset obtained from the questionnaire.

The main questions of this investigation are, as said before:

Is it possible to implement improvements in the practice of teaching clustering techniques (more specifically, k-means clustering), based on more visually presentations?

Which are the impact of a new learning method for students' performance and which are the students' opinion about that? Is it easier to learn? Is it more motivating to learn this way?

To answer to these questions, the questionnaire in a sample of students seems to be the perfect investigation to do, not only because it is a study in the field of education and in that students' opinion are very important but also to understand if the improvement in the developed model is useful for a better learning.

According to Richardson (2004), students' performance on questionnaires on approaches to studying show reasonable stability over time. Commonly, for a data collection of a study elaboration case studies and statistics analysis are used. In this dissertation, we used a quantitative method to extract, from the dataset obtained from the questionnaire, useful knowledge of the students learning about the k-means algorithm in NetLogo.

The questionnaire permits the collection of a significant amount of information about a sample from a population of a reasonable size, allowing an examination of the variables in an investigation. The analyze of the questionnaire is an essential part of this dissertation because it was developed to evaluate the students' performance, through learning analytics.

5.1. Study Details

The main goal of the questionnaire was to investigate if the improved k-means model in NetLogo could be a new approach to teaching this algorithm more efficiently. For that reason, the theme of the survey was k-means clustering algorithm in the field of data mining, is that a content constraint of the study. The target population are students of the University of Porto, not only because as a student of this university I have easier access to this population than other universities, but also and more important because this dissertation fits in the field of education and for that reason, the questionnaire had to be answered by students.

5.2. Structure of the questionnaire

As previously mentioned in the section of literature review and according to the improved model in the NetLogo environment, it was decided to divide the questionnaire in two parts:

1. **First part**: the students had to denominate their nationality, age, academic background, master course area, level of expertise in clustering and what is the main goal of the k-means clustering algorithm. Being these six variables defined as:
    - **Nationality**: qualitative nominal variable.
    - **Age**: quantitative continuous variable.
    - **Degree background**: qualitative nominal variable.
    - **Master course area**: qualitative nominal variable.

- **Level of expertise in clustering**: qualitative ordinal variable – it was defined a scale from 0 to 100, is that 0 means the students neither knows the term of clustering and above 0 know, and within this type of knowledge, the students who answered to know the term clustering above 50 were considered to have an excellent level of knowledge.
- **What is the main goal of the k-means clustering algorithm:** the nominal variable – it is a multiple-choice question, in which the participant can only signal one answer and only one answer is correct, according to the Appendix 2.

Between the first and second part of the questionnaire, it was made an intervention based on the improved k-means clustering model. It was proposed to the students for testing the simulation model and try for themselves to find the optimal number of clusters**.**

2. **Second part:** After testing the simulation in the NetLogo software, the students were put to the test with 2 questions about what they tested in the model for evaluating the impact in their knowledge about the algorithm, and finally the students were asked to evaluate the interaction with the improved model in their expertise in clustering:
    - **How the optimal number of clusters is reached:** the nominal variable - it is a multiple-choice question, in which the participant can only signal one answer and only one answer is correct, according to the Appendix 2.
    - **What happens when the algorithm finds the optimal number of clusters:** nominal variable - it is a multiple-choice question, in which the participant can only signal one answer and only one answer is correct, according to the Appendix 2.
    - **The interaction with the algorithm in NetLogo web help you understand what the k-means clustering method consists of:** qualitative ordinal variable – it was defined a scale from 1 to 5, being that:

      1) It didn't help anything.
      2) It helped a little bit.
      3) It helped reasonably.
      4) It helped a lot.
      5) It helped perfectly.

The exact questions of the questionnaire can be seen in Appendix 2.

5.3. The procedures of data collection

In the context of this dissertation, the questionnaire survey was applied via institutional e-mail to students of the University of Porto and via publications in social networks in specific groups of students of the University of Porto. The target population of the questionnaire was chosen because of my proximity to the students of the University of Porto, and therefore it is a more reliable population at the moment of data collection. The questionnaire was conducted in the period from May 1, 2018 to July 31, 2018, that is, the questionnaire was open to answers for three consecutive months.

This survey consists of eight closed questions and an open question where each respondent responds through given options. The open answer is where the student indicates his or her nationality. In the ninth and last question, the participant indicates how useful the improved model in learning the algorithm is to be perceived if the student learned the algorithm through the answers to the previous questions, that is, if he was correct in the answers related to what he saw in the simulation of the model, and whether he actually thought the model helped him. The questions presented in the questionnaire survey, which can be seen in Appendix 2, were selected according to the concepts discussed in the literature review and are related to the subject under investigation.

The questionnaire was answered by 206 students from different faculties of University of Porto within a time horizon of three months, between 1 of May of 2018 and 31 of July of 2018, as previously mentioned.

5.4. Data analysis

For the presentation of the data, tables and graphs were used, with the respective statistical data preceded by analysis. Since the objective of the research is to understand the impact of a new method of learning an algorithm, more specifically, the learning of the k-means clustering algorithm in NetLogo, we applied simulation-related questions that the student could test in the questionnaire. Thus, to analyze the data it is interesting to analyze the associations between the variables and their dependencies to see if the student understand the algorithm, according to their academic background, age and master course.

Data analysis was performed using descriptive and inferential statistics, using SPSS-24.0 software (IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.).

Taking into account the fulfillment of the necessary criteria for the performance of parametric tests, and after performing the Kolmogorov Smirnov test to assess the distribution of the variables (Maroco, 2014), it is concluded that the variables under study do not follow a normal distribution and do not fulfill the necessary criterion for the performance of tests parametric tests in order to test the association and heterogeneity among the different constituent variables of the study. Taking into account that the null hypothesis (H0) for the Kolmogorov Smirnov normality test is that the data are normally distributed, and since the p-value result was ($p < 0.05$) for the study variables as a function of the groups, we reject the null hypothesis (H0) and assume that the variables do not follow a normal distribution. In this way, non-parametric tests were used, namely the chi-square test, the Mann-Withney test and the Kruskall-Wallis test.

For the association of categorical variables, the chi-square test (X^2) is used to test whether two or more independent populations (or groups) differ in relation to a particular characteristic, i.e., if the frequency with which the elements of the sample are distributed by the classes of a qualitative variable is random or not. Continuity Correction was also used since they are 2x2 Tables (Maroco, 2014).

5.4.1. Characterization of the sample

Regarding the nationality of participants (Figure 9), the majority were Portuguese (n=177, 85.9%), followed by Brazilians (n=17, 8.3%).

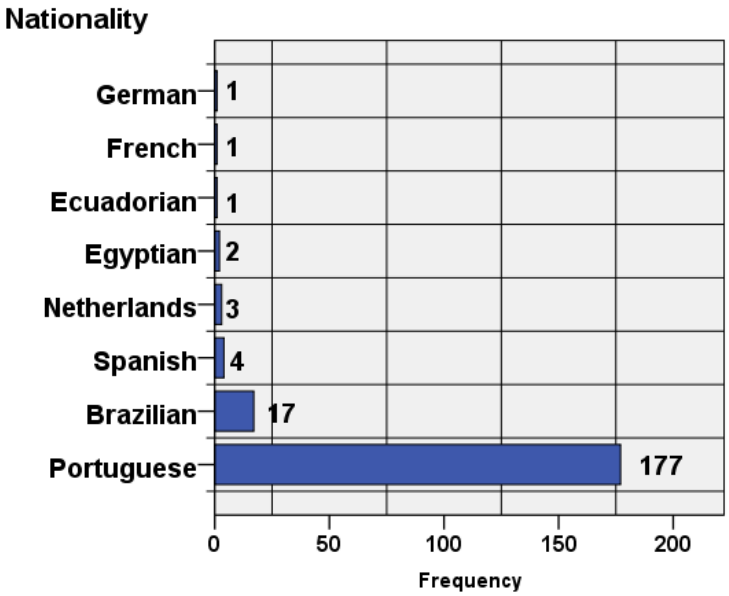

Figure 9 - Frequency of the nationality of participants

Regarding age (Figure 10), the majority of participants were between 18-24 years old (59.71%), followed by participants aged 25-34 years (30.58%).

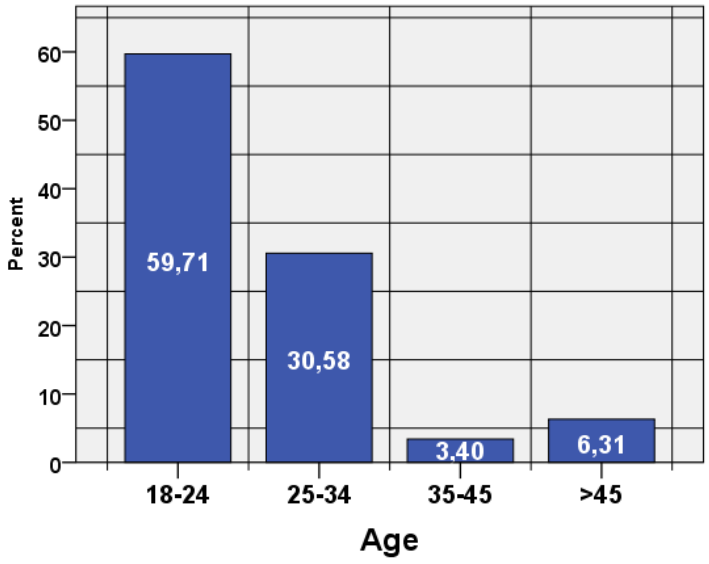

Figure 10 - Frequency of the age of participants

As for the background (Figure 11), the majority of participants were from Economics (35.44%), followed by Management (23.79%) and Engineering (11.65%).

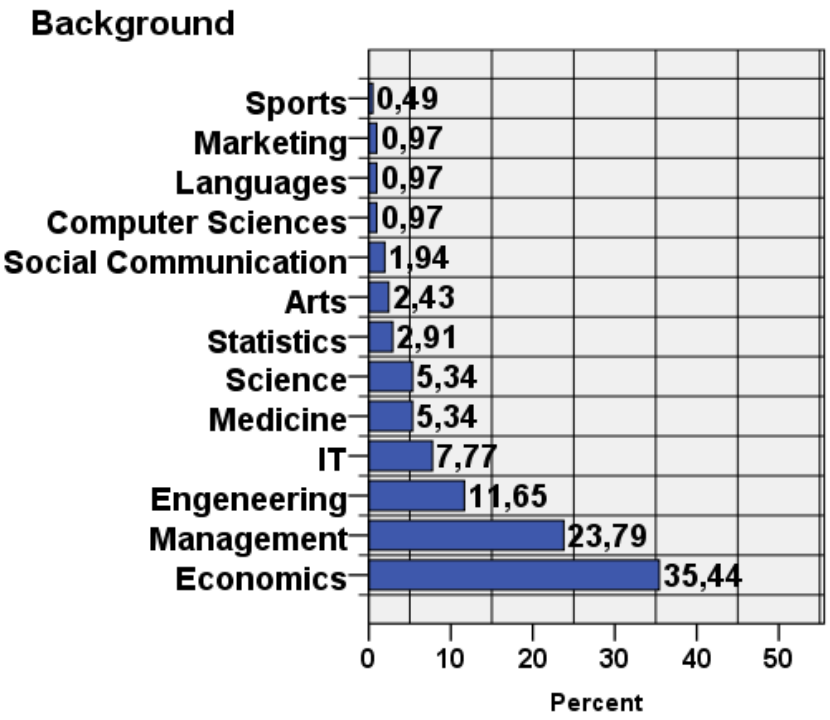

Figure 11 - Background of participants

With regard to the master course (Figure 12), the majority of the participants were from Data Analytics (34.18%) and Management (24.05%).

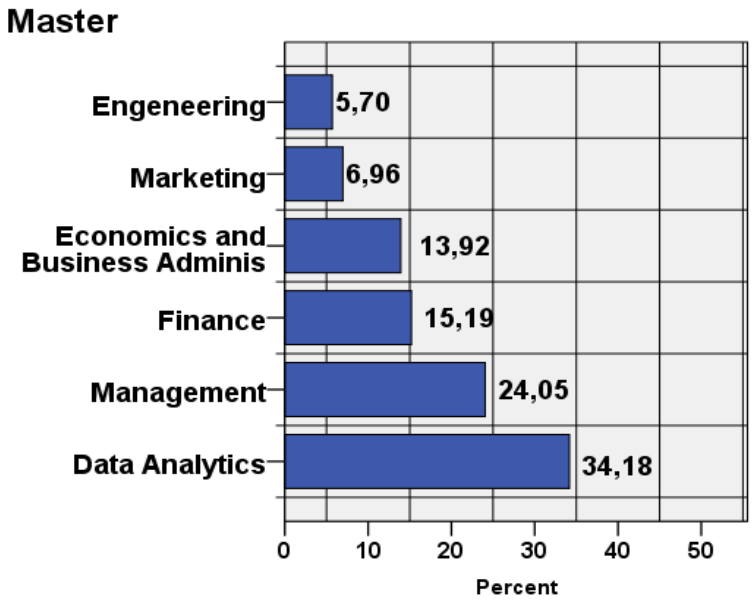

Figure 12 - Master area of participants

Regarding the fact that the participant, if he or she considers it an expertise in clustering (see Figure 13) - that is, the participant answered to having a level of knowledge in clustering above 0 – it was found that the majority of participants considered to be an expertise in clustering (n=172, 83.50%).

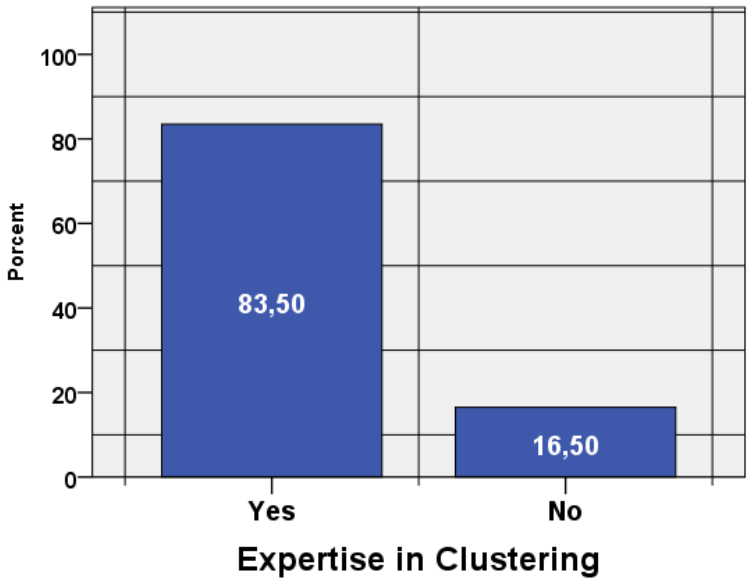

Figure 13 - Percentage of participants who consider themselves expertise or not in clustering

Participants who considered that they did not have expertise in clustering (n=34; 16.50%), considered that their level of Expertise in Clustering was zero. Of the participants who considered expertise in clustering (n=172; 83.50%), they considered that their level of expertise in Clustering was a minimum of 1 and a maximum of 100, a mean of 41.31 and a standard deviation of 30.268.

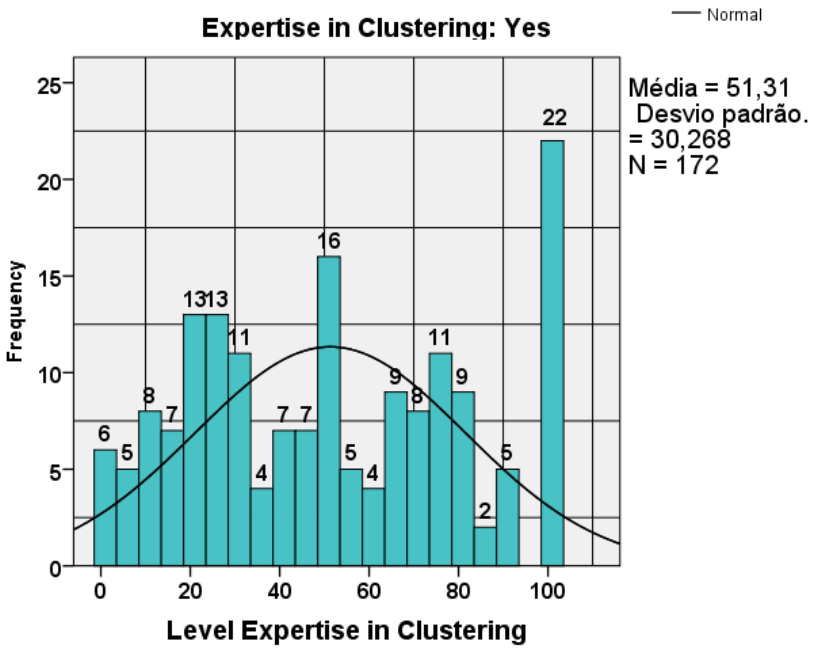

Figure 14 - Frequency of the level of expertise in clustering of the participants who consider themselves expertises in clustering (>1 in the level)

5.4.2. Exploratory analysis

From the association between the variable expertise in clustering and the type of responses of the variables - goal of k-means algorithm, optimal number of clusters and convergence criterion - and age (see Table 5) we found that there are two statistically significant associations, which suggests the dependence of the variables expertise in clustering and convergence criterion (p=0.022), with participants who considered expertise in clustering with a higher percentage of correct answers (88.1%); and between the variables expertise in clustering and age (p=0.037). The middle age and the older participants, aged 25-34 years and> 35 years respectively, had a majority of (92%) and (90%), respectively, in regard to considering themselves experts in clustering.

		Expertise in Clustering						
		Yes		No		Total		*p*
Goal of K-means algorithm	Wrong	37	75,5%	12	24,5%	49	100%	
	Right	135	86,0%	22	14,0%	157	100%	
Total		172	83,5%	34	16,5%	206	100%	0,133
Optimal number of clusters	Wrong	59	80,8%	14	19,2%	73	100%	
	Right	113	85,0%	20	15,0%	133	100%	
Total		172	83,5%	34	16,5%	206	100%	0,569
Convergence criterion	Wrong	53	74,6%	18	25,4%	71	100%	
	Right	119	88,1%	16	11,9%	135	100%	
Total		172	83,5%	34	16,5%	206	100%	**0,022**
Age	18-24	96	78,0%	27	22,0%	123	100%	
	25-34	58	92,1%	5	7,9%	63	100%	
	> 35	18	90,0%	2	10,0%	20	100%	
Total		172	83,5%	34	16,5%	206	100%	**0,037**

Table 5 - SPSS output from the Chi-squared test

From the comparison of the rating of the simulation in Netlogo between whether or not the participant considers an expertise in clustering (see Table 6), we find that there is no statistically significant difference.

	Expertise in Clustering				
	Yes (n=172)		No (n=34)		
	average	sd	average	sd	*p*
(a)The simulation in Netlogo helps?	3,62	1,06	3,26	0,96	0,070

(a) (1=It didn't help anything | 2=It helped a little bit | 3=It helped reasonably | 4=It helped a lot | 5=It helped perfectly)

Table 6 - SPSS output from Mann-Whitney test

From the comparison of the level of expertise in clustering between the type of answers to the question related to the goal of k-means algorithm (see Table 7), we find that there are statistically significant differences (p=0.000), whose mean was higher in the group of participants who answered correctly.

	Goal of K-means algorithm				
	Wrong (n=49)		Right (n=157)		
	average	sd	average	sd	***p***
(a) Level Expertise in Clustering	20,98	18,56	49,66	34,36	**0,000**

(a)(1=Low Level | 100 = Perfect Level)

Table 7 - SPSS output from the Mann-Whitney test across values of the answers to the question related to the Goal of K-means algorithm

From the comparison of the level of expertise in clustering between the type of answers to the question related to the optimal number of clusters founded in the simulation (see Table 8), we find that there are statistically significant differences (p=0.001), whose mean was higher in the group of participants who answered correctly.

	Optimal number of clusters				
	Wrong (n=73)		Right (n=133)		
	average	sd	average	sd	***p***
(a) Level Expertise in Clustering	31,37	27,13	49,14	35,20	**0,001**

(a)(1=Low Level | 100 = Perfect Level)

Table 8 - SPSS output from the Mann-Whitney test across values of the answers to the question related to the Optimal number of clusters founded in the simulation

Comparing the level of expertise in clustering between the type of answers to the question related to the convergence criterion of the Elbow method (see Table 9), we found that there are statistically significant differences (p=0.000), whose mean was higher in the group of participants who answered correctly.

	Convergence criterion				
	Wrong (n=71)		Right (n=135)		
	average	sd	average	sd	***p***
(a) Level Expertise in Clustering	29,66	27,36	49,77	34,57	**0,000**

(a)(1=Low Level | 100 = Perfect Level)

Table 9 - SPSS output from the Mann-Whitney test across values of the answers to the question related to the convergence criterion of the Elbow method

From the comparison of the level of expertise in clustering and the rating of the simulation in Netlogo between the age group (see Table 10), we find that there is a statistically significant difference in the level of expertise in clustering (p=0.001), whose

average level was higher in the group of participants >35 years (mean=61.45) and in the 25-34 years group (mean=49.24).

	Age						
	18-24 (n=123)		25-34 (n=63)		> 35 (n=20)		
	average	sd	average	sd	average	sd	*p*
(a) Level Expertise in Clustering	36,54	34,25	49,24	29,51	61,45	32,32	**0,001**
(b) The simulation in Netlogo helps?	3,48	1,00	3,68	1,01	3,65	1,46	0,325

(a)(1=Low Level | 100 = Perfect Level)
(b)(1=It didn't help anything | 2=It helped a little bit | 3=It helped reasonably | 4=It helped a lot | 5=It helped perfectly)

Table 10 - SPSS output from the Kruskal-Wallis test across the age groups

In order to obtain a better understanding of the relations between the different output variables seen in the above results, it is important to evaluate its relations one by one. With this goal in mind, it will be presented a series of charts depicting the relations of the values of each output variable.

The first chart (Figure 15) is a projection line chart, which represents the relation between the average of the level of expertise in clustering and students' age and the answer to the first question (before the simulation in NetLogo software) related to the goal of k-means algorithm. It can be seen, that older students (>35 years old) considers themselves, in average, with a higher level of expertise in clustering. Thus, it can also be seen that the students in this age range answered, always, correctly to this question.

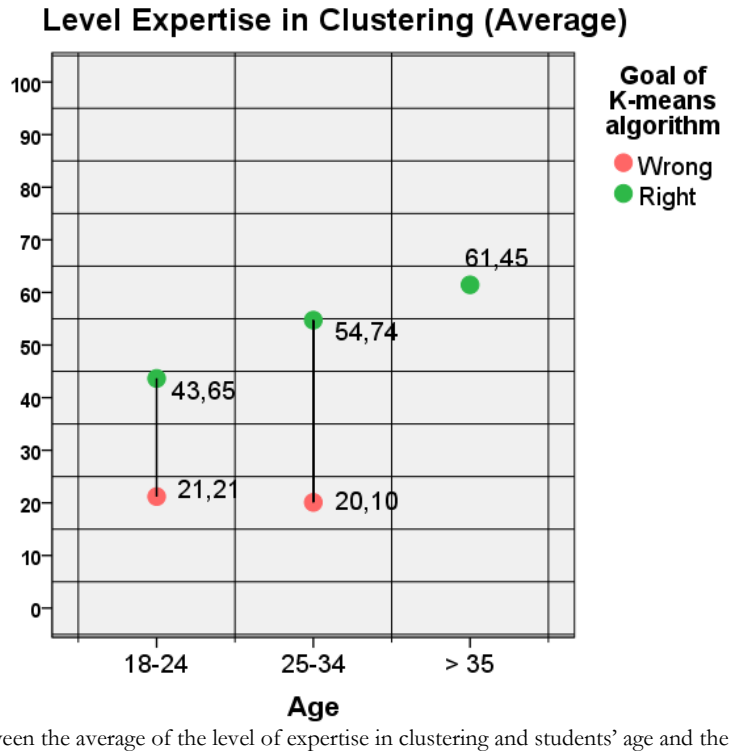

Figure 15 - Relation between the average of the level of expertise in clustering and students' age and the answer to the question related to the goal of k-means clustering algorithm

The second chart (Figure 16) is also a projection line chart, which represents the relation between the average of the level of expertise in clustering and students' age and, for this time, the answer to the question related to the optimal number of clusters founded in the simulation. It can be seen that older students (>35 years old) who answered correctly to this specific question consider themselves, on average, with a higher level of expertise in clustering. Thus, the discrepancy between wrong and right answers on optimal number of clusters question, depending on the level of expertise is high, with an average of 73,31 (from 0 to 100) in the level of expertise in clustering for the older students who answered correctly and with an average of 35,71 (from 0 to 100) in the level of expertise in clustering for the older students who answered wrongly.

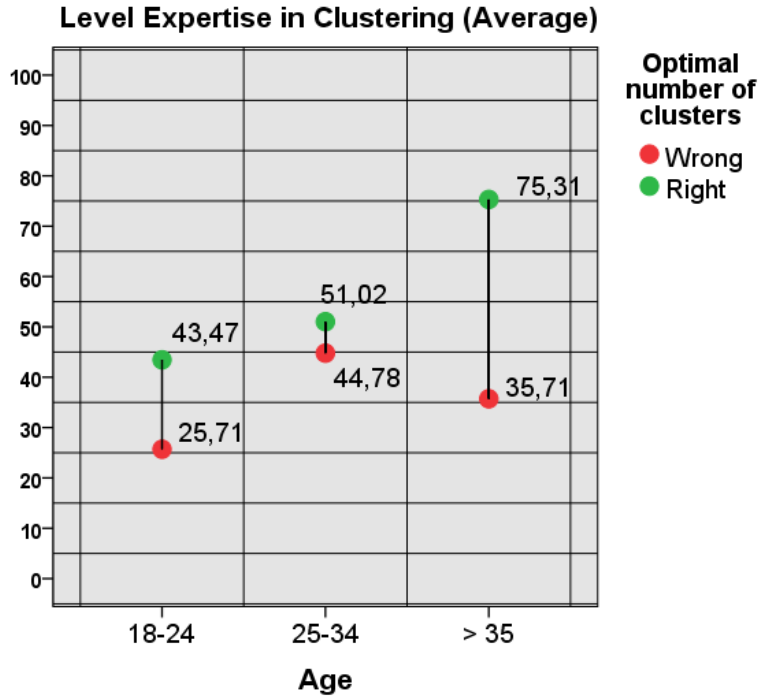

Figure 16 - Relation between the average of the level of expertise in clustering and students' age and, for this time, the answer to the question related to the optimal number of clusters founded in the simulation

The third chart (Figure 17) represents the relation between the average of the level of expertise in clustering and students' age and the answer to the question related to the convergence criterion of the Elbow method. It can be seen older students (>35 years old) consider themselves once again, on average, with a higher level of expertise in clustering. The discrepancy between wrong and right answers on convergence criterion question, depending on the level of expertise is low, with an average of 63,03 (from 0 to 100) in the level of expertise in clustering for the older students who answered correctly, and with an

average of 55 (from 0 to 100) in the level of expertise in clustering for the older students (>35 years old) who answered wrongly.

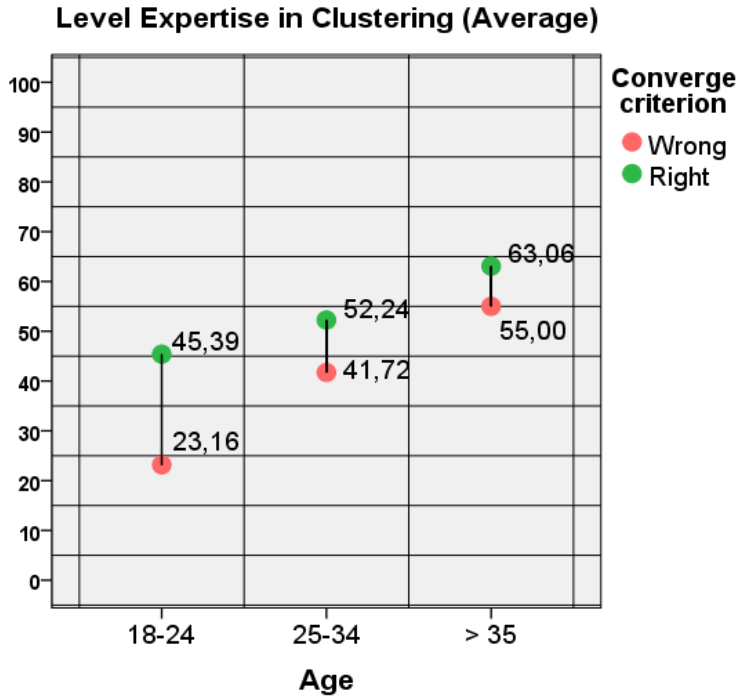

Figure 17 - Relation between the average of the level of expertise in clustering and students' age and the answer to the question related to the convergence criterion of the Elbow method

From the three charts above, it can be said that older students (>35 years old) have a self-perception of their level of expertise in clustering above younger students (18-24 years old). For proving this point, it is presented the following chart (Figure 18) where it can be seen, in average, older students (>35 years old) consider themselves with a 61.45 (from 0 to 100) level of expertise in clustering. On the contrary, younger students (18-24 years old) consider themselves with a 36.54 (from 0 to 100) level of expertise in clustering.

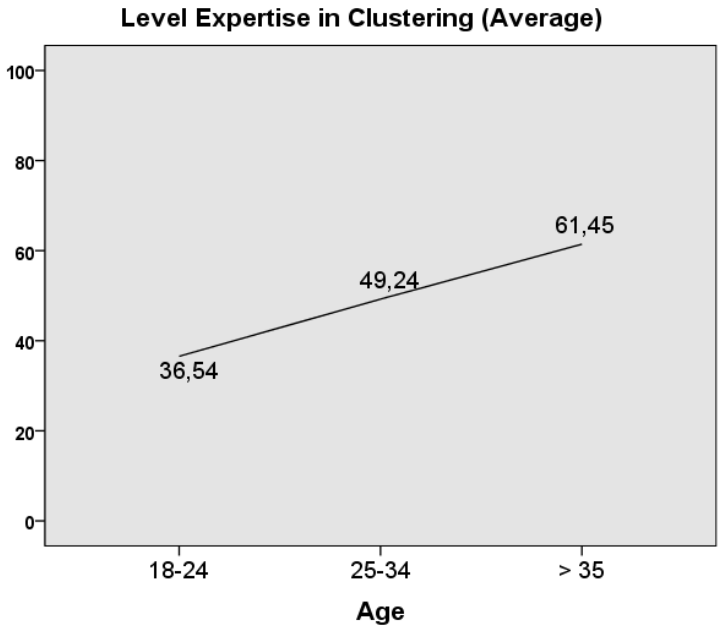

Figure 18 - Relation between the average of the level of expertise in clustering and students' age

The following charts represent the relation between the average of the effectiveness of the simulation on the learning process on the present study, through the rating of the simulation done, and their ages and the answers to the three technical questions of the questionnaire.

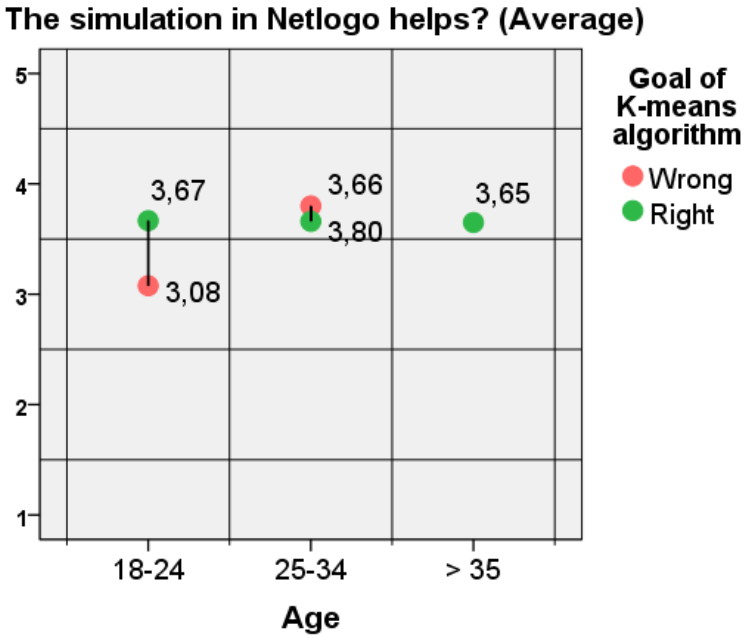

Figure 19 - Relation between the average of the rating of the simulation in NetLogo and students' age and the answers to the question related to the goal of k-means algorithm

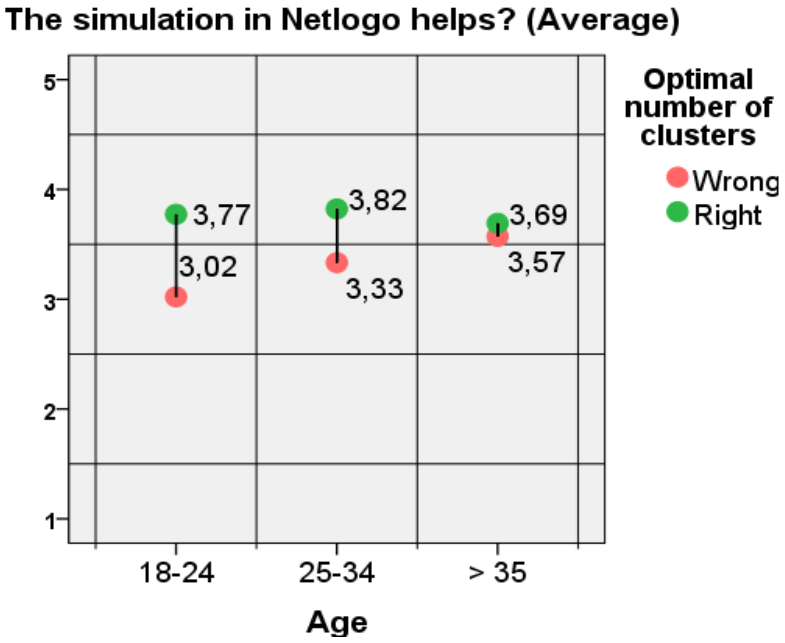

Figure 20- Relation between the average of the rating of the simulation in NetLogo and students' age and the answers to the question related to the optimal number of clusters founded in the simulation

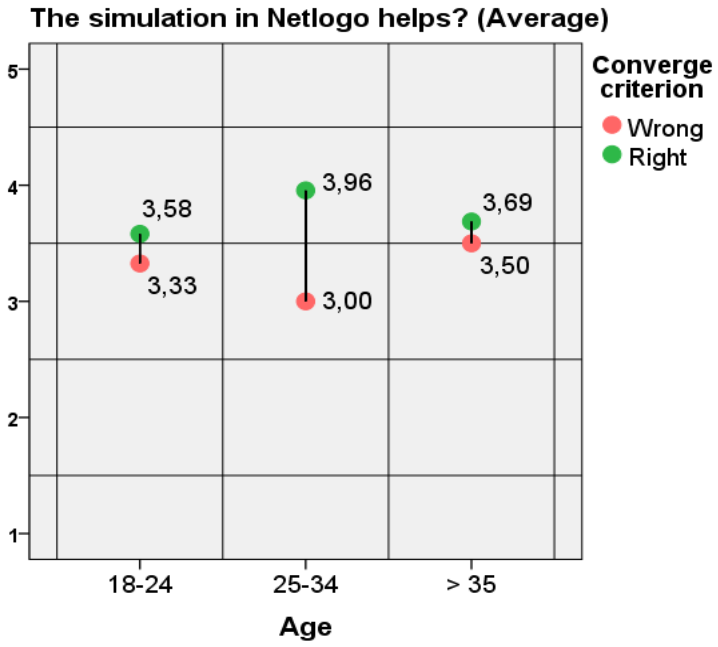

Figure 21 - Relation between the average of the rating of the simulation in NetLogo and students' age and the answers to the question related to the convergence criterion of the Elbow method

In three projections charts (Figure 19, 20 and 21), it can be seen that the learning process of the k-means algorithm in the NetLogo software was efficient because, in average, students rate it above the average, not forgetting that student could rate the simulation from 1 to 5, according to the definitions described on section 5.2 of this dissertation.

From the charts, it can be said that in two of the answers of the three questions, i.e., in two of three charts above, middle age students (25-34 years old) who answered correctly rate the simulation, in average, with a higher classification than younger and older students.

When comparing the relation of the average of the students' rating of the simulation and their age and the answer to the question related with the goal of k-means algorithm, it can be seen that older students (>35 years old) rate simulation in NetLogo, on average, above average and answered the question, always, correctly. The younger students (18-24 years old) present a discrepancy between right and wrong answers and their rating on the simulation, in average, but still the classification of the simulation is above average (from 1 to 5, according to section 5.2), as it can be seen in Figure 19.

When comparing the relation of the average of the students' rating of the simulation and their age and the answer to the question related with the optimal number of clusters founded in the simulation, the middle age students (25-34 years old) who answered correctly to the question rate the simulation above the average, but the lower discrepancy

between right and wrong answers and their rating on the simulation are presented for older students (>35 years old), with an average of 3.69 (from 1 to 5) on the rate of the simulation when they answer right and with an average of 3.57 (from 1 to 5) on the rate of the simulation when they answer wrong to the question of the optimal number of clusters, as it presented in Figure 20.

When comparing the relation of the average of the students' rating of the simulation and their age and the answer to the question related with the convergence criterion of the Elbow method, the middle age students (25-34 years old) who answered correctly to the question between rate, once again, the simulation above the average, but, in this specific question, this students present the higher discrepancy between right and wrong answers and their rating on the simulation with an average of 3.96 (from 1 to 5) on the rate of the simulation when students answer right to the question, and with an average of 3.00 (from 1 to 5) when students answer wrong to the question. The older students (>35 years old) present, once again, the lower discrepancy between right and wrong answers and their rating on the simulation, in average, with an average of 3.69 (from 1 to 5) on the rate of the simulation when students answer right to the question, and with an average of 3.50 (from 1 to 5) when students answer wrong to the question, as it can be seen in Figure 21.

Considering the main goal of this dissertation, it is particularly relevant to understand which was the impact of the simulation on NetLogo on the students' learning process of the algorithm. With this in mind, a decision tree was developed, which attempts to classify the rating of the simulation in NetLogo by the students based only on their level of expertise in clustering.

The dataset from the 206 questionnaires was separated into training and test data, with the training data composed of 150 observations and the test data is composed of 56 observations.

The decision tree was obtained using the R packages "rpart" an "rpart.plot" and it is presented in Figure 22.

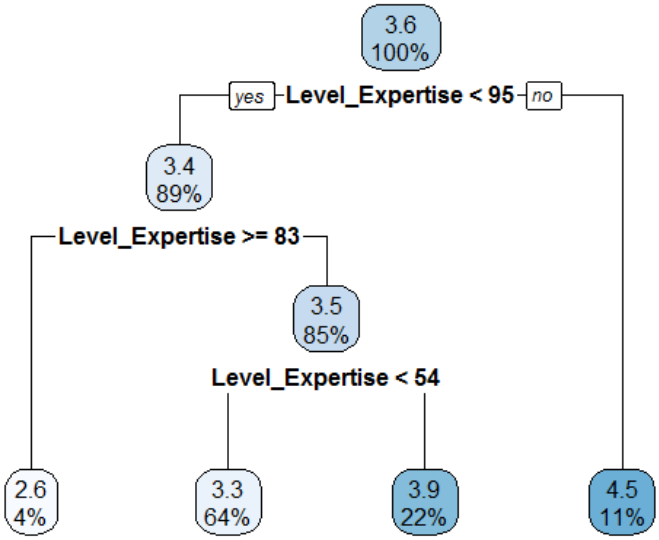

Figure 22 - Decision Tree for predicting the rank of the simulation based on the level of expertise in clustering

Based on the mean squared error (MSE) obtained from the classification tree on the training set, different degrees of post-pruning were tried and the presented tree in Figure 22 presents a MSE of 0.3409091, which is not considered low but also is not high. For that reason, we can say that the level of expertise in clustering has a credible weight on predicting the rating of the student on the simulation on NetLogo, i.e., the knowledge of the student on the algorithm has a certain weight when the student classifies the simulation performed on NetLogo.

In order to effectively evaluate the impact of the simulation in NetLogo on learning the algorithm of k-means clustering by the student, and taking into account the performance metric defined in section 4.5., we will analyze the results obtained in the responses of the 206 students in the questionnaire through the performance metric.

As the two questions taken into account in the performance metric were normalized and therefore are now two binary variables it is possible to have three types of results in the metric:

- 0 when the student responds incorrectly to both questions, i.e., has performed poorly.
- 0.5 when the student responds correctly to one of the two questions and therefore is considered to have performed well.
- 1 when the student correctly answers the two questions proposed, and therefore, had a perfect performance after the simulation.

The following pie chart, in Figure 23, shows the percentage of students corresponding to each type of performance referred to.

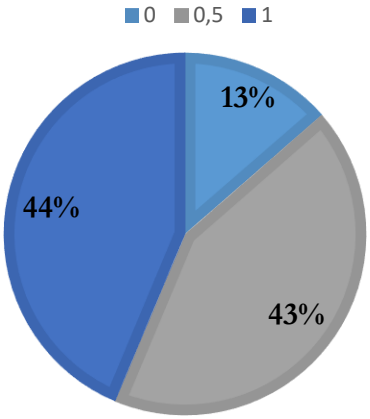

Figure 23 - Pie chart with the percent of students based on the performance metric

We can see from Figure 23 that most students scored at least one question, and therefore 87% of the students had a good performance in learning the algorithm through simulation.

	Performance metric
average	65,05%
sd	34,81%

Table 11 - Summary of performance metric

We can verify through Table 11 that, on average, students performed above average (65%), and the standard deviation is in 35%, which means that there is some dispersion of data in the sample relative to the average, but it is not very high.

An interesting comparison mentioned earlier, is to evaluate the students' performance results through the metrics against students' evaluation in relation to the degree of help that the simulation gave them in learning the algorithm. In order to make this comparison, the variable related to the simulation rate, which ranges from 1 to 5 (as defined in section 5.3), had to be normalized, that is, for the variable to comprise values between 0 and 1, we proceeded to the normalization of the observations according to $xnew = \frac{xobs-xmin}{xmax-xmin}$, with *xnew* being the normalized observation for the variable, *xobs* being the observation, *xmin* being the minimum value observed in the variable and *xmax* being the maximum value observed in the variable in study. Thus, the variable began to comprise values from 0 to 1, which on this scale and, according to section 5.3, correspond to:

- 1 has turned into 0 - It didn't help anything.
- 2 has turned into 0.25 - It helped a little bit.
- 3 has turned into 0.5 - It helped reasonably.
- 4 has turned into 0.75 - It helped a lot.
- 5 has turned into 1 - It helped perfectly.

With the observations of the variable normalized, it is, thus, possible to compare the results of this variable with the results obtained through the performance metric of the students.

	Rate of Simulation
average	63,96%
sd	26,30%

Table 12 - Summary of the normalized rate of simulation

We can verify that, on average, the students classified the simulation in a range between 0.5 – 0.75, but closer to 0.75, which means that, on average, the simulation helped a lot in learning the k-means algorithm clustering.

Taking into account the metric created to measure the student's performance and the variable that measures the degree of help that the simulation gives in learning the algorithm, and comparing them according to whether the student considers himself or herself an expert in clustering, we can verify in Figure 24 that students who consider themselves to be expert in clustering have a better performance in the questions after simulation and also consider that the simulation in NetLogo helps them learn the algorithm, when compared to those that do not consider themselves experts in clustering.

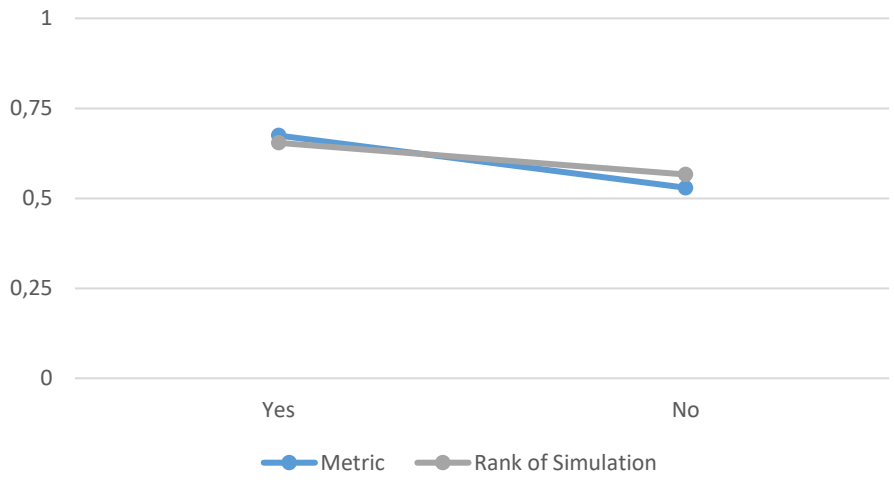

Figure 24 - Comparison of the metric with the rank of simulation depending on whether or not the student considers clustering expertise

Comparing the performance of the student and the variable that measures the degree of help that the simulation gives in learning the algorithm, and comparing them according to the student's age, we can see in Figure 25 that as age increases, the student presents a better performance after the simulation and considers that the simulation was useful for their learning process.

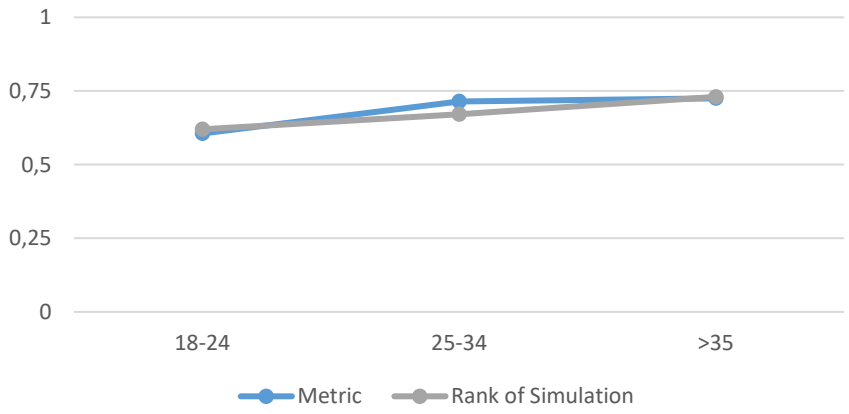

Figure 25 - Comparison of the metric with the rank of simulation regarding the students' age

Comparing the performance of the student and the variable that measures the degree of help that the simulation gives in learning the algorithm, and comparing them according to the academic background of the student, we can see in Figure 26 that the students with the best performance after the simulation have as background Statistics and Computer Science, and classify the simulation as a good help to learning the algorithm. On the other hand, the students with the worst performance are students with a background of Languages and Social Communication with a performance below the average, but even if they do not have an average performance, these students consider that the simulation helped them to learn the algorithm above the average.

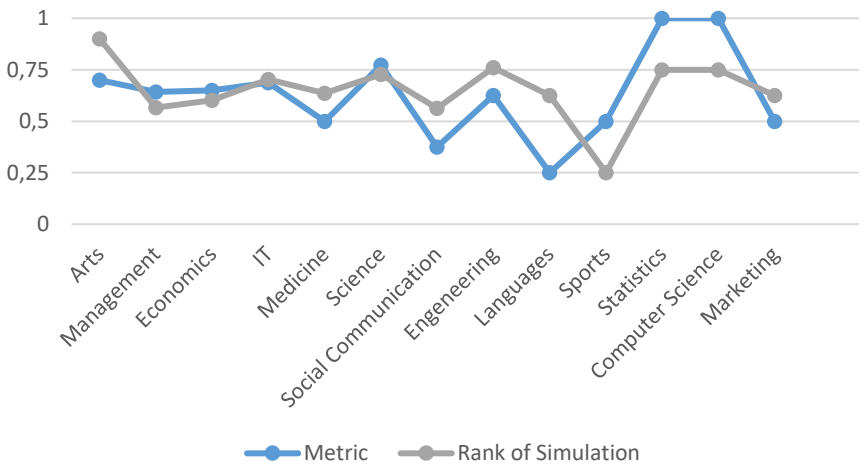

Figure 26 - Comparison of the metric with the rank of simulation regarding the students' background

Comparing the performance of the student and the variable that measures the degree of help that the simulation gives in learning the algorithm, and comparing them according to the masters course that the student attends or attended, we can verify in Figure 27 that the students with the best performance after simulation are of the masters of Tax and Finance, Electronic Engineering, Accounting and Mathematical Engineering and, they classify the simulation above average, this is, as a good help to the learning of the algorithm, except for the students of the area of master of Tax and Finance. In turn, the students with the worst performance are students in the master course of Energy, Law and Biology with a below-average performance simulation, but even though they do not have an average performance, these students consider that the simulation helped them learn the algorithm, with the exception of students in Law's master course.

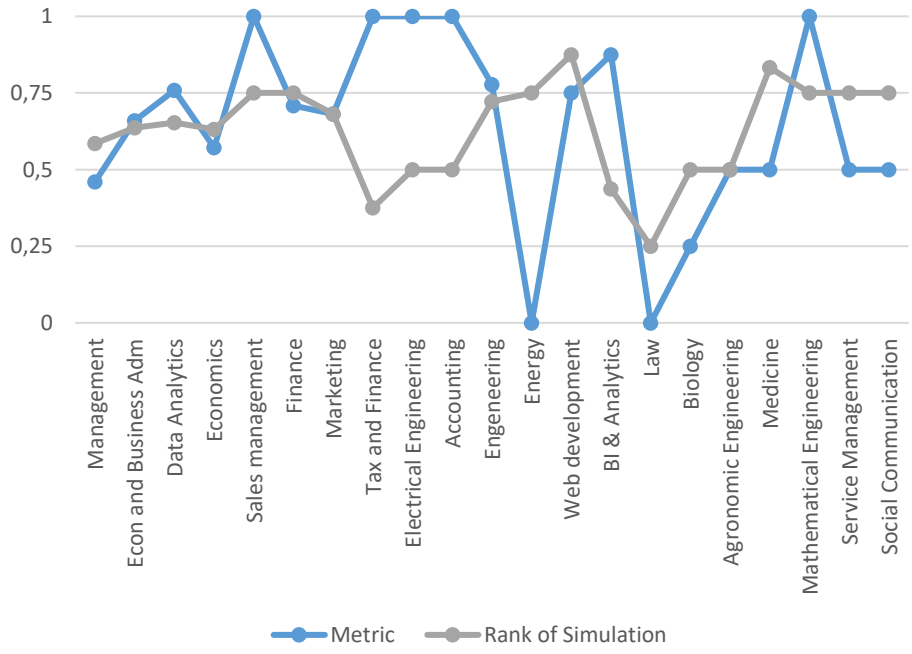

Figure 27 - Comparison of the metric with the rank of simulation regarding the students' master area

With the results obtained by using the performance metrics of the students and comparing them with the classifications made by the students about the degree of help that the simulation gave in their learning of the algorithm, it is possible to conclude that the students had a performance after simulation above the average and classified the improved model in NetLogo as useful in learning the algorithm k-means clustering.

5.5. Results discussion

Recall that the present study is only a pilot, with the main objective of analyzing in what sense a type of teaching approach different from the conventional one will be linked to a better learning of the student. The results obtained through the questionnaire were presented in the previous section. Learning analytics was adopted, as mentioned in the literature review of this dissertation since data from an educational database were analyzed. It was decided to briefly present and comment on the results obtained in the previous section, as they were presented. Thus, in this context, and throughout this section, the final discussion of the results obtained will be held.

To discuss the results of the questionnaire, it is important to highlight some hypothesis done with chi-squared test, for evaluating the independence between variables from Table 5.

The null hypothesis:

H0: There are independence between the variable level of expertise in clustering and the variable convergence criterion.

An alternative hypothesis:

H1: There are no independence between the variable level of expertise in clustering and the variable convergence criterion.

With a confidence level of 0.95, the null hypothesis is rejected because the results presented in Table 5 revealed the existence of statistically significant dependence between the variable level of expertise in clustering and the variable convergence criterion, with a p-value of 0.022, lower than 0.05, which is the confidence level usually adopted in research. This suggests the dependence of the two variables, and it can be validated in the same table that participants who considered themselves experts in clustering (answered to this question more than 1 on the level) had a higher percentage of correct answers in the question of the convergence criterion of the Elbow method (after the simulation). This proves that the improvement of a student's self-concept seems to be valued as an educational outcome in its own right (Shavelson et al, 1996) because the knowledge of the algorithm leads the student to success on the test.

The chi-squared test, for evaluating the independence between variables from Table 5, permits to test another hypothesis.

The null hypothesis:

H0: There are independence between the variable level of expertise in clustering and the variable age.

An alternative hypothesis:

H2: There are no independence between the variable level of expertise in clustering and the variable age.

The results presented in Table 5 revealed the existence of a statistically significant dependence between the variable level of expertise in clustering and participants' age, with a p-value of 0.037. Therefore, for a confidence level of 0.95, the null hypothesis is rejected, suggesting the dependence between the variables, also it can be seen in Table 5 that 92.1% of middle age participants (25-34 years old) and 90.0% of older (>35 years old) participants consider themselves experts in clustering. This proves what Bloom wrote in 2006: older people incorporate certain of the positive stereotypes of ageing into their self-concept (Bloom, 2006).

With the result of the nonparametric test of Mann-Whitney, on Table 6, it is important to note the following hypothesis test, which is applied in two independent samples (Marôco, 2014). The values calculated by the test evaluate the degree of interlacing of the data of the two groups after the ordering.

The null hypothesis:

H0: There are no differences in the average rating of the simulation between the fact that the participant considers or not an expertise in clustering.

An alternative hypothesis:

H3: There are differences in the average rating of the simulation between the fact that the participant considers or not an expertise in clustering.

With a confidence level of 0.95, the null hypothesis is not rejected because the results presented in Table 6 do not reveal a statistically significant difference, with a p-value of 0.07, higher than 0.05. This suggests there are no statistically significant differences between the two variables – the rating of the simulation and the level of expertise in clustering. Although there are no statistically significant differences in the rating of the simulation between the fact that the participant considers whether or not an expertise in clustering, it can be seen in Table 6 that considering the participant a clustering expertise or not, the average simulation rating in both cases it is above average because the simulation effectiveness rating has a range of values from 1 to 5, as specified in section 5.2 of this dissertation. Thus, being the classification 3 - it helped reasonably - it is possible to state

that, on average, the participants ranked simulation as useful in learning. It can also be seen, in Figure 24, that the students with the best performance after the simulation are those who consider themselves experts in clustering, and in turn, classify the simulation as a good help for learning the algorithm under study. Although it was not proved by the test, this suggests that the simulation facilitates the understanding of the content for those who are already familiar with the algorithm (Trindade, 2005).

From Table 7, which represents the results of a nonparametric test of Mann-Whitney, which in turn tests the hypothesis of existing statistically significant differences in the average level of expertise in clustering of the participant between the type of the answers to the question related to the goal of k-means algorithm.

The null hypothesis:

H0: There are no differences in the average level of expertise in clustering between the type of answers to the question related to the goal of k-means clustering.

An alternative hypothesis:

H4: There are differences in the average level of expertise in clustering between the type of answers to the question related to the goal of k-means clustering.

With a confidence level of 0.95, the null hypothesis is rejected because the results presented in Table 7 revealed the existence of statistically significant differences in the average level of expertise in clustering between the type of answers of the question related to the goal of k-means clustering, with a p-value of 0.000. As the p-value is lower than 0.05, this suggests there are statistically significant differences between the variable level of expertise in clustering and the type of answers related to the question of the goal of k-means clustering. It is possible to see in Table 7 that participants who answered correctly to the question of the goal of k-means clustering have a higher average value in the level of expertise in clustering (49.66 for those who scored in the answer versus 20.98 for those who did not answer correctly). This means that participants who, on average, consider themselves more expertise in clustering match, more frequently, in the answer related to the question of the goal of the k-means algorithm.

The same nonparametric test was done for comparing the level of expertise in clustering between the type of the variables with the answers to the questions conducted in the questionnaire after the simulation – the question related to the optimal number of clusters obtained in the simulation (Table 8) and the convergence criterion (Table 9).

From Table 8, the null hypothesis:

H0: There are no differences in the average level of expertise in clustering between the type of answers to the question related to the optimal number of clusters founded in the simulation.

An alternative hypothesis:

H5: There are differences in the average level of expertise in clustering between the type of answers to the question related to the optimal number of clusters founded in the simulation.

With a confidence level of 0.95, the null hypothesis is rejected because the results presented in Table 8 revealed the existence of statistically significant differences in the average level of expertise in clustering between the type of answers of the question related to the optimal number of clusters, with a p-value of 0.001. As the p-value is lower than 0.05, this suggests there are statistically significant differences between the variable level of expertise in clustering and the type of answers to the question of the optimal number of clusters founded. It is possible to see in Table 8 that participants who answered correctly to the question related to the optimal number of clusters have a higher average value in the level of expertise in clustering (49.14 for those who scored in the answer versus 31.37 for those who did not answer correctly). This means that participants who, on average, consider themselves more expertise in clustering match, more frequently, in the answer to the question of the optimal number of clusters.

From Table 9, the null hypothesis:

H0: There are no differences in the average level of expertise in clustering between the type of answers to the question related to the convergence criterion of the elbow method.

An alternative hypothesis:

H6: There are differences in the average level of expertise in clustering between the type of answers to the question related to the convergence criterion of the elbow method.

With a confidence level of 0.95, the null hypothesis is rejected because the results presented in Table 9 revealed the existence of statistically significant differences in the level of expertise in clustering between the type of answers of the question related to the convergence criterion of the elbow method, with a p-value of 0.000. Therefore, for a level of significance of 0.05, the result suggests there are statistically significant differences between the variable level of expertise in clustering and the type of answers to the question of the convergence criterion of the elbow method. It is possible to see in Table 9 that participants who answered correctly to the question related to the convergence criterion of

the elbow method have a higher average value in the level of expertise in clustering (49.77 for those who scored in the answer versus 29.66 for those who did not answer correctly). This means that participants who, on average, consider themselves more expertise in clustering match, more frequently, in the answer to the question of the convergence criterion of the elbow method. These results are in line with Cavalcanti's 2003 study, which mentions that a positive self-perception of the student will lead to a higher performance on any subject.

The non-parametric test of Kruskal-Wallis was used to do a comparison of the level of expertise in clustering between the age group. From Table 10, it is possible to do the following hypothesis test:

The null hypothesis:

H0: There are no differences in the level of expertise in clustering between the age group.

An alternative hypothesis:

H7: There are differences in the level of expertise in clustering between the age group.

With a confidence level of 0.95, the null hypothesis is rejected when comparing the variable level of expertise in clustering between the age group, because the results presented in Table 10 revealed the existence of statistically significant differences, with a p-value of 0.001. As the p-value is lower than 0.05, this suggests there are statistically significant differences between the variable level of expertise in clustering and the age group. It is also possible to verify in the results presented in Table 10, that as the participant's age increases, the participant considers, on average, himself in a higher level of expertise in clustering, with an average on the level of expertise in clustering of 61.45 for the older participants (>35 years old), of 49.24 for the middle age group of participants (25-34 years old) and with an average of 36.54 for the younger students (18-24 years old). To reinforce these conclusions and according to the performance metrics developed based on the questions of the questionnaire after the simulation, we observe, in Figure 25, that the older students are the ones that perform better after the simulation. Once again, Bloom's study in 2006 indicates that the similarities in self-perception with age outweigh the differences (Bloom, 2006).

The same nonparametric test was done for comparing the rating of the simulation between the age group.

From Table 10, the null hypothesis:

H0: There are no differences in the rating of the simulation between the age group.

An alternative hypothesis:

H8: There are differences in the rating of the simulation between the age group.

With a confidence level of 0.95, the null hypothesis is not rejected when comparing the variable rating of the simulation between the age group, because the results presented in Table 10 do not revealed the existence of statistically significant differences, with a p-value of 0.325, higher than 0.05. This suggests there are no statistically differences in the rating of the simulation between the age group. Although there are no differences in the rating of the simulation between the age group, it can be seen in Table 10 that a student, of any group age, rated the simulation, on average, above average because the simulation is ranged from 1 to 5 as defined in section 5.2.

In this regard, it is important to note that related analysis has been done in another educational dataset around the world in recent years, is this an important part of the educational data mining process. According to Romero and Ventura (2013), in an educational point of view, the application process of learning analytics can be seen as an iterative cycle of hypothesis and tests formation, as it was done in the results discussion of this dissertation. In this process, the goal is not just to transform data in knowledge, but also to filter the extracted knowledge to decision-making about how to change the educational environment and how to improve the learning process. The goal of educational data mining is to improve education by creating models capable of identifying the main problems associated with poor learning and the results are used as strategies to do an improvement in teaching techniques, for better learning as a future goal. Learning analytics, in its turn, applies known models to response questions related to learning environments and it does not create models. One of the most important applications of learning analytics is to predict students' performance to help, mainly, students identified as students at risk of failure. According to Baker (2010), it is essential to use human judgment in learning analytics using automatized tools, as it was done with the simulation of an improved model of k-means algorithm in an interactive and visual software as NetLogo and the evaluation of the self-perception of the participants' expertise on that.

Regarding the methodological approach used in this dissertation and taking into account the two research questions made in the beginning:

Is it possible to implement improvements in the practice of teaching clustering techniques (more specifically, k-means clustering), based on more visually presentations?

Which are the impact of a new learning method for students' performance and which are the students' opinion about that? Is it easier to learn? Is it more motivating to learn this way?

With the results obtained, it is possible to conclude that improvements in the practice of teaching clustering algorithms through more visual and motivational methods for student learning is conceivable. The results of this research work prove to us that improved teaching can lead to improved student performance, which answers the second research question of this work. We have seen throughout this pilot study that student's performance after a new method of teaching the k-means clustering algorithm was positive and that students rated this method as being useful in their learning experience.

6. Concluding Remarks

Learning analytics is an increasingly useful practice and it is defined as the measurement and segmentation of data about learners used to optimize their learning experience. It is possible to extract useful knowledge from the data collect about learners to analyzing and for try to found patterns to create personalized teaching strategies for the student's weaknesses and strengths. While educational data mining is concerned with the development of methods for exploring the different types of data that come from educational features and using those methods to understand in a better way students and their learning contexts and features, learning analytics is concerned with the analysis and communication of data about learners and their contexts for optimizing the learning process and its environment, as it is defined by the societies of educational data mining and learning analytics, respectively.

Keeping in mind the relevance of understanding the learning process and experience influencing the students' performance, the goal of this dissertation has been to study the students' performance through a non-conventional learning approach of the k-means algorithm in specific, by using a multi-agent programming language and modelling environment. The analysis performed throughout this dissertation, by applying the developed model in the questionnaires to students for evaluating their performance, have led to some conclusions:

i. The performance of the student in a given test is dependent on:
    a. The student's level of knowledge on the subject being tested, being, on average, higher the test scores the higher the level of expertise in the subject, as it is proved through the performance metric defined in section 4.5.
    b. The learning process and experience, because the environment and the way the subject is approached is very important for the student to retain the knowledge received.

Considering these conclusions, it is interesting to consider their implications regarding learning analytics research. In fact, while teachers may have no control over the students' performance in a given subject, they can still influence the learning experience and environment regarding the effectiveness of their teaching skills and the innovative teaching tools they perform. As the main purpose of learning analytics is analyze data about learners

for optimizing the learning process, teachers could try, by applying new methodologies in their teaching practices, improve the learning experience of their students in order to increase their academic achievements. With non-conventional approaches on some subject, teachers can gain students' attention and increase their grades. In addition, with these approaches, teachers motivate students to study more and to really understand the topic addressed, because nowadays the role of the teacher is to partner with the students.

Despite the identifiable limitations of the improved k-means clustering model in the NetLogo software, this dissertation still satisfies its final purpose, which is to analyze students' performance based on learning through this model, and in function of their knowledge regarding the algorithm. In fact, the improved model developed in NetLogo software has an innovative character when used to teach the procedure of the specific algorithm in a simulation environment. In addition, the analysis of the results obtained through the students' responses to the questionnaire provided interesting and coherent insights into the innovative theme defined as learning analytics. As such, the results obtained through the questionnaire about this new form of the teaching of the algorithm under study may prove useful for future research on the optimization of students' learning processes. There are a lot of potential future research paths to possibly follow, built on the results of the present dissertation:

1. The effectiveness of pedagogical design can be greatly improved through feedback from students using learning analytics.
2. Ensuring that pedagogical education is aligned with students' needs and their well-being can be analyzed through learning analytics.
3. Students can take responsibility for their learning by optimizing the learning process through learning analytics.
4. Evaluating the efficiency of a specific educational institution by measuring the impacts of optimizing learning processes on student performance with learning analytics.
5. There is a growing need for more and better metrics to measure students' knowledge about specific topics. This may be the closest future research to happen to the existing knowledge base about learning analytics today.

Summing up, we believe that this exploratory analysis helped us understand the extent to which the optimization of a specific learning process has an impact on students' performance.

Appendixes

1. Model Programming Code:

```
breed [datapoints datapoint]
breed [centroids centroid]

globals [any-centroids-moved?]

to setup
  clear-all
  set-default-shape datapoints "circle"
  set-default-shape centroids "x"
  generate-datapoints
  reset-centroids
end

to setup2
  clear-all
  set-default-shape datapoints "circle"
  set-default-shape centroids "x"
  generate-datapoints
  reset-centroids3
  reset-ticks
end

to generate-datapoints
  repeat num-datapoints [
    let center-x random-xcor / 1.5
    let center-y random-ycor / 1.5
    create-datapoints num-datapoints [
      setxy center-x center-y
      set heading random 360
       ]]
end
```

```
to reset-centroids
  set any-centroids-moved? true
  ask datapoints [ set color grey ]

  let colors base-colors
  ask centroids [die]
  create-centroids num-centroids [
    move-to one-of datapoints
    set size 5
    set color last colors + 1
    set colors butlast colors
  ]
  ;;clear-all-plots
  reset-ticks
end

to reset-centroids2
  set any-centroids-moved? true
  ask datapoints [ set color grey ]

  let colors base-colors
  ask centroids [die]
  create-centroids num-centroids [
    move-to one-of datapoints
    set size 5
    set color last colors + 1
    set colors butlast colors
  ]
end

to reset-centroids3
  set any-centroids-moved? true
  ask datapoints [ set color grey ]

  let colors base-colors
  ask centroids [die]
```

```
  create-centroids 1 [
    move-to one-of datapoints
    set size 5
    set color last colors + 1
    set colors butlast colors
  ]
end

to go
  if not any-centroids-moved? [stop]
  set any-centroids-moved? false
  assign-clusters
  update-clusters
  tick
end

to go2
  set num-centroids (ticks + 1)
  reset-centroids2
  while[any-centroids-moved?][
    set any-centroids-moved? false
    assign-clusters
    update-clusters2
  ]
  tick
end

to assign-clusters
  ask datapoints [set color ([color] of (closest-centroid) - 2)]
end

to update-clusters
  let movement-threshold 0.1
  ask centroids [
    let my-points datapoints with [ shade-of? color [ color ] of myself ]
    if any? my-points [
```

```
      let new-xcor mean [ xcor ] of my-points
      let new-ycor mean [ ycor ] of my-points
      if distancexy new-xcor new-ycor > movement-threshold [
        set any-centroids-moved? true
      ]
      setxy new-xcor new-ycor
    ]
  ]
  update-plots
end

to update-clusters2
  let movement-threshold 0.1
  ask centroids [
    let my-points datapoints with [ shade-of? color [ color ] of myself ]
    if any? my-points [
      let new-xcor mean [ xcor ] of my-points
      let new-ycor mean [ ycor ] of my-points
      if distancexy new-xcor new-ycor > movement-threshold [
        set any-centroids-moved? true
      ]
      setxy new-xcor new-ycor
    ]
  ]
end

to-report closest-centroid
  report min-one-of centroids [ distance myself ]
end

to-report square-deviation
  report sum [ (distance myself) ^ 2 ] of datapoints with [ closest-centroid = myself ]
end
to-report mean-square-deviation
  report mean [ square-deviation ] of centroids
end
```

2. Questionnaire:

Questions	**Possible Answers**
Q1. What is your nationality?	Open Answer
Q2. How old are you?	o 18-24 years old o 25-35 years old o >35 years old
Q3. What is your academic background? Or, if you are in your bachelor degree what area are you attending?	o Economics o Engineering o Medicine o Management o IT o Science o Arts o Social Communication o Languages o Marketing and Publicity o Other (need to specify)
Q4. What master's area are you currently attending or thinking about attending in the future?	o Data Analytics o Economics o Finance o Management o Marketing o Economics and Business Administration o Other (need to specify)
Q5. Classify your level of expertise about clustering?	Range from 0 to 100, where 0 means the students neither know the term of clustering and above 0 know, and within this type of knowledge, the students who answered to know above 50 have an excellent level of knowledge.

Q6. What is the main goal of a k-means clustering method?	○ The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. ○ K-means clustering is a type of a supervised learning, which is used when you have labeled data. ○ The algorithm assigns each data point to one of the firsts groups founded. Data points are clustered based on feature dissimilarity.
Q7. How the optimal number of clusters is reached?	○ The curve presented in the left chart began to decrease exponentially and then stabilized, and it is at this point of stabilization that we find the optimal number of clusters. ○ The curve presented in the left chart began to decrease exponentially and it is at this point that the decrease begins that we find the optimum number of clusters. ○ The curve presented in the left chart began to decrease exponentially and then stabilized, and it is at the end of the stabilization that we find the

	optimal number of clusters.
Q8. What happens when the algorithm finds the optimal number of clusters?	○ The convergence criterion is reached and the square error tends to decrease significantly. ○ The convergence criterion is reached and the square error remains. ○ The convergence criterion is reached and the square error no longer decreases significantly and there is no re-assignment of a pattern occurs from one cluster to another.
Q9. Did the interaction with the algorithm in NetLogo web help you understand what the k-means clustering method consists of? Classify from 1 to 5.	Range from 1 to 5, where: 1) It didn't help anything; 2) It helped a little bit; 3) It helped reasonably; 4) It helped a lot; 5) It helped perfectly.

References

Afonso, N. (2005), *Investigação Naturalista em Educação: Guia prático e crítico*. Porto: Asa Editores

Backer, E. & Jain, A. (1981), "A clustering performance measure based on fuzzy set decomposition". *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 3, nº 1, pp. 66–75.

Baker, B. (2007), *A conceptual framework for making knowledge actionable through capital formation*. United States, University of Maryland University College. D.Mgt. dissertation.

Baker, R., D'Mello, S., Rodrigo, M. & Graesser, A. (2010), "Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments." *International Journal of Human-Computer Studies*, vol. 68, nº4, pp. 223-241.

Baker, R. S., & Inventado, P. S. (2014), "Educational data mining and learning analytics.". *Learning analytics*, pp. 61-75. New York.

Bausmeister, R. F. (2014), "Self-regulation, ego depletion, and inhibition.", in *Neuropsychologia*. Florida: Florida State University. pp. 313-319.

Berk, R.A. (2005) "Survey of 12 Strategies to Measure Teaching Efectiveness." *International Journal of Teaching and Learning in Higher Education*, vol. 17, nº1, pp. 48-62.

Bholowalia, P., & Kumar, A. (2014), "EBK-means: A clustering technique based on elbow method and k-means in WSN." *International Journal of Computer Applications*, vol. 105, nº9.

Bloom, K. L. (2006), "Age and the self-concept." *American Journal of Psychiatry*, vol. 118, nº6, pp. 534–538.

Bonabeau, E. (2002), "Agent-based modeling: Methods and techniques for simulatin human systems.". *Proceedings of the National Academy of Sciences*, vol. 99, nº 3, pp.7280-7287.

Bradley, P. S., & Fayyad, U. M. (1998), "Refining Initial Points for K-Means Clustering". *International Conference on Machine Learning.* , vol. 98, pp. 91-99.

Cavalcanti, M. J. A. (2003), Aprendizagem & auto-estima. Available at http://docslide.com.org/documents/aprendizagem-e-auto-estima.html. Accessed on 10.07.2018.

CommLab (2010), How to measure the efectiveness of e-learning courses. Available at https://www.commlabindia.com/resources/article/e-learning-programs.php. Accessed on 25.06.2018.

Desmarais, M. & Baker, R. (2012), "A review of recent advances in learner and skill modeling in intelligent learning environments." *User Model. User-Adapt. Interact*, pp. 9 38.

Dron, J. & Anderson, T. (2009), "On the design of collective applications." *Proceedings of the 2009 International Conference on Computational Science and Engineering*, vol. 4, pp. 368-374. Canada. August 29-31.

Elias, T. (2011), "Learning analytics: Definitions, processes, and potential." *Unpublished paper.* Available at: http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

Epstein J. M. & Axtell R. L. (1996), "Growing Artificial Societies: Social Science from the Bottom Up". *MIT Press.* Cambridge.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010), "The ODD protocol: a review and first update." *Ecological modelling*, vol. 221, nº23, pp.2760-2768.

Hamerly, G., & Elkan, C. (2004), "Learning the k in k-means.", *Advances in neural information processing systems*, pp. 281-288.

Hendricks, M., Plantz, M. C., & Pritchard, K. J. (2008), "Measuring outcomes of United Way–funded programs: Expectations and reality.", in J. G. Carman & K. A. Fredericks (editors), *Nonprofits and evaluation. New Directions for Evaluation*, vol. 119, pp. 13–35.

Hjorth, A., Head, B. and Wilensky, U. (2014). *NetLogo K-Means Clustering model.* Available at http://ccl.northwestern.edu/netlogo/models/K-MeansClustering. Accessed on 12.05.2018.

Johnson, L., R. Smith, H. Willis, A. Levine, Haywood, K. (2011). *Learning Analytics. The 2011 Horizon Report.* Austin, Texas: The New Media Consortium. Available at https://www.nmc.org/pdf/2011-Horizon-Report.pdf. Accessed on 6.12.2017.

Kodinariya, T. M., & Makwana, P. R. (2013), "Review on determining number of Cluster in K-Means Clustering." *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, nº6, pp. 90-95.

IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

MacQueen, J. B. (1967), "Some Methods for classification and Analysis of Multivariate Observations." *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297. 1University of California Press. pp. 281–297. Berkeley. January 7.

Oblinger, D. G. & Campbell, J. P. (2007). *Academic Analytics.* Available at www.educause.edu/ir/library/pdf/PUB6101.pdf. Accesed on 23.12.2017.

Romero, C. & Ventura, S. (2013), "Data Mining in Education." *WIREs Data Mining and Knowledge Discovery*, vol. 3, pp. 12-27.

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010), "Application of k-Means Clustering algorithm for prediction of Students Academic Performance." *International Journal of Computer Science and Information Security*, vol. 7, nº1, pp. 292-295.

Pelánek, R. (2015), "Metrics for Evaluation of Student Models." *Journal of Educational Data Mining*, vol. 7, nº2, pp. 1-19.

Proceedings of the 1st International Conference on Learning Analytics and Knowledge. (2011) ACM, New York, NY, USA. February 27 – March 1.

Richardson, J. T. (2004), "Methodological issues in questionnaire-based research on student learning in higher education", *Educational psychology review*, vol. 16, nº4, pp. 347-358.

Rogers, P. C., McEwen, M.R. & Pond, S. (2008), "The design and evaluation of distance education." in Veletsianos, *G.* (editor) *Emerging technologies in distance education*, pp. 231-247.

Sakellariou, I., Kefalas, P., & Stamatopoulou, I. (2008), "Teaching intelligent agents using NetLogo. "*ACM-IFIP IEEIII*, pp. 209-221.

Selim, S. Z., & Ismail, M. A. (1984), "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality." *IEEE Transactions on pattern analysis and machine intelligence,* vol. 1, pp. 81-87.

Siemens, G. & Baker, R. (2010), "Learning analytics and educational data mining: Towards communication and collaboration." *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 252–254. Canada. April 29 – May 2.

Siemens, G. & Long, P. (2011), "Penetrating the Fog: Analytics in Learning and Education."*Educause review*, vol. 46, nº5.

Shavelson, R. J., Hubner, J. J., & Stanton, G. (1976), "Selfconcept: Validation of construct interpretations." *Review of Educational Research*, vol. 41, pp-46-407.

Shulman, L. (1987), "Knowledge and Teaching: Foundations of the New Reform." *Harvard Educational Review*, vol. 57, nº1, pp. 1–23

Steinbach, M., Karypis, G., & Kumar, V. (2000), "A comparison of document clustering techniques." *KDD workshop on text mining*, vol. 400, nº 1, pp. 525-526.

Tisue, S., & Wilensky, U. (2004), "Netlogo: A simple environment for modeling complexity. "*, International conference on complex systems*, vol. 21, pp. 16-21.

Tisue, S., & Wilensky, U. (2004), "NetLogo: Design and implementation of a multi-agent modeling environment." *Proceedings of the Agent 2004 Conference on Social Dynamics: Interaction, Reflexivity and Emergence*. Chicago.

Trindade, A. R. (2005), "A eficácia do ensino: indicadores, métodos e instrumentos."

Vaz Serra, A. (1988), "O auto-conceito." *Análise Psicológica*, vol. 4, nº2, pp. 101-110.

Wilensky, U. (1999), NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Scheuer, O., & McLaren, B. M. (2012), "Educational data mining.". *Encyclopedia of the Sciences of Learning*, pp. 1075-1079). Boston.

Xu, R., & Wunsch, D. (2005), "Survey of clustering algorithms." *IEEE Transactions on neural networks*, vol. 16, nº3, pp. 645-678.