UNIVERSIDADE ABERTA

DEPARTAMENTO de CIÊNCIAS e TECNOLOGIA

Educational Data Mining e *Learning Analytics* na melhoria do ensino *online*

Dissertação para obtenção do Grau de Mestre em

Estatística, Matemática e Computação

Especialização – Estatística Computacional

Susana Maria Sousa Martins Leite de Faria

Orientador: Professor Doutor Angel Juan

Co-orientador: Professor Doutor Amílcar Manuel do Rosário Oliveira

fevereiro 2014

Agradecimentos

Será sempre pouco o espaço e o tempo para agradecer àqueles cujo contributo tornou possível a concretização deste trabalho, que é simultaneamente um projeto de vida pessoal. Desta forma, gostaria de expressar a minha gratidão a todos os que deram o seu apoio, uns mais presentes que outros, porém, todos se revelaram valiosos para a execução desta tese.

Tentarei não pecar por omissão, mas gostaria de deixar alguns agradecimentos particulares, nomeadamente ao Professor Doutor Angel Juan e ao Professor Doutor Amílcar Manuel do Rosário Oliveira, que contribuíram para uma experiência gratificante, dando-me a liberdade intelectual para elaborar o meu trabalho, mas exigindo sempre uma elevada qualidade em todos os desenvolvimentos. Agradeço-lhes as palavras de incentivo e otimismo, mesmo quando nada parecia correr bem. Não me deixaram "sair" do caminho, e tiveram a generosidade de me guiar com sugestões e comentários que contribuíram decisivamente para que fosse possível atingir os objetivos desta dissertação.

E, por último, mas igualmente fundamentais, agradeço ao meu marido, familiares e amigos, que suportaram este longo processo comigo, apoiando-me sempre com compreensão e motivação extra!

A todos eles, que sempre me incentivaram, obrigada por tudo!

"Dai-me uma alavanca e um ponto de apoio e moverei o mundo."

(Arquimedes)

Resumo

Educação é um dos temas mais importantes e discutidos em todo o mundo. Sendo um processo de aquisição de conhecimento e/ou aptidões, tem sofrido grandes alterações ao longo dos tempos.

Na última década, os avanços das tecnologias de informação e computação têm permitido às pessoas interagirem e aprenderem de uma nova forma.

Com as inovações tecnológicas, as escolas e universidades estão a alterar a forma como transmitem e partilham conhecimentos. Ao passo que, até ao ensino secundário, as Escolas disponibilizam uma plataforma *Moodle*, onde os professores divulgam e partilham alguns documentos e tarefas que servem de apoio às suas práticas letivas; já no ensino superior as alterações são mais significativas. As Universidades chegam mesmo a alterar a metodologia dos seus cursos. Para além do ensino tradicional optam por outras modalidades de ensino: *b-learning* (ensino simultaneamente presencial e à distância) e/ou *e-learning* (ensino à distância).

Os modelos de ensino/aprendizagem assentes em ambientes *online* permitem aos alunos terem acesso ao conhecimento a qualquer hora e em qualquer lugar. No entanto, também têm os seus desafios, devido à ausência de contacto humano direto e às insuficiências que isso pode gerar. Contudo, os propugnadores do *e-learning* defendem que a criação de comunidades virtuais que interagem através de *chats*, fóruns, *emails*, etc, compensam essa carência, enriquecendo o processo relacional entre pessoas com o mesmo interesse, mas com diferentes visões e localizadas em regiões e países distintos.

Com o aumento do uso de ambientes *online* e outras tecnologias para apoio ao processo de ensino e aprendizagem, um grande volume de dados tem sido gerado a partir das diferentes

interações no sistema, envolvendo estudantes e professores. Com a análise destes dados, podemos obter uma quantidade de informação e conhecimento pertinente e essencial para a melhoria da qualidade do ensino. Nomeadamente, no combate ao insucesso e ao abandono escolar. Diversos estudos têm sido promovidos e desenvolvidos de modo a identificarem e analisarem as causas do insucesso escolar. Como consequência têm sido desenvolvidos programas e medidas que visam a promoção do sucesso. Uma das medidas consiste no acompanhamento adequado e personalizado dos alunos ao longo do seu percurso académico.

Neste trabalho é proposto um modelo de análise de dados, com base em cartas de controlo, regressão logística e análise de *clusters*, com vista à extração de conhecimento, relevante na previsão do desempenho escolar, no ensino *online*.

Abstract

Education is one of the most important and widely discussed subjects all over the world. Provided that it is a process of knowledge and/or skills acquisition, education has undergone many changes over time.

Over the last decade, improvements in information and computer technologies have enabled people to interact and learn in a different way.

Due to technological advance, schools and universities have been changing the way they transmit and share knowledge. Whereas up to high school education schools provide a Moodle platform, in which teachers spread and share some documents and tasks that support their classroom practices, in higher education the technological changes are more significant. Universities even change the methodology of their courses. Besides the traditional way of teaching, they choose other types of education: b-learning (both classroom and online learning) and/or e-learning (online learning).

The teaching/learning models based on online environments allow students to have access to knowledge at anytime and anywhere. However, it also has its challenges due to the absence of direct human contact and the insufficiencies this might create. Nevertheless, the proponents of e-learning argue that the creation of virtual communities that interact through chat rooms, forums, email, etc, surpasses that absence as they enrich the relational process among people who share the same interest but have different views and may be located in different regions and countries.

With the increasing use of online environments and other technologies supporting the educational process, a large amount of data has been generated from different interactions in the system involving students and teachers. From the analysis of these data it is possible to get

a considerable and relevant amount of information and knowledge which are essential for improving the quality of teaching, particularly as regards the prevention of school failure and dropout. Several studies have been promoted and developed in order to identify and analyse the causes of school failure. Consequently, some programs and measures aimed at reaching school success have been developed. One of them consists of appropriate and personalized support for students throughout their academic career path.

In this work it is proposed a model of data analysis based on control charts, logistic regression, and cluster analysis in order to extract relevant knowledge for the prediction of school performance on the online teaching.

Simbologia e Notações

ANOVA – *Analysis of Variance*;

AIED – *Artificial Intelligence in Education;*

AVA – *Ambiente Virtual de Aprendizagem;*

DT – *Data Mining*;

EDM – *Educational Data Mining*;

ITS – *Inteligent Tutoring System;*

KDD – Knowledge Discovery in Databases;

LA – *Learning Analytics*;

LC – Linha central;

LCI – Limites do controlo inferior;

LCS – Limites de controlo superior;

LMS – Sistema de Gestão de Aprendizagem;

MMV – Método de Máxima Verosimilhança;

PLE – Ambiente de Aprendizagem Pessoal;

ROC – *Receiver Operating Characteristic*;

SPSS – *Statistical Package for the Social Sciences.*

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

1 Introdução

Neste capítulo introdutório será realizada uma apresentação do enquadramento e motivação do tema abordado ao longo deste trabalho, bem como os objetivos e a sua organização.

1.1 Motivação

Na conceção tradicional de Educação, o estudante chega à escola com a "cabeça vazia" e compete ao professor transmitir-lhe conhecimentos, para que mais tarde seja submetido a provas e exames, de modo a avançar para uma nova fase. Portanto, o professor desempenha o papel principal neste modelo de educação.

Com os avanços das tecnologias de informação e comunicação, deixamos de ter um modelo de educação centrado no professor e passamos a ter o estudante como ator ativo e central no seu processo de aprendizagem. Desta forma, os estudantes passam a construir os seus conhecimentos segundo os seus estilos de aprendizagem, utilizando para isso, sistemas interativos com apoio tecnológico, onde a motivação para a aprendizagem surge no estudante. Cabendo à escola/universidade dotá-lo de capacidades que lhe permita no seu futuro

profissional aprender qualquer assunto que lhe interesse. Portanto, o professor passa a desempenhar o papel de guia, conselheiro e parceiro do aluno na procura de informação e da verdade.

Antes da chegada da Internet, surgiram vários tipos de formação com o advento das novas tecnologias. Apareceram cursos ministrados pela "Tele-Escola", onde os estudantes assistiam às aulas pela televisão e no final realizavam questionários elaborados pelos seus profissionais, para testarem o conhecimento adquirido. Este método permitia que em escolas sem profissionais em educação ou com falta deles, os estudantes pudessem ter acesso a uma educação credível sem saírem de suas casas. No entanto, tinha as suas desvantagens, tais como: falta de interatividade, rigidez dos horários e monotonia.

Posteriormente, surgiram cursos de formação à distância, onde os estudantes tinham manuais, cassetes e vídeos gravados fornecidos pelos centros de formação. Os estudantes quando tinham dúvidas contactavam o professor pelo telefone ou correio. Este método tinha os seus inconvenientes, pois os estudantes nunca viam as suas dúvidas devidamente esclarecidas e em tempo útil.

Entretanto, surgiram programas de multimédia interativos com fins educativos. Por exemplo, dicionários, enciclopédias, manuais escolares, etc.

Com a chegada da Internet, muitas instituições viram um excelente meio de fornecer formação sem grandes investimentos, com a vantagem de não estar restringido a uma dada localização geográfica, nem a um horário definido e nem à existência de deficiência física. Possibilitou a pessoas de todo o mundo, terem acesso à formação, tendo apenas ligação à Internet. Várias foram as instituições (universidades, centros de formação profissional), que aproveitaram esta vaga para lançarem mãos à obra investindo, seriamente, neste tipo de formação. Surgem assim, algumas universidades dedicadas unicamente ao ensino *online,* e algumas das universidades que, para além do ensino tradicional, enriquecem a sua oferta pedagógica com outras modalidades de ensino, o ensino *b-learning* (ensino simultaneamente presencial e à distância) e o ensino *e-learning* (ensino à distância).

Enquanto que as conversas de corredor e as aulas expositivas tendem a evaporar-se assim que terminam, cada clique nas redes sociais para atualização de dados, cada interação social e cada página lida *online* pode deixar uma pegada digital.

A aprendizagem *online* cria trilhos de dados dos estudantes que podem fornecer informações valiosas sobre como está a decorrer o processo de aprendizagem, nomeadamente informações sobre os alunos que, estão em risco de abandono escolar ou precisam de atividades diferenciadas para aumentar o seu sucesso e confiança. Para além deste tipo de

informações, cada vez mais as instituições de ensino estão sob uma crescente pressão para reduzir custos e aumentar a qualidade. Desta forma, a extração de dados / análise promete ser uma lente importante, contribuindo para o sucesso de todas as partes intervenientes do processo de ensino.

Educational Data Mining e Learning *Analytics* são duas áreas antigas e novas ao mesmo tempo. Antigas, porque tratam de um problema que existe desde os tempos de Platão: como melhorar a forma de aprendizagem dos alunos. Novas, porque estas ferramentas utilizadas para alcançar este objetivo existem apenas há 10 anos.

Vários investigadores têm-se dedicado a explorar técnicas semelhantes em *Educational Data Mining* e *Learning Analytics* na área da educação para melhorar o ensino *online*. Os estudantes poderão usufruir destas melhorias bem como os professores e administradores das universidades.

Ao passo que a *Educational Data Mining* tem-se centrado em técnicas e tecnologias, a *Learning Analytics* tem sido prática e focada na aplicação.

Independentemente das diferenças entre *Learning Analytics* e *Educational Data Mining*, as duas áreas têm uma sobreposição significativa, tanto nos objetivos do investigador como nos métodos e técnicas que são utilizados na investigação.

Learning Analytics baseia-se numa gama mais ampla de disciplinas académicas de ensino de mineração de dados, incorporando conceitos e técnicas da ciência da informação e sociologia, além da ciência da computação, estatística, psicologia e ciências da aprendizagem. Ao contrário de *Educational Data Mining*, *Learning Analytics*, geralmente, não enfatiza a redução do aprendizado em componentes mas procura compreender sistemas inteiros e apoiar a tomada de decisão humana em vez de buscar respostas automáticas.

Neste trabalho, pretende-se aplicar técnicas e métodos de *Educational Data Mining* e *Learning Analytics* para a previsão do desempenho escolar no ensino *online*.

O modelo proposto pretende prever o desempenho escolar dos estudantes, através de um conjunto de variáveis de entrada. Com este estudo pretende-se ajudar a comunidade educativa a melhorar a gestão de recursos e a aplicação de estratégias, por forma a melhorar o desempenho dos estudantes.

1.2 Objetivos

O principal objetivo deste trabalho consiste na identificação de padrões úteis, com vista à extração de conhecimento sobre o desempenho escolar a partir da base de dados sobre

estudantes do ensino *online*. O objetivo é tentar prever o desempenho escolar e, se possível, alertar os professores para a necessidade de travar o abandono escolar. Para o conseguir, recorrer-se-á a técnicas de *Educational Data Mining*.

1.3 Organização da Dissertação

A dissertação encontra-se dividida em duas partes fundamentais, isto é, uma visão geral de toda a fundamentação teórica que serviu de apoio à realização do trabalho prático (capítulos 2 e 3) e o trabalho prático utilizado para aplicação dos objetivos inicialmente propostos, sendo também analisados e discutidos os resultados obtidos (capítulo 4).

Esta dissertação está organizada em quatro capítulos (excluindo o capítulo introdutório) e possui 3 anexos.

Capítulo 2 – *Educational Data Mining* e *Learning Analytics*

Neste capítulo, fornece-se uma visão geral sobre *Educational Data Mining* e *Learning Analytics*.

Capítulo 3 – Técnicas de *Educational Data Mining*

Neste capítulo, apresenta-se uma visão pormenorizada de algumas técnicas de *Educational Data Mining*.

Capítulo 4 – Um modelo de análise de dados com base em Análise de *Clusters*, Cartas de Controlo e Regressão Logística para monitorizar o ensino *online*.

Neste capítulo, fornece-se o modelo proposto, acrescentando à ferramenta de *Education Data Mining* – SAMOS – estas três técnicas mostrando ser uma mais-valia para o modelo inicial.

Capítulo 5 – Considerações Finais e Perspetivas Futuras

Pretende-se, com este capítulo, apresentar uma síntese do trabalho realizado, sendo discutidas as conclusões mais importantes relativas ao trabalho desenvolvido. O capítulo termina com a apresentação das contribuições e recomendações para trabalhos futuros.

No anexo A, é apresentado o questionário realizado.

No anexo B, é exibida uma análise estatística da base de dados.

No anexo C, é aduzida uma análise descritiva dos dados (C1), a análise de *clusters* (C2) e a regressão logística binária (C3).

Capítulo 2

2 *Educational Data Mining* e *Learning Analytics*

Já sabemos que são cada vez mais usados dados para a tomada de decisão; as empresas transformam os dados, guardados nos seus sistemas, em informações importantes para a tomada de decisão através da tecnologia *Business Intelligence*. Esta tecnologia pode discernir padrões e tendências históricas de dados e pode criar modelos que preveem tendências e padrões futuros.

Um dos primeiros exemplos do uso de dados para estudar o comportamento *online* foi *Web Analytics*. No entanto, as empresas já desenvolveram técnicas mais sofisticadas para rastrear as interações das pessoas com os seus *sites*. Com esta monitorização, surgiram mudanças significativas; emergiram, por exemplo, os *e-books* com mais passagens e as páginas *web* mais prováveis de interesse. Relativamente às redes sociais *Twitter*, *Facebook* ou *blogs*, os comentários postados podem ser rastreados e analisados.

Para analisar estes novos eventos, precisamos de novas técnicas, nomeadamente, para trabalhar com textos não estruturados e com os dados em forma de imagem, dados de várias fontes, e com grandes quantidades de dados.

Os dados podem ser colocados de forma estruturada, sendo fáceis de trabalhar, e de forma não estruturada, em que o computador tem dificuldade em trabalhá-los sem ajuda humana. Um exemplo simples é uma mensagem de *email* que é composta por partes

estruturadas (de, para e a data de envio) e algumas partes não estruturadas (o assunto e o corpo da mensagem).

Com o aumento da pesquisa em *Machine Learning* surgiram técnicas fundamentais na descoberta do conhecimento e em *Data Mining (DM)*, pois emergem novas e potentes informações que conseguem extrair-se das grandes quantidades de dados não estruturados. Com estas técnicas, descobrem-se padrões nos dados e, em seguida, constroem-se modelos que preveem um resultado.

Educational Data Mining (EDM) e *Learning Analytics* (LA) são duas áreas em desenvolvimento, que trabalham especificamente com grandes volumes de dados em educação. Embora não haja nenhuma distinção dura e rápida entre estes dois campos, eles tiveram diferentes histórias de pesquisa e de desenvolvimento, daí serem considerados como duas áreas de pesquisa distintas.

Em geral, *Educational Data Mining* procura novos padrões nos dados e desenvolve novos algoritmos e/ou novos modelos; enquanto que, *Learning Analytics* aplica modelos preditivos já conhecidos nos sistemas de ensino e aprendizagem. Seguidamente ambas as áreas serão apresentadas detalhadamente.

2.1 Educational Data Mining

A área emergente de *Educational Data Mining* procura desenvolver ou adaptar métodos e algoritmos de mineração já existentes, de modo a ajudar a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem. Com estes métodos pretende-se, por exemplo, entender melhor o estudante no seu processo de aprendizagem, analisando-se a sua interação com o ambiente (Baker et al., 2012).

Educational Data Mining pode ser esboçada como a combinação de três áreas principais (ver Figura 1): ciências da computação, educação e estatística. A intersecção dessas três áreas também torna outras subáreas intimamente relacionadas com *Educational Data Mining*, como: a educação baseada em computador, *Data* Mining e Machine *learning* e *Learning Analytics* (Romero e Ventura, 2013).

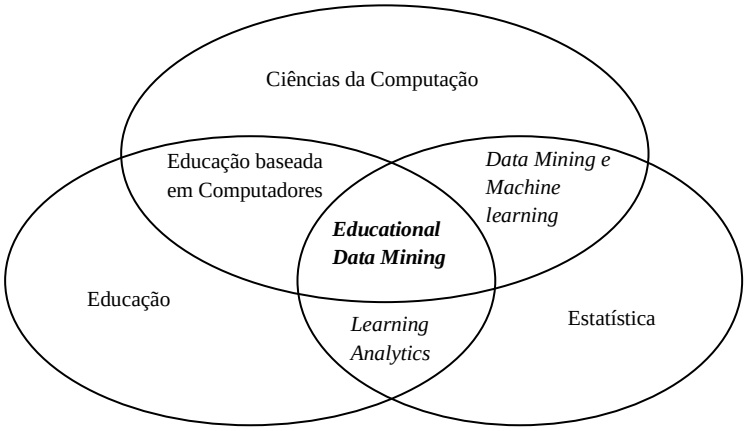

Figura 1. Principais áreas relacionadas com a mineração de dados educacionais (adaptado de [Romero e Ventura, 2013])

De todas as áreas acima mencionadas (ver Figura 1), o campo mais relacionado com *Educational Data Mining* é *Learning Analytics* (Romero e Ventura, 2013).

2.1.1 Origens de *Educational Data Mining*

Após algumas iniciativas em participar em *workshops* específicos dentro das Conferências sobre *Artificial Intelligence in Education* (AIED) e *Inteligent Tutoring System* (ITS), foi apenas em 2005, em Pittsburgh, EUA, que foi organizado o primeiro *Workshop* em *Educational Data Mining* como parte integrante da *20th National Conference on Artificial Intelligence*. Entre 2006 e 2007, realizou-se várias vezes este *workshop*. Em 2006, realizaram-se dois, um em Jhongli, Taiwan e o outro, em Boston, EUA. Em 2007, realizaram-se três, em Niigata, Japão, na Califórnia, EUA e em Crete, Grécia. O último *workshop* surgiu com o nome *Workshop on Applying Data Mining in e-Learning* (Romero e Ventura, 2013).

Em 2008, em Montreal, Canadá, realizou-se a primeira conferência em EDM: *First International Conference on Educational Data Mining*. Devido ao impacto que este evento teve, ganhou regularidade de realização anual, estando agora em 2013 na sua sexta edição. A 1ª conferência (2008) realizou-se em Montréal, Québec, Canadá; a 2ª conferência (2009) realizou-se em Cordoba, Espanha; a 3ª conferência (2010) realizou-se em Pittsburgh, PA, EUA; a 4ª conferência (2011) realizou-se em Eindhoven, Holanda; a 5ª conferência (2012) realizou-se em Chania, Grécia e a última realizada até ao momento, a 6ª, teve lugar em Memphis, Tennessue, EUA.

Existem várias sociedades internacionais sobre *Educational Data Mining*. Em 2011, emerge a sociedade cientista para EDM, *International Educational Data Mining Society* (http://www.educationaldatamining.org) e a *Society for Learning Analytics Research* (SoLAR) (http://www.solaresearch.org/). Em 2012, surge *IEEE Task Force of Educational Data Mining* (EDM-TF) (http://datamining.it.uts.edu.au/edd/).

2.1.2 Principais publicações de *Educational Data Mining*

Atualmente, apenas foram publicados dois livros sobre *Educational Data Mining*. O primeiro, intitulado *Data Mining in e-learning*, com 17 capítulos voltados para ambientes educacionais baseados na *web*. O segundo, intitulado *Handbook of Educational Data Mining*, com 36 capítulos voltados para diferentes tipos de ambientes educacionais (Romero e Ventura, 2013).

Há várias pesquisas em periódicos e capítulos de livros sobre *Educational Data Mining*. A primeira e mais popular revisão de pesquisas em *Educational Data Mining* foi apresentada por Romero e Ventura, no ano 2007, no Jornal *Expert Systems with Applications*, volume 33 com o título: *Educational data mining: a survey from 1995 to 2005*. De seguida, Baker e Yacef, no ano 2009, os mesmos autores lançaram a mais completa revisão teórica de EDM, no Jornal *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications & Reviews 2010*, volume 40, intitulada *Educational Data Mining: a review of the state-of-the-art* (Romero e Ventura, 2013).

A primeira e mais ampla revisão que surgiu num capítulo de um livro, orientada para a aplicação de *Data Mining* em *e-learning*, foi escrita por Castro, Vellido, Nebort & Mugica, em 2007, com o título *Applying data mining techniques to e-learning problems*, no livro *Evolution of Teaching and Learning Paradigms in Intelligent Environment*, volume 62 da *Springer-Verlag* (Romero e Ventura, 2013).

Um segundo capítulo, surgiu num formato mais curto e mais orientado para *Intelligent Tutoring System* (ITS) foi escrito por Baker, em 2010, com o título *Data mining for education*, no livro *International Encyclopedia of Education*, 3ª edição, volume 7 da Oxford.
Um terceiro capítulo de um livro, considerado o mais genérico mas o mais curto, foi escrito por Scheuer e McLaren, em 2011, com o título *Educational Data Mining*, no livro *The Encyclopedia of the Sciences of Learning* da Springer.

Finalmente, foi publicado um relatório, sobre como melhorar o ensino e a aprendizagem através da *Educational Data Mining* e *Learning Analytics*, escrito por Bienkowski, Feng e

Means, em 2012, com o título *Enhancing teaching and learning through educational data mining and learning analytics: na issue brief*, do Departamento da Tecnologia Educacional dos EUA.

Há uma grande variedade de revistas internacionais e de prestígio onde foram publicados vários trabalhos sobre EDM. Segundo Romero e Ventura (2013), os que tiveram mais impacto foram: *Journal of Educational Data Mining*, *Journal of Artificial Intelligence in Education*, *Journal of the Learning Sciences*, *Computer and Education*, *IEEE Transactions on Learning Technologies*, *IEEE Transactions on Knowledge and Data Engineering*, *ACM Special Interest Group on Knowledge Discovery and Data Mining*, *Explorations*, *User Modeling and User-adapted Interaction*, *Internet and Higher Education*, *Decision Support Systems*, *Expert Systems with Applications* e *Knowledge-Based Systems.*

De todos eles, o mais específico é o *Journal of Educational Data Mining* - http://www.educationaldatamining.org/JEDM/ - que foi lançado em 2009 como um jornal *online* e gratuito.

Por outro lado, os 10 artigos, sobre *Educational Data Mining*, mais citados, no Google Académico, são: *Educational data mining a survey from 1995 to 2005* (com 408 citações), *Data mining in course management systems: moodle case study and tutorial* (com 267 citações), *Web usage mining for a better Web-based learning environment* (com 206 citações), *Off-task behavior in the cognitive tutor classroom: when students game the system* (com 203 citações), *Building a recommender agent for e-learning systems* (com 202 citações), *Detecting student misuse of intelligent tutoring systems* (com 184 citações), *The ecological approach to the design of e-learning environments: purpose-based capture and use of information about learners* (com 153 citações), *Student modeling and machine learning* (com 142 citações), *Smart recommendation for an evolving e-learning system: architecture and experiment* (com 129 citações) e *Towards evaluating learners´ behavior in a Web-based distance learning environment* (com 126 citações) (Adaptado de [Romero e Ventura, 2013]).

2.1.3 Principais investigadores de *Educational Data Mining*

Há um grande número de investigadores importantes na área de *Educational Data Mining*. Ryan Baker, do Instituto Politécnico de Worcester, EUA, e presidente da sociedade de EDM. Kalina Yacef, da Universidade de Sydney, Austrália, e editor da revista JEDM e membro do comité da direção da sociedade EDM, junto com Tiffany Barnes da Universidade

da Carolina do Norte, EUA; Joseph E. Beck do Instituto Politécnico Worcester, EUA; Cristobal Romero e Sebastian Ventura, Professores da Universidade de Cordoba, Espanha; Michel Desmarais, da Escola Politécnica de Montreal, Canadá; Neil Heffernan, do Instituto Politécnico de Worcester, EUA; Agathe Merceron, da Beuth Universidade de Ciências Aplicadas, na Alemanha e Mykola Pechenizkiy da Universidade de Tecnologia de Eindhoven, na Holanda.

Outros autores relevantes são: Osmar Zaiaine da Universidade Alberta do Canadá; John Stamper, Kenneth Koedinger e Jack Mostow da Universidade Carnegie Mellon dos EUA; Judy Kay da Universidade de Sydney na Austrália; Rafi Nachmias da Universidade de Tel Aviv em Israel; Gord McCalla da Universidade de Saskatchewan no Canadá; Arthur Graesser da Universidade de Hemphis dos EUA, entre outros.

2.1.4 Processo de Descoberta de Conhecimento Educacional

De acordo com Romero et al. (2010), citado por Romero e Ventura (2013), o processo de aplicação de *Data Mining* para sistemas de ensino pode ser interpretado segundo diferentes pontos de vista.

Por um lado, segundo um ponto de vista educacional e experimental, pode ser visto como um ciclo iterativo de formação de hipóteses, testes e aperfeiçoamento (ver figura 2).

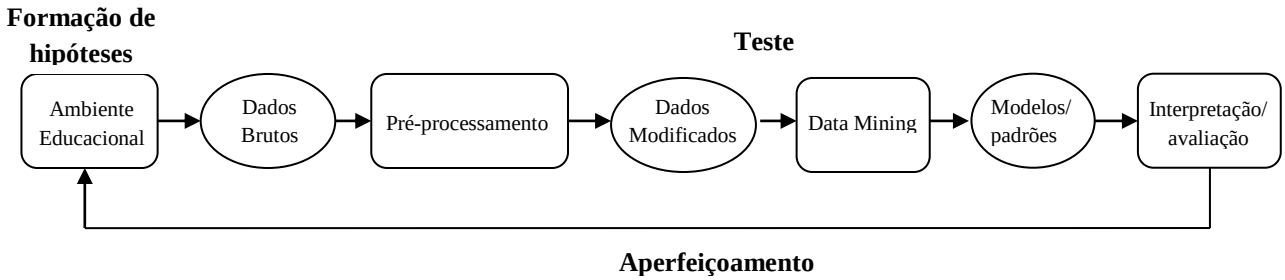

Figura 2. Descoberta de conhecimento educacional e processo de *data mining*. (adaptado de [Romero e Ventura, 2013])

Neste processo, o objetivo não é apenas transformar dados em conhecimento, mas também filtrar o conhecimento extraído para a tomada de decisões sobre como modificar o ambiente educacional, de modo a melhorar a aprendizagem do aluno. Este é um tipo de avaliação formativa de um programa educativo, enquanto ele ainda está em desenvolvimento, e com a finalidade de melhorar continuamente o programa. Analisar a forma como os alunos

utilizam o sistema é uma maneira de avaliar o *design* institucional de uma forma formativa e pode ajudar os *designers* educacionais a melhorar os materiais institucionais. Por exemplo, as técnicas de *Educational Data Mining* são usadas para descobrir modelos/ padrões que, por sua vez, podem ser usados para ajudar os *designers* educacionais a estabelecer uma base pedagógica para uma tomada de decisões e para projetar ou modificar a abordagem pedagógica de um ambiente educacional *online*.

Por outro lado, a partir de um ponto de vista da *Data Mining*, que se assemelha à descoberta de conhecimento e ao processo de *Data Mining*, embora haja diferenças importantes e características específicas em cada passo, como é descrito mas subsecções seguintes.

2.1.4.1 Ambiente Educacional

De acordo com Romero et al. (2010), citado por Romero e Ventura (2013), dependendo do tipo de ambiente educacional (educação tradicional em sala de aula, educação baseada no computador ou na *Web*) e do sistema de informação que o suporta (gestão de aprendizagem, *intelligent tutoring* ou *adaptive hypermedia system*) diferentes tipos de dados podem ser recolhidos para resolver diferentes problemas educacionais. Todos estes dados podem ter diversas fontes, incluindo dados administrativos, questionários, medições recolhidas de experiências controladas, entre outras. Reunir e integrar esses dados brutos para a *Data Mining* são tarefas que fazem parte da etapa seguinte: o pré-processamento.

2.1.4.2 Pré-processamento

De acordo com Bienkowski et al. (2012), em contextos educativos, o pré-processamento de dados é uma tarefa muito importante e complicada e que ocupa, por vezes, mais de metade do tempo total despendido na resolução de um problema de *Data Mining*.

Primeiro, os dados educacionais disponíveis (crus, originais e primários) para resolver um problema não estão sob a forma adequada. Segundo, devido à natureza heterogénea e hierárquica dos dados, a determinação do formato dos dados é uma tarefa a ter em conta, pois a melhor organização dos dados depende do tipo de problema a ser resolvido. Assim, é necessário converter os dados para um formato adequado (dados modificados) para resolver

um problema educacional específico. Esta tarefa inclui a escolha dos dados a recolher, pensando sempre no tipo de questões que pretendemos responder e ter a certeza que os dados estão de acordo com as perguntas. Por outro lado, os ambientes educacionais podem armazenar uma enorme quantidade de dados possíveis a partir de várias fontes com diferentes formatos e com diferentes níveis de "granularidade" (do grão grosso para grão fino) ou múltiplos níveis de hierarquia que fornecem mais ou menos dados. Normalmente, também está disponível um grande número de variáveis/atributos com informações sobre cada aluno, que poderão ser integrados numa tabela para melhor se poderem analisar. Questões de tempo, sequência e o contexto também desempenham papéis importantes no estudo de dados educacionais. O tempo é importante para termos dados como a indicação do tempo em que os estudantes treinaram ou simplesmente do tempo que precisaram para aprender uma determinada matéria. O contexto é importante para explicar os resultados e saber onde o modelo pode ou não ser aplicado.

Finalmente, é importante manter e proteger a confidencialidade das informações dos estudantes, suprimindo algumas informações pessoais, tais como nome, *email*, número de telefone, etc. Desta forma, os dados utilizados passam a ser anónimos e, por exemplo, podemos usar uma sequência numérica para identificar os estudantes.

2.1.4.3 *Data Mining*

De acordo com Cabena et al. (1998), citado por Baker et al. (2012), *Data Mining* é uma área disciplinar, usando principalmente conhecimentos de análise estatística de dados, *Machine Learning*, reconhecimento de padrões e visualização de dados.

Segundo Romero e Ventura (2013), a maioria das técnicas tradicionais de *Data Mining* – *classification*, *clustering*, *relationship mining*, entre outras – já foram aplicadas com sucesso na educação, por Baker RSJd em *Data Mining for Education* (2010). Segundo os autores Romero et al. (2010), a técnica de *Data Mining* mais útil e mais antiga na educação é a *classification*. No entanto, os sistemas educativos têm características especiais que exigem um tratamento diferente do problema de mineração, requerendo diferentes métodos (*hierarchical data mining* e *longitudinal data modeling*). Como consequência, são necessárias algumas técnicas específicas de *Data Mining* que possam ser aplicadas na aprendizagem e a outro tipo de dados sobre os estudantes. Contudo, *Educational Data Mining* ainda é uma área de investigação emergente e podemos prever que o seu desenvolvimento posterior irá resultar numa melhor compreensão dos desafios específicos a esta área e irá ajudar os investigadores

envolvidos em *Educational Data Mining* a ver que técnicas podem ser adotadas e que novas técnicas têm de ser desenvolvidas. Por outro lado, há métodos de *Data Mining* que são mais adequados para resolver alguns dos tipos de problemas educativos, como vem descrito no capítulo 2.1.5 (Romero e Ventura, 2013).

2.1.4.4 Interpretação dos resultados

Esta etapa final é muito importante para aplicar os conhecimentos adquiridos para a tomada de decisão sobre a forma de melhorar o ambiente ou o sistema educativo. Assim, os modelos obtidos pelos algoritmos de *Data Mining* têm de ser abrangentes e úteis para o processo de tomada de decisão.

Finalmente, em vez de mostrar o modelo de *Data Mining* obtido, opta-se por apresentar aos utilizadores uma lista de sugestões ou conclusões sobre os resultados e como aplicá-los.

2.1.5 Métodos de *Educational Data Mining*

Em *Educational Data Mining* existem vários métodos que decorrem diretamente da análise de dados que surgem da interação dos estudantes com os ambientes de aprendizagem.

Atualmente, há uma série de métodos populares dentro de *Educational Data Mining*, estando a sua escolha dependente do tipo de objetivo do *Educational Data Mining*. Alguns deles são amplamente reconhecidos como universal em todos os tipos de mineração de dados e outros têm especial destaque dentro de *Educational Data Mining*.

Ryan Baker (2010) classifica os métodos de *Educational Data Mining* da seguinte forma:

- *prediction*:
  - *classification*;
  - *regression*;
  - *density estimation*;
- *clustering*;
- *relationship mining*:
  - *association rule mining*;
  - *correlation mining*;
  - *sequential pattern mining*;

    - *causal data mining*;
- *distillation of data for human judgment*;
- *discovery with models*.

Romero e Ventura (2013) referem que, dos métodos acima indicados, os mais usados em *Educational Data Mining* são: *prediction, clustering, relationship mining, distillation of data for human judgment* e *discovery with models* e, ainda acrescenta os seguintes métodos:

- *outlier detection*;
- *social network analysis*;
- *process mining*;
- *text mining*;
- *knowledge tracing*; e
- *nonnegative matrix factorization*.

Todos os métodos indicados têm o seu objetivo e aplicação em *Educational Data Mining* (ver tabela 1).

Método	Objetivos/Descrição	Aplicação em EDM
Prediction	Inferir um atributo de destino ou aspeto único dos dados (variável prevista) de alguma combinação de outros aspetos dos dados (variáveis de previsão). Existem três tipos de métodos de *prediction*: *classification* (quando a variável prevista é categórica), *regression* (quando a variável prevista é contínua) e *density estimation* (quando a variável prevista é uma função de densidade de probabilidade).	Prever o desempenho do estudante e detetar os seus comportamentos.
Clustering	Formar grupos de dados, de forma que os objetos contidos nos dados fiquem agrupados naturalmente de acordo com a semelhança entre eles.	Agrupar materiais semelhantes dos cursos ou agrupar os estudantes com base na sua aprendizagem e padrões de interação. Promover a aprendizagem colaborativa e agrupar os estudantes a fim de lhes dar tarefas diferenciadas de acordo com as suas capacidades e outras características.
Relationship mining	Identificar relações entre as variáveis e codificá-las em regras para uso posterior. Existem quatro tipos de *relationship mining*: *association rule mining* (relação entre variáveis), *correlation mining* (correlação linear entre as variáveis), *sequential pattern mining* (associações temporais entre as variáveis) e *causal data mining* (relação de causalidade entre as variáveis).	Identificar relações nos padrões de comportamento dos estudantes e diagnosticar as dificuldades de aprendizagem dos estudantes ou erros que ocorrem frequentemente.

Tabela 1. Métodos de *Educational Data Mining*, seus objetivos/descrição e aplicações (continuação) [adaptado de Romero e Ventura (2013)]

Método	Objetivos/Descrição	Aplicação em EDM
Distillation of data for human judgment	Representar os dados de forma mais legível e visual para facilitar a compreensão humana e assim apoiar decisões importantes baseadas em dados. Por um lado, é relativamente fácil obter estatísticas descritivas de dados educacionais para obter características globais, resumos e relatórios sobre o comportamento dos estudantes. Por outro lado, as técnicas de visualização de informação e gráficas ajudam a ver, explorar e compreender grandes quantidades de dados educativos de uma só vez.	Ajudar os educadores/professores a visualizarem e analisarem as atividades de curso dos estudantes e o uso de informação.
Discovery with models	Utilizar um modelo gerado por um método de *prediction*, tal como *classification*, ou por um método de *clustering* e, em seguida, esse modelo é utilizado como componente, ou ponto de partida, em outra técnica de *prediction* ou *relationship mining*.	Apoiar a identificação de relações entre os comportamentos dos estudantes e as características dos estudantes ou variáveis contextuais. Apoia a análise de questões de pesquisa através de uma ampla variedade de contextos, bem como a integração de estruturas de modelagem psicométricas em modelos *machine learning*.
Outliers Detection	Descobrir pontos de dados que sejam significativamente diferentes do resto dos dados. Um *outlier* é uma observação diferente (ou medida) que normalmente toma um valor maior ou menor que os outros dados.	Detetar os estudantes com dificuldades de aprendizagem, desvios nas ações ou comportamentos dos professores ou dos estudantes. Detetar irregularidades nos processos de aprendizagem.
***Social Network analysis* (SNA)**	Compreender e medir as relações entre entidades em informação obtida em rede. SNA vê relações sociais em termos de teoria de rede consistindo em nós (representando atores individuais dentro da rede) e conexões ou ligações (que representam relações entre os indivíduos, tais como amizade, parentesco, posição organizativa, etc).	SNA pode ser usada para interpretar e analisar a estrutura e relações em tarefas colaborativas e interações com as ferramentas de comunicação.
Process mining	Extrair conhecimento relacionado com o processo a partir do registro de eventos gravado por um sistema de informação para ter uma representação visual clara de todo o processo. Este consiste em três subáreas: verificação de conformidade, descoberta de modelos e extensão de modelos.	Pode ser usado para refletir o comportamento dos estudantes no que diz respeito à sua evolução e desempenho ao longo do seu percurso académico.
Text mining	Obter informações de alta qualidade a partir dos textos. Este método inclui tarefas como: categorização de textos, agrupamento de textos, extração do conceito/entidade, análise de sentimentos, síntese de documentos, entre outros.	Analisar o conteúdo dos fóruns de discussão, fóruns, *chats*, páginas da *web*, documentos e assim por diante.
knowledge tracing	Estimar a competência do estudante em determinadas áreas do conhecimento. Para isso, usa-se um modelo cognitivo que mapeie um item de resolução de problemas perante as competências necessárias e regista as respostas corretas e incorretas dos estudantes como prova do seu conhecimento numa determinada competência.	Acompanhar o conhecimento do aluno ao longo do tempo.

Tabela 2. Métodos de *Educational Data Mining*, seus objetivos/descrição e aplicações (continuação) [adaptado de Romero e Ventura (2013)]

Método	Objetivos/Descrição	Aplicação em EDM
***Nonnegative matrix factorization* (NMF)**	Interpretar, de uma forma simples e direta, em termos de matrizes quadradas, também denominado modelo de transferência. Há muitos algoritmos NMF que podem dar origem a diferentes soluções. NMF consiste numa matriz de números positivos, resultante do produto de duas matrizes mais pequenas.	A matriz M $(M = Q * S)$ que representa o resultado observado do teste de um examinando, que pode ser decomposta em duas matrizes: a matriz Q que representa a matriz dos itens e S que representa o domínio de competências de cada estudante.

Tabela 3. Métodos de *Educational Data Mining*, seus objetivos/descrição e aplicações [adaptado de Romero e Ventura (2013)]

Os métodos de *Educational Data Mining*, segundo Baker et al. (2011) têm sido, frequentemente, utilizados para fornecer suporte e mensagens de *feedback* aos professores, orientar os estudantes com recomendações, identificar os grupos de estudantes com características comuns e prever o desempenho e o risco de abandono.

De acordo com Bienkowski et al. (2012), as tecnologias e aplicações de EDM podem construir modelos para responder às seguintes questões:

- Que sequência de tópicos é mais eficaz para um determinado estudante?

- Que ações estudantis estão associadas a um maior nível de aprendizagem?

- Que ações estudantis indicam satisfação, envolvimento, progresso na aprendizagem, etc.?

- Que características de um ambiente de aprendizagem *online* levam a uma melhor aprendizagem?

- O que irá prever o sucesso estudantil?

2.1.6 Mudança dos temas dos artigos ao longo dos tempos

De acordo com Romero e Ventura (2007), citado por Baker e Yacef (2009), os métodos de *relationship mining* (*association rule mining, correlation mining, sequencial pattern mining* e *causal data mining*) foram o tipo de investigação em *Educational Data Mining* mais importantes entre 1995 e 2005. 43% dos artigos nesses anos envolviam métodos de *relationship mining*. A *prediction* era o segundo método mais proeminente, com 28% dos artigos a envolverem vários tipos de métodos de *prediction* (*classification, regression, density estimation*). *Human judgment* e *clustering* seguiam-se com 17% e 15% respetivamente.

De acordo com Baker et al. (2008) e Barnes et al. (2009), citado por Baker e Yacef (2009), um padrão muito diferente é visto nos artigos dos primeiros dois anos da conferência sobre *Educational Data Mining*. Ao passo que *relationship mining* foi dominante entre 1995 e

2005, em 2008-2009, este método deslizou para o 5º lugar, com apenas 9% dos artigos. *Prediction*, que estava em 2º lugar entre 1995 e 2005 mudou-se para a posição dominante em 2008-2009, representando 42% dos artigos EDM de 2008. *Human judgement* e *clustering* permanecem aproximadamente na mesma posição, quer em 2008-2009 quer em 1995-2005, com 12% e 15% dos artigos, respetivamente.

Um novo método, significativamente mais preponderante em 2008-2009 do que em anos anteriores, é o *discovery with models*. Enquanto nenhuns artigos na pesquisa de Romero e Ventura envolviam *discovery with models*, em 2008-2009, este tornou-se a segunda categoria mais comum de pesquisa em EDM representando 19% dos artigos.

Educational Data Mining parece estar a crescer rapidamente e a tendência é continuar a desenvolver-se nos próximos anos.

2.1.7 Técnicas de *Educational Data Mining*

Na sua maioria, as técnicas utilizadas na área de *Educational Data Mining* são oriundas da área de *Data Mining* (Baker, 2012). No entanto, na maioria das vezes, há a necessidade de adaptá-las devido às particularidades existentes em ambientes educacionais e aos seus dados.

As técnicas de *Educatonal Data Minig* mais utilizadas são: árvores de decisão, máquinas de vetores de suporte, regressão logística, redes bayesianas, regressão linear, redes neurais, máquinas de vetores de suporte de regressão, algoritmo *k-means*, algoritmo genético, mistura gaussiana, *quantitative association rule*, *generalized association rule*, *sequencial patterns* e *association rules extended with negative* (adaptado de Romero (2010)).

2.1.8 Principais Aplicações de *Educational Data Mining*

Há muitos exemplos de aplicações ou tarefas em ambientes educacionais que podem ser resolvidos através de *Data Mining*. Entre todos eles, prever o desempenho dos estudantes é a mais antiga e a mais popular aplicação de *Educational Data Mining* (Romero e Ventura, 2013). No entanto, nos últimos anos, a EDM foi aplicada para resolver um grande número de novos e diferentes problemas.

Os autores Baker (2010) e Baker et al. (2012) consideraram como principais as seguintes aplicações:

- **modelagem do estudante** - armazenar informação sobre as características dos alunos, tais como conhecimento, motivação, atitudes, personalidade, além de questões sociais. As técnicas de EDM podem ser utilizadas para dar uma maior precisão no modelo de estudante e proporcionar uma maior personalização e adaptação dos serviços oferecidos por um Ambiente de Aprendizagem Virtual (AVA). Modelar as diferenças existentes entre os estudantes possibilita acompanhar a aprendizagem de forma individualizada, melhorando-a significativamente em cada estudante. Recorrendo aos métodos de EDM é possível modelar características do estudante em sistemas em tempo real.

- **modelagem do domínio** – descrever o domínio de instrução em termos de conceitos, habilidades, itens de aprendizagem e suas inter-relações (Pavlik et al., 2009; Baker et al., 2012).

- **suporte pedagógico** – descobrir que tipos de suporte pedagógico são mais eficientes para a globalidade ou para grupos específicos de estudantes. Baker e Yacef (2009) consideram que a última situação se torna mais complexa devido às particularidades de cada estudante

- **investigação científica** - desenvolver e testar teorias científicas educacionais, para formular novas hipóteses científicas, ou seja, procurar desenvolver melhores sistemas de apoio ao ensino e à aprendizagem (Siemens e Baker, 2012; Baker et al., 2012).

Romero e Ventura (2013), para além das quatro aplicações supra mencionadas, acrescenta outros exemplos de aplicações em *Educational Data Mining*, tais como:

- prever o desempenho do aluno(1);
- fornecer *feedback* para apoiar os professores(2);
- personalizar aos estudantes(3);
- recomendar aos alunos(4);
- criar alertas para os intervenientes interessados(5);
- agrupar os estudantes(6);
- construir cursos(7);
- planear e calendarizar(8);
- estimar parâmetros(9).

Os mesmos autores, em 2011, mencionaram outro exemplo de aplicação em EDM: personalizar aos estudantes(3) adaptando automaticamente a aprendizagem, navegação, conteúdo e apresentação a cada estudante individualmente.

De acordo com Tang et al. (2012), citado por Romero e Ventura (2013), outra aplicação pertinente de EDM é Recomendar aos estudantes(4) – onde se pretende fazer recomendações

aos estudantes no que diz respeito às suas atividades ou tarefas – páginas a visitar, problemas ou cursos a serem feitos, etc.

Os autores Kotsiantis et al. (2010), citado por Romero e Ventura (2013) aplicam os métodos de EDM para criar alertas para os intervenientes interessados(5), de modo a monitorizar o progresso de aprendizagem dos estudantes, para detetar em tempo real, comportamentos estudantis indesejáveis, tais como baixa motivação, uso indevido, copiar, abandono, etc.

Outra aplicação que surgiu foi agrupar os estudantes de acordo com as suas características(6), dados pessoais de aprendizagem, entre outras (Ayers et al., 2009; Tang e McCalla, 2005) promovendo a aprendizagem colaborativa (Tang e McCalla, 2005) para descobrir padrões que reflitam o comportamento dos estudantes, a fim de lhes dar tarefas diferenciadas (Hamalainen et al., 2004).

Segundo Garcia et al. (2009), a construção de cursos(7) foi uma aplicação criada para ajudar os professores e os responsáveis pelo desenvolvimento dos cursos a levarem a cabo o processo de construção/desenvolvimento destes e aplicar o respetivo conteúdo de aprendizagem automaticamente.

Os autores Hsia et al. (2008), citado por Romero e Ventura (2013), desenvolveram planeamento e calendarização(8) para planearem futuros cursos, a calendarização de curso do estudante, planear a atribuição de recursos, monitorizar processos de admissão e aconselhamento, desenvolver o currículo, entre outros.

Os autores Wauters et al. (2011), citado por Romero e Ventura (2013), implementaram a estimação de parâmetros(9), com o objetivo de inferir parâmetros de modelos probabilísticos a partir de dados fornecidos para prever a probabilidade de acontecimentos de interesse.

Relativamente à modelagem do estudante, os autores Frias-Martinez et al. (2006) acrescentam uma outra aplicação, nesta área, a fim de desenvolver e afinar os modelos cognitivos dos estudantes que representam as suas habilidades e conhecimentos declarativos. Baker et al. (2008), por sua vez, criaram uma aplicação para detetar comportamentos inadequados dos estudantes. Eles verificam se o estudante está "usurpando/burlando o sistema" (do inglês *gaming the system*), ou seja, o estudante tenta obter sucesso pedindo diversas dicas, em vez de aprender as matérias e usar esse conhecimento para descobrir a resposta a um determinado problema. Os autores, D´Mello et al. (2008), verificam se um estudante está entediado, confuso ou frustrado na utilização do sistema; isto é feito por meio da análise de atributos extraídos da interação dos estudantes com o sistema, como por exemplo, informação temporal e informação das respostas.

Outras aplicações surgiram, nomeadamente, em 2008, os autores Romero, Ventura e Garcia, organizaram algumas aplicações, desenvolvidas por outros investigadores, mediante as técnicas de EDM.

Os autores, Romero et al. (2008), com a aplicação da técnica *Association Rule Mining*, indicam uma série de aplicações importantes para o ensino *online*. Desta forma, referiram a necessidade da criação de um agente recomendador que pudesse indicar atividades de aprendizagem *online* ou atalhos com base nas ações dos estudantes anteriores. Esta recomendação pode ser uma atividade *online*, como fazer um exercício, ler mensagens postadas ou executando uma simulação *online*, ou pode ser simplesmente um recurso *web* (Zaiane, 2002). Outras aplicações surgiram com o intuito de orientar automaticamente as atividades do estudante criando e recomendando, de forma inteligente, materiais didáticos (Lu, 2004), para determinar que materiais didáticos são os mais adequados para serem recomendados aos estudantes (Markellou et al., 2005), para identificar atributos que caracterizam os padrões de disparidade de desempenho entre diferentes grupos de estudantes (Minaei-Bidgoli et al., 2004). Segundo Romero et al. (2004), citado por Romero et al. (2008), outra aplicação relevante é a de descobrir relações interessantes do uso de informação do estudante, a fim de fornecer um *feedback* ao professor. De acordo com Merceron e Yacef (2004), e citado por Romero et al. (2008), surgiu também a necessidade de encontrar os erros dos estudantes que muitas vezes ocorrem em conjunto.

Romero et al. (2008), com a utilização da técnica *Sequencial Pattern Mining*, indicam uma série de aplicações. Nomeadamente, para avaliar as atividades do estudante na adaptação e na personalização da entrega de recursos (Zaiane e Luo, 2001); para a criação de atividades personalizadas para diferentes grupos de estudantes (Wang et al., 2004), para identificar sequências de interação indicadoras de problemas, de modo que se possa usá-los para ajudar os grupos de estudantes no reconhecimento precoce de problemas e identificar padrões que são marcadores de sucesso a fim de indicar melhorias durante o processo de aprendizagem (Kay et al., 2006).

Romero et al. (2008) utilizam a técnica *Text Mining* para apoiar as pessoas que lidam com uma grande quantidade de documentos, na recolha e preparação dos seus materiais (Grobelnik et al., 2002); para auxiliar a aprendizagem colaborativa e os fóruns de discussão com avaliação entre pares (Ueno, 2004); para classificar as discussões dos estudantes de acordo com um conjunto de padrões, detetando o foco da conversa, identificar tópicos de discussão que podem ter confusões e perguntas sem resposta e estimar a profundidade técnica das contribuições (Kim et al., 2006).

2.1.9 Ferramentas de *Educational Data Mining*

Com o aumento de informações digitais disponíveis, aumenta também o interesse na descoberta de conhecimento implícito no cruzamento dessas informações. Desta forma, multiplicam-se as ferramentas que auxiliam nas diferentes etapas de *Knowledge Discovery in Databases* (KDD) e, em especial, na etapa de *Data Mining*.

Existem diversas ferramentas, gratuitas e comerciais, disponíveis. Dentro das comerciais, temos, por exemplo: *DBMiner*, *Clemetine*, *Oracle Data Mining*, *IBM DB 2 Intelligent Miner* e o *WizRule*; quanto às gratuitas, ou seja, de código aberto, temos: *WEKA* e *RapidMiner*.

As ferramentas supracitadas não são projetadas especificamente para fins e problemas pedagógicos/educativos. Ou seja, são complexas para um professor utilizar porque estão projetadas mais para o poder e flexibilidade do que para a simplicidade. No entanto, tem sido desenvolvido um grande número de ferramentas especificamente orientadas para resolver diferentes problemas educativos. Estas foram criadas por diversos investigadores e vêm mencionadas abaixo, por ordem cronológica.

Em 2004, Romero et al. criaram uma ferramenta, *EPRules*, com o intuito de descobrir regras de previsão/predição nas bases de dados, para melhorar o *software* educacional, de modo a dar *feedback* aos seus autores. Esse conhecimento pode ser muito útil para o professor ou o autor do curso, que poderia decidir que modificações seriam as mais adequadas para melhorar a eficácia do curso (Romero et al., 2004). No mesmo ano, Mazza e Milani, criaram uma ferramenta, *GISMO*, para ajudarem os professores a visualizarem o que está a acontecer nas suas aulas de ensino à distância, de modo a tomarem consciência e proporcionarem um melhor apoio aos seus estudantes (Mazza e Milani, 2004).

Em 2005, Merceron e Yacef desenvolveram a ferramenta *TADA-ED*, para ajudar os professores a identificarem padrões relevantes nos exercícios que os estudantes fazem *online*. Estes podem ser utilizados para ajudar os professores na gestão das suas aulas, compreendendo a aprendizagem dos seus estudantes e refletindo sobre o seu método de ensino e são utilizados também para apoiar a reflexão dos estudantes e fornecer-lhes um *feedback* proactivo, (Merceron e Yacef, 2005). No mesmo ano, os autores Avouris et al. criaram a ferramenta *Synergo/ColAT*, para analisar e produzir visões interpretativas das atividades de aprendizagem (Avouris et al., 2005); e Mostow et al. criaram *LISTEN Mining tool* com o intuito de explorar os grandes registos de interação entre o estudante e o professor/tutor (Mostow et al., 2005). Becker et al. (2005), citado por Romero e Ventura (2013), criaram *O3R* para extrair e interpretar padrões de navegação sequenciais.

Em 2006, Os autores Bellaachia e Vommina (2006), segundo Romero e Ventura, (2013), conceberam a ferramenta *MINEL*, uma estrutura para personalizar o conteúdo de aprendizagem num ambiente adaptável através da análise do comportamento de navegação e do desempenho do estudante de modo a fazê-lo alcançar o seu objetivo de aprendizagem.

Em 2007, Jovanovic et al. criaram a ferramenta *LOCO-Analyst*, uma ferramenta educativa para dar *feedback* aos professores sobre aspetos relevantes do processo de aprendizagem que ocorre num ambiente de aprendizagem baseado na *web* (Jovanovic et al., 2007).

Em 2008, Hershkovitz et al. desenvolveram a ferramenta *Measuring tool*, para medir a motivação dos estudantes *online* (Hershkovitz et al., 2008); Koedinger et al. criaram a ferramenta *DataShop*, com o intuito de armazenar e analisar dados obtidos pelo fluxo de cliques, de modo a entenderem melhor os estados cognitivos e afetivos do estudante (esses resultados têm sido usados para redesenhar o modelo de ensino e melhorar a aprendizagem do estudante) (Koedinger et al., 2008). Selmoune e Alimazighi (2008), segundo Romero e Ventura (2013), conceberam *Decisional tool* para descobrir fatores que contribuem para os índices de sucesso/insucesso dos estudantes.

Em 2009, Garcia et al. desenvolveram a ferramenta *CIECoF*, para dar instruções aos criadores dos *softwares* de curso sobre como melhorá-los. Trata-se de uma ferramenta de fácil utilização e assim os professores apenas precisam de se focar na análise dos resultados e tomar decisões sobre como melhorar os cursos *e-learning* (Garcia et al., 2009). Juan et al., criaram a ferramenta *SAMOS*, que possibilita aos professores acompanharem o estado das aprendizagens dos estudantes e as suas atividades de grupo, através dos relatórios semanais gerados automaticamente a partir dos dados armazenados no servidor. Assim, os estudantes com um nível baixo de atividade são facilmente identificados, o que permite ao professor socorrê-los a tempo, evitando possíveis desistências e conflitos no interior dos grupos de trabalho (Juan et al., 2009). Romero et al., conceberam a ferramenta *AHA! Mining Tool*, para recomendar aos estudantes os *links*/ páginas da *web* mais apropriados numa próxima pesquisa (Romero et al., 2009). No mesmo ano, Johnson e Barnes criaram *EDM Visualization Tool*, que permite aos professores visualizar o processo no qual os estudantes resolveram problemas processuais, na lógica, utilizando um sistema tutorial inteligente. O objetivo desta ferramenta é permitir que os professores consigam navegar, explorar e ganhar perceções sobre o desempenho dos estudantes. Isto permite-lhes uma melhor compreensão dos pontos fortes e fracos dos mesmos para que possam ser feitos ajustes nas aulas ou trabalhos de casa de modo a ajudar ao máximo a aprendizagem (Johnson e Barnes, 2009). Gaudioso et al. (2009), citado

por Romero e Ventura (2013), desenvolveram a ferramenta *PDinamet*, a fim de apoiar os professores na modelagem colaborativa do estudante.

Em 2011, Rabbany et al. criaram a ferramenta *Meerkat-ED*, para analisar a participação dos estudantes nos fóruns de discussão, utilizando técnicas de análise de redes sociais (Rabbany et al., 2011). Pedraza-Perez, Romero e Ventura conceberam a ferramenta *MMT too*l, para facilitar não só aos especialistas em *Data Mining*, como também aos principiantes, a execução de todas as etapas do processo de *Data Minin*g dos dados do *Moodle* (Pedraza-Perez et al., 2011). Graf et al. desenvolveram a ferramenta *AAT*, cujo objetivo é aceder e analisar dados sobre o comportamento dos estudantes em sistemas de aprendizagem. Esta ferramenta pode fornecer informação útil acerca dos processos de aprendizagem dos estudantes permitindo a identificação de material didático difícil ou inadequado e pode, por isso, contribuir significativamente para a criação de atividades e recursos de apoio ao estudante melhores ou mais desenvolvidos (Graf t al., 2011). Bakharia e Dawson (2011), citado por Romero e Ventura (2013), criaram *SNAPP* para visualizar a evolução de relações entre participantes dentro dos fóruns de discussão.

De acordo com Zafra et al. (2012), em 2012 foi criada a ferramenta *DRAL* com o intuito de descobrir as atividades realizadas em sistema *e-learning* mais relevantes para um estudante obter aprovação com base em características extraídas de dados registados num sistema educativo baseado na *web* (Romero e Ventura, 2013).

Garcia-Saiz e Zorrilla (2013), citado por Romero e Ventura (2013), criaram em 2013, a ferramenta *E-learning Web Miner* a fim de descobrir os perfis e modelos de comportamento dos estudantes, relativamente ao modo como trabalham em cursos virtuais.

Embora as técnicas de *Educational Data Mining* tenham sido utilizadas em alguns cursos e instituições com sucesso, é preciso passar do laboratório para o mercado em geral, e para alcançar este objetivo é necessário, num trabalho futuro, que as ferramentas de EDM estejam disponíveis gratuitamente para *download*, de modo que sejam utilizadas por uma população muito mais vasta e ampla. Pois a maioria das atuais ferramentas específicas de EDM não estão disponíveis para *download*. As ferramentas de EDM devem ser inclusivas e integradas nos seus próprios sistemas educacionais baseados em computador, juntamente com os outros instrumentos e devem ser de fácil manuseamento para o utilizador. É também importante que professores e instituições desenvolvam uma cultura de utilização de dados para tomar decisões educacionais e melhorar a educação (Romero e Ventura, 2013).

2.2 Learning Analytics

Learning Analytics é um conceito jovem e em desenvolvimento. De acordo com a 1ª Conferência Internacional sobre *Learning Analytics & Knowledge*, "*Learning Analytics* é a medição, recolha, análise e comunicação de dados sobre os alunos e seus contextos, para fins de compreender e otimizar a aprendizagem e os ambientes em que acorre".

Learning Analytics está a ser definida como uma área de pesquisa e aplicação, e está relacionada com a *academic analytics*, *action analytics* e *predictive analytics*.

"Análise de Aprendizagem refere-se à interpretação de uma grande variedade de dados produzidos e recolhidos em nome dos estudantes, de modo a avaliar o progresso académico, prever desempenho futuro e detetar possíveis problemas. Os dados são recolhidos a partir de ações estudantis explícitas, tais como: completar tarefas e fazer exames, e a partir de ações subentendidas, incluindo interações sociais online, atividades extracurriculares, posts nos fóruns de discussão, e outras atividades que não são diretamente avaliadas como parte do progresso educacional do estudante. Os modelos de análise que processam e exibem os dados auxiliam os membros do corpo docente e funcionários da universidade/instituição na interpretação dos mesmos. O objetivo de Learning Analytics é permitir aos professores e às escolas adequarem oportunidades educativas ao nível da necessidade e capacidade de cada estudante."

"Learning Analytics precisa mais do que simplesmente se centrar no desempenho dos estudantes. Também pode ser usada para avaliar currículos, programas e instituições. Poderia contribuir para esforços de avaliação existentes num campus, ajudando a fornecer uma análise mais profunda, ou poderia ser usada para transformar a pedagogia numa forma mais radical. Ela também poderia ser utilizada pelos próprios estudantes, criando oportunidades para uma síntese holística através de atividades de aprendizagem formais e/ou informais. " (Johnson et al., 2011)

Siemens e Long (2011) consideram que *Learning Analytics* é essencial para penetrar na neblina que se instalou sobre a maior parte do ensino superior. Para os professores, a disponibilidade de informações em tempo real sobre o desempenho dos alunos, nomeadamente, os estudantes que estão em situação de risco, podem ser uma ajuda importante no planeamento das atividades de ensino. Para os estudantes, receber as informações sobre o seu desempenho em relação aos seus colegas ou sobre o seu progresso em relação aos seus objetivos pessoais, pode ser motivador e encorajador. Finalmente, os administradores e os responsáveis pela tomada de decisão são hoje confrontados com uma

incerteza em virtude de cortes no orçamento e competição global no ensino superior. *Learning Analytics* pode penetrar a neblina da incerteza em torno de como distribuir recursos, desenvolver vantagens competitivas e, mais importante, melhorar a qualidade e o valor da experiência de aprendizagem.

2.2.1 Origens de *Learning Analytics*

A partir de 2010, tem-se realizado, todos os anos, uma conferência sobre Learning Analytics (*The International Conference on Learning Analytics and Knowledge*), e em 2011 foi fundada uma sociedade profissional, *The Society for Learning Analytics Research* (SoLAR) (www.solaresearch.org). Até ao momento, ou seja, até 2014, realizaram-se 3 conferências. A 1ª conferência (2011) realizou-se em *Banff*, *Alberta*, Canadá; a 2ª, em 2012, realizou-se em Vancouver, Canadá e a 3ª, 2013, teve lugar em *Leuven*, Bélgica. Este ano realizar-se-á a 4ª conferência em *Indianapolis*, Indiana, EUA.

2.2.2 Principais publicações de *Learning Analytics*

Os documentos sobre *Learning Analytics* mais citados no Google Académico são: *The Horizon Report: 2011 Edition* de Johnson et al. (2011) com 694 citações; *Academic analytics: A new tool for a new era* de Campbell et al. (2007) com 110 citações; *Signals: applying academic analytics* de Kimberly Arnold (2010), com 52 citações; *Penetrating the fog: Analytics in learning and education* de Phillip D. Long e George Siemens (2011) com 51 citações; *The state of learning analytics in 2012: A review and future challenges* de Ferguson (2012) com 43 citações; *Using learning analytics to assess students' behavior in open-ended programming tasks* de Blikstein (2011) com 26 citações e *Academic analytics and data mining in higher education* de Baepler e Murdoch (2010) com 21 citações (Adaptado de [Ferguson, 2012]).

2.2.3 Principais investigadores de *Learning Analytics*

Embora *Learning Analytics* seja uma área de pesquisa e aplicação muito recente, Ferguson (2012), com base nos trabalhos que foram submetidos para a segunda Conferência Internacional Sobre *Learning Analytics & Knowledge* e mediante o número de citações, considerou que os investigadores com mais realce em *Learning Analytics* são: Kimberly

Arnold, especialista em Avaliação Educacional do grupo das Tecnologias de ensino e aprendizagem da Universidade de Purdue, localizada em West Lafayette, Indiana, EUA; Ryan Baker, do Instituto Politécnico de Worcester, EUA, e presidente da sociedade de EDM; John P. Campbell, Reitor Adjunto e Diretor de Informações da Universidade de West Virgínia, EUA; Phill Long, professor de Inovação e Tecnologia Educacional na Escola de ITEE e da Escola de Psicologia, diretor fundador do Centro de Inovação e Tecnologia Educacional (CEIT) da Universidade de Queensland, Austrália; George Siemens, diretor associado do Instituto de Investigação de Tecnologia de Conhecimento avançado na Universidade de Athabasca, Canadá e Etienne Wenger, professor convidado nas Universidades de Manchester e Aalborg, Inglaterra e Dinamarca, respetivamente.

2.2.4 Processo típico de *Learning Analytics*

Segundo Dyckhoff (2012) um processo típico de *Learning Analytics* é semelhante ao que está ilustrado na figura 3.

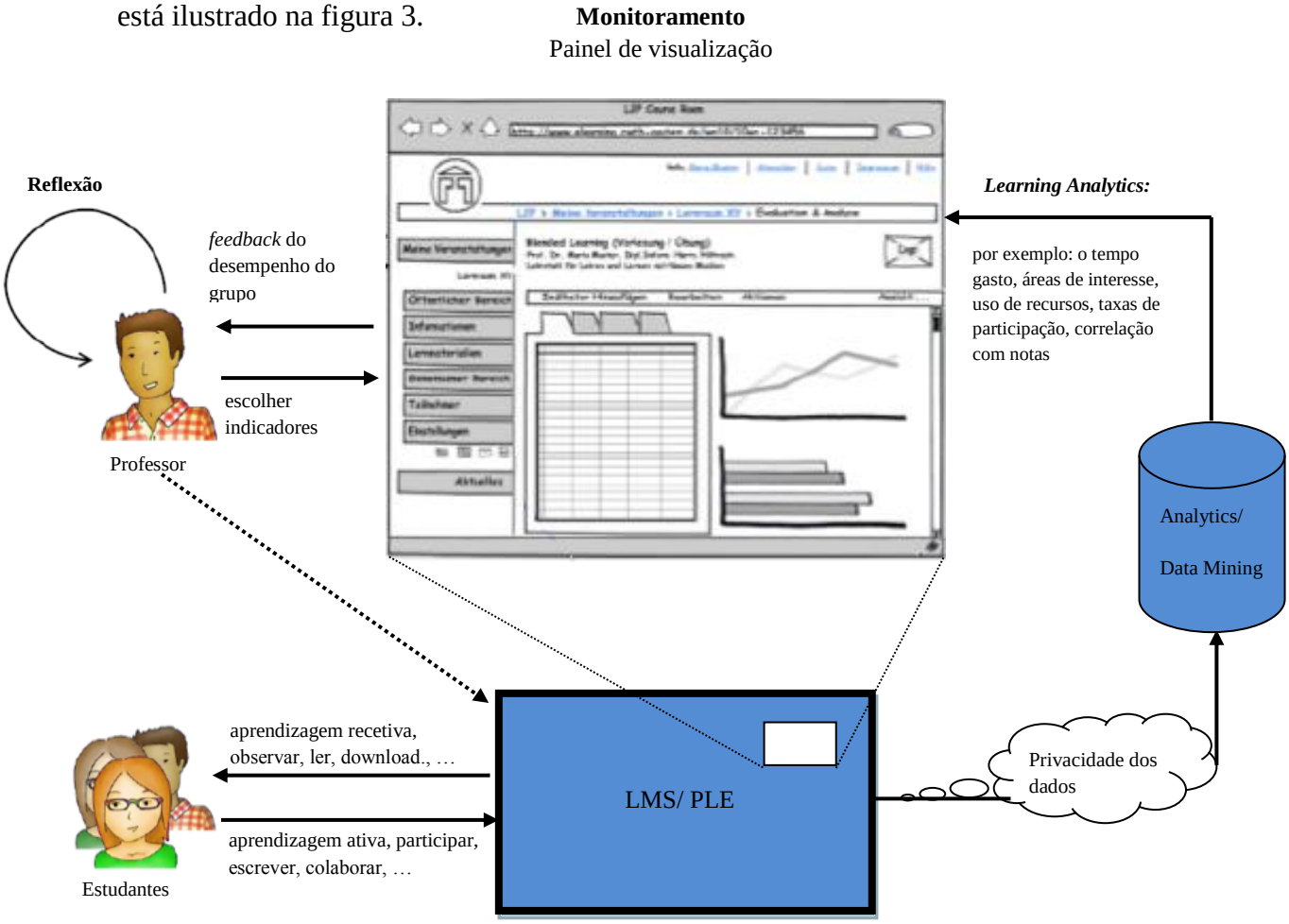

Figura 3. Processo de *Learning Analytics* (adaptado de [Dyckhoff, 2012])

O processo começa com a etapa de recolha de dados. Nesta etapa, os dados são recolhidos a partir de diferentes atividades efetuadas pelos estudantes quando interagem com elementos de aprendizagem dentro de um Ambiente de Aprendizagem Virtual (AVA), Sistema de Gestão de Aprendizagem (LMS) ou um Ambiente de Aprendizagem Pessoal (PLE). Exemplos dessas atividades incluem a participação em exercícios colaborativos, escrevendo um *post* no fórum ou lendo um documento. Nesta etapa, é fundamental respeitar as questões de privacidade dos dados. Geralmente, o *output* da etapa da extração e do pré--processamento dos dados é transferido para uma nova base de dados.

A segunda etapa do processo de *Learning Analytics* é a mineração dos dados pré--processados, com base em diferentes técnicas de *Data Mining*, tais como: *clustering*, *classification*, *association rule mining* e *social network analysis*.

Seguidamente, os resultados do processo de extração podem ser apresentados como um elemento que pode ser integrado num AVA, num painel de instrumentos ou num PLE. Com base nas visualizações gráficas apropriadas dos dados analisados, é suposto que os professores sejam capazes de interpretar mais rapidamente a informação visualizada, refletir sobre o impacto do seu método de ensino no comportamento da aprendizagem e desempenho dos seus estudantes e retirar as primeiras conclusões acerca da eficácia da sua forma de ensino, ou seja, considerar se os seus objetivos foram alcançados. Além disso, os resultados inesperados devem motivar os professores a melhorar, de forma iterativa, as suas intervenções pedagógicas. No entanto, ter uma visualização gráfica não garante que os professores sejam capazes de interpretar corretamente a informação representada. Os indicadores devem ser traçados e avaliados cuidadosamente. Do mesmo modo, o sistema deve fornecer instruções de interpretação.

2.2.5 Métodos de *Learning Analytics*

Ao contrário de *Educational Data Mining*, *Learning Analytics*, geralmente, não aborda o desenvolvimento de novos métodos computacionais para a análise de dados; aborda sim a aplicação de métodos e modelos conhecidos para responder a questões importantes que afetam a aprendizagem do estudante e sistemas de aprendizagem organizativos.

Por oposição à *Educational Data Mining*, que enfatiza respostas automatizadas aos estudantes, *Learning Analytics* permite a adaptação de respostas, tais como através da adaptação do conteúdo instrucional, intervindo com estudantes em situação de risco e fornecendo *feedback* (Bienkowski et al., 2012).

Uma aplicação chave de *Learning Analytics* é monitorizar e prever o desempenho da aprendizagem dos estudantes e detetar possíveis problemas com antecedência para que possam ser providenciadas intervenções a fim de identificar os estudantes em risco de reprovar num curso ou programa de estudos (EDUCAUSE, 2010; Johnson et al., 2011).

Vários modelos de *Learning Analytics* têm sido desenvolvidos para identificar o nível de risco dos estudantes em tempo real a fim de aumentar a probabilidade de sucesso dos estudantes. Exemplos de tais sistemas incluem sistema de Sinais de Curso da Universidade de Purdue (Arnold, 2010) e o sistema *Moodog* sendo usado a nível do curso na Universidade da Califórnia, Santa Barbara e a nível institucional da Universidade do Alabama (EDUCAUSE, 2010). As instituições de ensino superior têm mostrado maior interesse em *Learning Analytics*, uma vez que, eles têm sido alertados para mais transparência e maior escrutínio das suas práticas de recrutamento e retenção.

Segundo Bienkowski et al. (2012), os métodos técnicos utilizados em *Learning Analytics* são variados. Para além dos usados em *Educational Data Mining*, *Learning Analytics* contempla:

- *Social network analysis* (exemplo: análise de relações e interações de estudante para estudante e de estudante para professor para identificar estudantes "desligados", influenciadores, etc.)
- *Social or "attention" metadata* (para determinar a forma como um utilizador está envolvido).

De acordo com Bienkowski et al. (2012), os métodos de *Learning Analytics* permitem responder a questões como:

- Quando é que os estudantes estão prontos para avançar para o próximo tópico?

- Quando é que os estudantes estão a ficar para trás numa determinada unidade curricular?

- Quando é que um aluno está em risco de não concluir um curso/unidade curricular?

- Que nota um estudante poderá tirar sem intervenção?

- Qual é o melhor curso que um determinado estudante pode tirar a seguir?

- Será que um estudante deve ser encaminhado para um conselheiro, a fim de obter ajuda?

2.3 Principais diferenças entre *Educational Data Mining* e *Learning Analytics*

De acordo com os autores Siemens e Baker (2012) e Romero e Ventura (2013), *Educational Data Mining* e *Learning Analytics* têm alguns objetivos comuns e interesses semelhantes. No entanto, existem diferenças que as distinguem. De seguida será apresentada uma breve comparação entre as duas áreas a diferentes níveis:

- Técnicas e métodos: em *Learning Analytics*, as técnicas mais utilizadas são SNA, *sentiment analysis*, *influence analytics*, *discourse analysis*, *learner success prediction*, *concept analysis* e *sensemaking models*. Em EDM, as técnicas mais utilizadas são a *classification*, *clustering*, *bayesian modeling*, *relationship Mining*, *discovery with models* e *visualization.*

- Origens: *Learning Analytics* tem origens mais fortes na *Web* Semântica, "currículo inteligente", previsão de resultados e intervenções sistémicas. *Educational Data Mining* tem fortes origens no *software* educacional e modelagem do estudante com uma comunidade significante na previsão dos resultados.

- Ênfase: *Learning Analytics* tem mais ênfase na descrição dos dados e resultados. Por sua vez, *Educational Data Mining* tem mais ênfase na descrição e comparação das técnicas utilizadas em *Data Mining*.

- Tipo de descoberta: em *Learning Analytics*, aproveitar o julgamento humano é fundamental; a descoberta automatizada é uma ferramenta utilizada para alcançar este objetivo. Por oposição, na *Educational Data Mining* valoriza-se mais a descoberta automatizada e o aproveitamento do julgamento humano é uma ferramenta utilizada para alcançar este objetivo.

Capítulo 3

3 Técnicas de *Educational Data Mining*

3.1 Análise de *clusters*

3.1.1 Introdução

Classificar é uma atividade basilar dos seres humanos. Desde muito cedo, as crianças aprendem a classificar objetos pertencentes ao seu meio envolvente associando os resultados dessa classificação a palavras da sua linguagem (Reis, 2001).

A Análise de *Clusters* é um procedimento da Estatística Multivariada que tenta agrupar um conjunto de dados em grupos homogéneos, chamados *clusters*; os dados podem ser objetos ou variáveis. Ou seja, cada observação pertencente a um determinado *cluster* é idêntica a todas as outras pertencentes a esse *cluster* e é diferente das observações pertencentes aos outros *clusters*.

Uma questão que surge frequentemente aos investigadores é como organizar dados observados, em estruturas com significado. A análise de *clusters* é usada com esse objetivo por investigadores de várias áreas: para descobrir uma estrutura nos dados sem uma explicação/interpretação prévia.

3.1.2 Exemplos de aplicação da análise de *clusters*

Para termos uma perceção mais sólida da relevância da análise de *Clusters*, apresenta-se, em seguida, alguns exemplos da sua aplicação:

- na Arqueologia, a compreensão das civilizações antigas pode resultar do agrupamento e identificação de utensílios usados por povos já desaparecidos;
- nas Ciências Sociais, a antropologia foi influenciada pela análise de *clusters* de modo a identificar áreas culturais semelhantes;
- na Sismologia, a análise de *clusters* é utilizada para se preverem futuros abalos sísmicos;
- na *Data Mining*, a análise de *clusters* é um dos métodos mais utilizados para formar grupos de dados com características semelhantes;
- na Biologia e na Química, a análise de *clusters* pode ajudar na definição de classificação nomeadamente a taxonomia relativa a minerais, insetos, plantas etc.;
- na Medicina, na Psicologia, na Psiquiatria, a análise de *clusters* pode contribuir para a obtenção de melhores tratamentos com base na classificação dos diagnósticos e da sintomatologia;
- na Análise de Mercados, os segmentos de consumidores ou produtos são considerados *clusters*, cuja análise é pertinente para perceber as formas de mercado;
- Em Marketing, a análise de *clusters* tem sido aplicada para dividir o mercado em pequenos grupos de acordo com as características geográficas, demográficas, sociais e económicos dos consumidores. Esta técnica permite antecipar o comportamento do mercado face à introdução de determinados produtos, e assim servir de referência na previsão de vendas;
- um diretor de marketing bancário pode aplicar a análise de *clusters* para identificar grupos de risco quanto aos créditos concedidos;
- um psicólogo organizacional pode utilizar a análise de *clusters* para identificar grupos de eleitores e classificá-los quanto às suas posições políticas.
- na classificação de documentos, a análise de *clusters* aplicada a grandes bases de dados, por exemplo a *Web*, facilita a pesquisa de informação.

3.1.3 Etapas da análise de *clusters*

Segundo Reis (2001), a análise de *clusters* compreende cinco etapas:

1. A seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
3. A definição de uma medida de semelhança ou distância entre cada dois indivíduos;
4. A escolha de um critério de agregação ou desagregação dos indivíduos, isto é, a definição de um algoritmo de partição/classificação;
5. Por último, a validação dos resultados.

De seguida, discutem-se os aspetos fundamentais de cada uma destas etapas.

3.1.4 Seleção de objetos

A seleção de objetos depende dos objetivos da análise. Caso se utilizem dados semelhantes aos de análises anteriores, devemos excluir os objetos que não tenham pertinência para o estudo, tendo o cuidado de manter os objetos importantes para o mesmo. Quando o conjunto de objetos é uma amostra da população, devemos ter em atenção que esta seja representativa de modo a que os grupos resultantes possam representar os grupos existentes na população.

3.1.5 A Seleção das variáveis

São as variáveis que caracterizam os objetos sendo a sua seleção um dos aspetos que mais influencia os resultados da análise de *clusters*. Para isso o investigador deverá ter um conhecimento prévio do assunto a estudar – o objeto de estudo – de modo que os dados disponíveis sejam os mais significativos na abordagem do problema. O responsável pelo estudo deverá ter em atenção o tipo de variáveis utilizadas sobretudo quando estas estão definidas em diferentes unidades de medida.

Relativamente ao número de variáveis, há diferentes opiniões quanto ao número ideal. Este deve corresponder à situação mais "equilibrada" da análise de *clusters*, de modo a que seja robusta para outras bases de dados da mesma população.

Admita-se que se pretende agrupar n indivíduos (que podem ser pessoas, animais, plantas, empresas, países ou mesmo palavras). Segundo Reis (2001), os algoritmos de agrupamento operam, geralmente, sobre dois tipos de estrutura de dados: o primeiro tipo apresenta os indivíduos sob a forma de uma matriz de dimensão $n \times p$ correspondendo as n linhas aos indivíduos e as p colunas aos seus atributos ou características; o segundo tipo consiste numa apresentação sob a forma de um quadro de dimensão $n \times n$ cujos elementos medem as proximidades entre cada par de indivíduos. Estas proximidades poderão ser semelhanças (quando medem a semelhança entre cada par de indivíduos) ou distâncias (quando medem o grau de afastamento ou diferença).

Quando o *input* inicial se encontra na primeira forma, há que atender ao tipo de variáveis – quantitativas (discretas ou contínuas) ou qualitativas – para se escolher o algoritmo de agrupamento adequado. Quando, adicionalmente, as variáveis se apresentam definidas em diferentes escalas de medida e se aplica a análise de *clusters* sem uma estandardização prévia, qualquer medida de semelhança/ distância vai refletir sobretudo o peso das variáveis que maiores valores e maior dispersão apresentam.

O processo de estandardização poderá não ser aconselhado em muitas situações. Se uma variável estiver definida numa escala de medida numericamente mais elevada, a sua dispersão será também mais elevada, resultando daí um maior peso na estrutura final. Com a estandardização, todas as variáveis terão o mesmo peso. Em algumas aplicações poderão existir variáveis com uma importância intrínseca superior, importância essa que deve ser medida e não anulada. Só a experiência e o conhecimento do assunto em estudo poderão ajudar a encontrar a solução mais correta para cada caso. No entanto, quando está ao alcance do investigador recolher dados primários é de toda a conveniência que, logo à partida, as perguntas estejam definidas na mesma unidade de medida (Reis, 2001).

3.1.6 Os métodos de análise de *clusters*

Segundo Reis (2001), na aplicação do método, é necessário identificar a técnica de análise mais apropriada. É possível dividir as técnicas disponíveis em vários grupos:

1. ***Técnicas de otimização***: define-se um critério de agrupamento e a sua otimização indica qual deverá ser o grupo onde cada caso será incluído, pressupondo que todos os casos pertencem a um número k predeterminado de grupos.

2. ***Técnicas hierárquicas***: podem-se subdividir em técnicas aglomerativas e divisivas, ambas partindo de uma matriz de semelhanças ou dissemelhanças (distâncias) entre os casos; estes métodos conduzem a uma hierarquia de partições $P_1, P_2, \ldots, P_n$ do conjunto de n objetos em $1, 2, \ldots, n$ grupos. Os métodos dizem-se hierárquicos porque, para cada par de partições, P_i e P_{i+1}, cada grupo da partição P_{i+1} está incluído num grupo da partição P_i.
3. ***Técnicas de densidade*** *(density or mode-seeking)*: os grupos são formados através da procura de regiões que contenham uma concentração relativamente densa de casos.
4. ***Outras técnicas***: que incluem aquelas em que se permite que haja sobreposição dos grupos (*fuzzy clusters*) e todas as restantes que não foram incluídas nas anteriormente definidas.

Vejamos, de seguida, com maior detalhe, os dois primeiros tipos de técnicas.

3.1.6.1 Métodos de Otimização

Esta metodologia consiste em dividir os diferentes casos de uma matriz de dados em k grupos mais ou menos homogéneos, constituindo cada grupo uma população bem definida. Nesta perspetiva, este método baseia-se diretamente na escolha antecipada de um número de agrupamentos que conterão todos os casos. Posteriormente, são divididos todos os casos pelos k grupos preestabelecidos e a melhor partição dos n casos será aquela que otimiza o critério escolhido.

Nos casos em que há apenas uma variável a caracterizar os indivíduos, considera-se como critério de otimização minimizar, para cada grupo, o somatório do quadrado dos desvios de cada elemento do grupo relativamente à média desse mesmo grupo. Ou seja, minimizar a soma de quadrados dentro dos grupos:

$$\sum_{j=1}^{k}\sum_{u=1}^{n_j}\left(X_{ju} - \overline{X_j}\right)^2$$

em que, $\overline{\mathrm{X}_\mathrm{j}}$ é o valor médio da variável para o grupo j, X_{ju} e são os valores da variável para todos os indivíduos pertencentes ao grupo j $\left(u = 1, 2, \ldots, n_j\right)$.

Para toda a amostra considerada $\left(u = 1, 2, \ldots, n_j\right)$, caracterizada por uma variável X, dividida em k grupos mutuamente exclusivos, pode demonstrar-se que

$$\sum_{u=1}^{n}\left(X_u - \overline{X}\right)^2 = \sum_{j=1}^{k}\sum_{u=1}^{n_j}\left(X_{ju} - \overline{X_j}\right)^2 + \sum_{j=1}^{k}\left(\overline{X_j} - \overline{X}\right)^2$$

sendo $\overline{\mathrm{X}}$ o valor médio da variável X para toda a amostra. Isto é, o somatório do quadrado dos desvios de uma variável em relação à sua média é igual ao somatório do quadrado dos desvios dentro dos grupos em que essa amostra se encontra dividida mais o somatório dos quadrados dos desvios entre os grupos.

A aplicação de um critério de otimização que divida uma amostra em k grupos homogéneos tem como objetivo a afetação dos elementos aos vários grupos de modo a minimizar a primeira parcela do somatório anterior ou, o que é equivalente, a maximizar a segunda. Simplificando, pretende-se que dentro de cada grupo os elementos sejam o mais semelhante possível e o mais diferente possível de elementos de outros grupos.

Quando os indivíduos são caracterizados por p variáveis $(p \geq 2)$, o somatório dos quadrados dos desvios dos valores das variáveis, relativamente à respetiva média, correspondem à diagonal principal de uma matriz, chamada matriz de soma de quadrados e produtos cruzados $(\mathbf{T})$, e que se obtém do seguinte modo

$$\mathbf{T} = \sum_{u=1}^{n}\left(\underset{\sim}{X}_u - \underset{\sim}{\overline{X}}\right)\left(\underset{\sim}{X}_u - \underset{\sim}{\overline{X}}\right)'$$

sendo $\underset{\sim}{\mathrm{X}}_{\mathrm{u}}$ o vetor dos valores das p variáveis para o individuo u e $\underset{\sim}{\overline{\mathrm{X}}}$ o vetor de médias para os n indivíduos da amostra.

Se dividirmos a população em k grupos é possível construir uma matriz $\mathbf{W_j}$ para cada um dos grupos $(j = 1, 2, \ldots, k)$.

$$\mathbf{W_j} = \sum_{\mathrm{u}=1}^{\mathrm{n_j}}\left(\underset{\sim}{\mathrm{X}}_{\mathrm{uj}} - \underset{\sim}{\overline{\mathrm{X}}}_{\mathrm{j}}\right)\left(\underset{\sim}{\mathrm{X}}_{\mathrm{uj}} - \underset{\sim}{\overline{\mathrm{X}}}_{\mathrm{j}}\right)'.$$

Os elementos da diagonal principal de cada uma destas matrizes $\mathbf{W_j}$, representam os somatórios dos quadrados dos desvios de cada variável relativamente à média do grupo j. O somatório dos desvios dentro dos grupos poderá ser obtido a partir da diagonal principal de uma matriz soma de todas as matrizes $\mathbf{W_j}$.

$$\mathbf{W} = \sum_{j=1}^{k}\mathbf{W_j}$$

De modo idêntico, os desvios entre os grupos poderão ser representados pelos elementos da diagonal principal de uma matriz de soma de quadrados dentro dos grupos (**B**).

O objetivo da aplicação de um critério de otimização é encontrar uma partição que minimize a dispersão dentro dos grupos. Isso pode ser conseguido minimizando o traço da matriz **W** ou, o que é equivalente, maximizando a dispersão entre os grupos, isto é, o traço da matriz **B**.

Segundo Reis (2001), um outro critério de otimização muito utilizado consiste em maximizar as distâncias de *Mahalanobis* entre os grupos. A partição ótima é encontrada quando se maximiza a seguinte distância

$$d_{lj} = \left(\underset{\sim}{X}_l - \underset{\sim}{X}_j \right)' \mathbf{S}^{-1} \left(\underset{\sim}{X}_l - \underset{\sim}{X}_j \right)$$

Sendo $\underset{\sim}{X}_l$ e $\underset{\sim}{X}_j$ os vetores dos valores das variáveis para os grupos *l* e *j*, e **S** a matriz agregada de variância/covariância amostral construída a partir das matrizes de cada grupo.

Todos esses critérios podem ser otimizados com o mesmo algoritmo, denominado algoritmo de transferência, uma vez que se transfere um caso de um grupo para outro se isso resultar numa melhoria do critério de agrupamento.

As técnicas de otimização têm duas desvantagens. Primeiro, não há garantias que o algoritmo forneça um ótimo global e não apenas um ótimo local. Em segundo lugar, é necessário um substancial tempo de computação, dado que o modo mais lógico de o fazer seria considerar todas as partições $[k = 2,3,4,...]$ e escolher a melhor de todas elas, o que torna necessária uma capacidade informática considerável. A sua aplicabilidade só se torna possível com a definição, *a priori*, do número de grupos pretendido.

3.1.6.2 Métodos Hierárquicos

Segundo Reis (2001), estes métodos conduzem a uma hierarquia de partições $P_1, P_2, ..., P_n$ do conjunto total de objetos em $1,2,...,n$ grupos. A denominação de hierárquicos advém do facto de, para cada par de partições P_j e P_{j+1}, cada grupo da partição P_{j+1} estar sempre incluído num grupo da partição P_j.

Este tipo de técnica baseia-se na construção de uma matriz de semelhanças ou diferenças em que cada elemento da matriz descreve o grau de semelhança ou diferença entre cada dois casos com base nas variáveis escolhidas. Dentro dos métodos hierárquicos encontramos os

aglomerativos e os *divisivos*. Nos aglomerativos, parte-se de n grupos de apenas um indivíduo cada, que vão sendo agrupados sucessivamente até se encontrar apenas um grupo que incluirá a totalidade dos n indivíduos. Nos métodos divisivos, considerado um processo inverso, parte-se de um grupo que inclui todos os indivíduos em estudo e, através de um processo sistemático de divisões sucessivas, obtém-se n grupos de 1 elemento cada.

Os métodos hierárquicos aglomerativos são os mais divulgados e utilizados na análise de *clusters*, uma vez que os métodos divisivos, tal como os de otimização, são extremamente pesados em termos de capacidade informática.

O terceiro problema a resolver em qualquer análise de *clusters* é a construção de uma matriz de semelhanças ou de distâncias, sendo esta o ponto de partida comum a todos os métodos hierárquicos.

3.1.7 Definição de medidas de semelhança/ distância

Segundo Tversky (1977), citado por Reis (2001), a análise teórica das relações de semelhança tem sido dominada pelos modelos geométricos. Estes modelos representam os objetos como pontos num qualquer espaço de coordenadas de forma que as dissemelhanças observadas entre objetos correspondam a distâncias métricas entre os respetivos pontos. Os métodos de classificação exigem que os índices de semelhança respeitem as propriedades das métricas, que são:

1. ***Simetria***: dados dois objetos, x e y, a distância entre eles verifica a propriedade $d(x,y)=d(y,x)\geq 0$
2. ***Desigualdade triangular***: dados três objetos, x, y e z, as distâncias entre eles satisfazem as propriedades: $d(x,y)\leq d(x,z)+d(z,y)$
3. ***Diferenciabilidade de não idênticos***: dados dois objetos, x e y, $d(x,y)\neq 0 \Rightarrow x\neq y$
4. ***Indiferenciabilidade de idênticos***: dados dois objetos idênticos, x e x', $d(x,x')=0$.

Segundo Aldenderfer e Blashfield (1985), citado por Reis (2001), os índices de (dis)semelhança normalmente utilizados podem ser classificados em quatro categorias:

- coeficiente de correlação;
- medidas de distância;
- coeficiente de associação;

- medidas de semelhança probabilística.

Todas estas medidas têm vantagens e desvantagens, mas os mais utilizados nas ciências sociais são o coeficiente de correlação e as medidas de distância.

3.1.7.1 Coeficientes de correlação

Os coeficientes de correlação, sendo de fácil interpretação geométrica, são das medidas de semelhança mais utilizadas nas ciências sociais. Em particular o coeficiente de correlação de Pearson (este coeficiente é definido, em geral, para duas variáveis e mede o grau de associação linear entre elas), definido para dois indivíduos l e j, caracterizados por um conjunto de p atributos:

$$r_{lj} = \frac{\sum_{v=1}^{p}\left(X_{lv} - \overline{X_l}\right)\left(X_{jv} - \overline{X_j}\right)}{\sqrt{\sum_{v=1}^{p}\left(X_{lv} - \overline{X_l}\right)^2\left(X_{jv} - \overline{X_j}\right)^2}}$$

sendo

$X_{lv} =$ valor da variável v para o indivíduo l, $(v = 1,..., p)$;

$X_{jv} =$ valor da variável v para o indivíduo j;

$\overline{X_l} =$ média de todas as variáveis para o indivíduo l;

$\overline{X_j} =$ média de todas as variáveis para o indivíduo j;

$p =$ número de variáveis.

Este coeficiente assume apenas valores entre -1 e +1, significando o valor zero não existir correlação entre os indivíduos. Este coeficiente é particularmente insensível às diferenças de escala das variáveis, uma vez que o cálculo da média de todas as variáveis para cada indivíduo impõe a estandardização prévia dessas variáveis. No entanto, é sensível às diferenças de forma de cada indivíduo e à dispersão dos valores das variáveis em torno das respetivas médias (Reis, 2001).

Outra desvantagem do coeficiente de correlação reside no fato de uma média de valores diferentes de variáveis não ter significado claro e, daí, calcular correlações entre casos pode não ter qualquer significado estatístico. Para além de tudo isto, este coeficiente não satisfaz a propriedade de desigualdade triangular das métricas. No entanto, a seguinte transformação do coeficiente pode dar lugar a uma métrica:

$$d_{lj} = \left[0{,}5\left(1 - r_{ij}\right)\right]^{1/2}$$

resultando $d_{lj} = 0$ para $r_{lj} = +1$

e $d_{lj} = 1$ para $r_{lj} = -1$.

Apesar das desvantagens, segundo Hamer e Cunningham (1981), citado por Reis (2001), o coeficiente de correlação tem sido utilizado com sucesso, precisamente quando se pretende que os resultados da classificação não sejam afetados por diferenças de dispersão e de escala das variáveis.

3.1.7.2 Medidas de distância

Existem várias medidas que podem ser utilizadas como medidas de distância ou dissemelhança entre os elementos de uma matriz de dados. De acordo com Cormack (1971), citado por Reis (2001), existe uma série de medidas possíveis de entre as quais se podem destacar como mais utilizadas:

1. ***Distância Euclideana***: a distância entre dois casos (i e j) é a raiz quadrada do somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis $(v = 1,2,..., p)$.

   $$d_{ij} = \sqrt{\sum_{v=1}^{p}\left(X_{iv} - X_{jv}\right)^2} \ .$$

   A distância euclidiana pode ser calculada a partir das variáveis originais, ou de variáveis estandardizadas. No primeiro caso as varáveis com maior amplitude apresentarão maiores distâncias euclideanas, enquanto que no 2º caso, a distância euclidiana não é influenciada pela amplitude das variáveis.

2. ***Quadrado da Distância Euclideana***: a distância entre dois casos (i e j) é definida como o somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis $(v = 1,2,..., p)$

   $$d_{ij}^2 = \sum_{v=1}^{p}\left(X_{ij} - X_{jv}\right)^2 .$$

3. ***Distância absoluta ou City-Block Metric***: a distância entre dois elementos (i e j) é a soma dos valores absolutos das diferenças entre os valores das variáveis $(v = 1,2,..., p)$ para aqueles dois casos:

$$d_{ij} = \sum_{v=1}^{p} \left| X_{iv} - X_{jv} \right|$$

4. ***Distância de Minkowski***: definida a partir da medida anterior, pode ser considerada como a generalização da distância Euclideana (as duas coincidem quando $r = 2$):

$$d_{ij} = \left(\sum_{v=1}^{p} \left| X_{iv} - X_{jv} \right|^{r} \right)^{1/r}$$

5. ***Distância de Mahalanobis***: também chamada distância generalizada. Esta medida, ao contrário das apresentadas anteriormente, considera a matriz covariância $\sum$ para o cálculo das distâncias:

$$d_{ij} = \left(\underset{\sim}{X}_i - \underset{\sim}{X}_j \right)' \sum{}^{-1} \left(\underset{\sim}{X}_i - \underset{\sim}{X}_j \right)$$

Sendo $\underset{\sim}{X}_i$ e $\underset{\sim}{X}_j$, respetivamente, os vetores de valores das variáveis para os indivíduos i e j.

6. ***Distância de Chebishev****:* a distância entre dois casos i e j é o valor máximo para todas as variáveis, das diferenças entre esses dois indivíduos.

$$d_{ij} = \max_{v} \left| X_{iv} - X_{jv} \right| \ .$$

A cada passo do processo aglomerativo, a matriz de semelhanças/distâncias é recalculada de modo a saber-se qual é a relação entre os grupos já formados e os elementos ainda não agrupados. Segundo Johnson (1967), citado por Reis (2001), é nesta altura, quando se calcula a relação entre os grupos já formados e os casos restantes, que os métodos aglomerativos apresentam diferenças entre si. Mais precisamente, neste momento do processo, deverá ser satisfeita a seguinte fórmula de recorrência:

$$d_{k(i,j)} = \alpha_i . d_{ki} + \alpha_j . d_{kj} + \beta . d_{ij} + \gamma \left| d_{ki} - d_{kj} \right|$$

Em que $d_{k(i,j)}$ é a distância entre o grupo k e o grupo (i, j) formado pela fusão dos grupos (ou elementos) i e j. Embora a fórmula de recorrência seja sempre a mesma, os coeficientes $\alpha_i, \alpha_j, \beta$ e σ diferem conforme o método aglomerativo escolhido.

Independentemente da medida de distância utilizada, existem vários problemas na sua utilização, sendo o mais relevante o efeito que as diferenças de escala das variáveis provocam sobre o valor das distâncias.

As variáveis que apresentam variações e unidades de medida elevadas facilmente anularão o efeito das outras variáveis. Uma forma de evitar esta situação, como já se referiu anteriormente, é proceder-se à estandardização das variáveis, de modo a tornar a sua média nula e o seu desvio-padrão unitário.

3.1.8 Critérios de agregação e desagregação dos casos

Após escolher a medida de distância, torna-se necessário decidir o critério de (des)agregação dos indivíduos a aplicar. Poder-se-á dizer que os vários métodos pretendem dar resposta aos seguintes tópicos:

- distância entre indivíduos do mesmo grupo e distância entre indivíduos de grupos diferentes;
- dispersão dos indivíduos dentro do grupo;
- densidade dos indivíduos dentro e fora dos grupos.

A grande diferença entre os vários métodos de agregação dos indivíduos é a forma como estimam distâncias entre grupos já formados e outros grupos ou indivíduos por agrupar. Pode dizer-se que o processo de agrupamento de indivíduos já agrupados depende da distância entre os grupos. Ou seja, diferentes definições destas distâncias poderão resultar em diferentes soluções finais.

Segundo Reis (2001), não existe aquilo a que se possa chamar o melhor critério de (des)agregação dos casos em análise de *clusters*. É comum utilizar-se vários critérios e fazer a comparação dos resultados. Se estes forem semelhantes, é possível concluir que se obtiveram resultados com elevado grau de estabilidade e, portanto, fiáveis. Os critérios de agregação mais utilizados são os seguintes:

1. ***Menor Distância* (*Single linkage* ou *Nearest neighbor*)**: define semelhança entre dois grupos – a semelhança máxima entre dois casos pertencentes a esses grupos. Ou dito de outro modo, dados dois grupos (i, j) e (k), a distância entre os dois é a menor das distâncias entre os elementos dos dois grupos:

$$d_{(i,j)k} = \min\{d_{ik}; d_{jk}\}$$

   Deste modo, qualquer grupo é definido como o conjunto de casos em que um elemento é mais semelhante a um outro elemento do mesmo grupo do que a qualquer elemento de outro grupo.

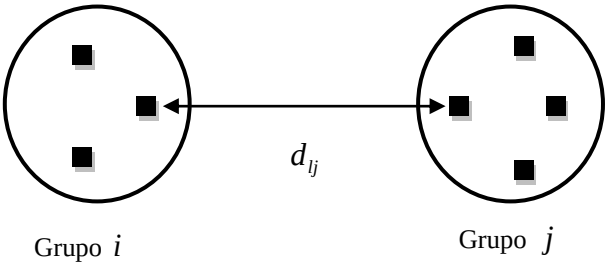

Figura 4. Método de *Single linkage* [adaptado de Reis (2001)]

De acordo com Lance e Williams (1967), citado por Reis (2001), este método torna-se, assim, um sistema contrator do espaço uma vez que cada caso terá mais tendência para se agrupar a um grupo já definido do que para formar o núcleo de um novo grupo. Contudo, para Cormack (1971), Lance & Williams (1967) e Sneath e Sokal (1973), citado por Reis (2001), esta característica torna-se numa desvantagem do método face à possibilidade dos grupos finais se assemelharem a cadeias de elementos quando representados num espaço multidimensional. Razão que tem relegado para segundo plano a utilização do método de *single linkage* como método preferencial de agregação de casos nas ciências sociais. A maior vantagem deste método é ser insensível a transformações monótonas da matriz de distâncias e ainda por não ser afetado pela existência de relações nos dados iniciais.

2. ***Maior Distância* (*Complete linkage* ou *farthest neighbor*)**: atua de forma inversa, uma vez que a distância entre dois grupos é agora definida como sendo a distância entre os seus elementos mais afastados ou menos semelhantes.

   Dados dois grupos (i, j) e (k), a distância entre eles é a maior das distâncias entre os seus elementos:

$$d_{(i,j)k} = \max\{d_{ik}; d_{jk}\}$$

   De acordo com esta estratégia, cada grupo passa a ser definido como um conjunto de elementos em que cada um é mais semelhante a todos os restantes elementos do grupo do que a qualquer dos elementos dos restantes grupos.

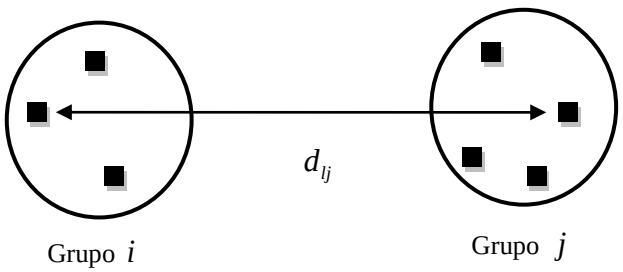

Figura 5. Método do *complete linkage* [adaptado de Reis (2001)]

Este método tem tendência para encontrar *clusters* compactos compostos de indivíduos muito semelhantes entre si. Isto dá-nos uma visão nítida dos diferentes grupos encontrados mas nem sempre apresenta um elevado grau de concordância com a estrutura inicial dos dados.

3. **Distância média entre *clusters* (*Average linkage between groups*)**: esta técnica de agrupamento define a distância entre dois grupos, i e j, como sendo a média das distâncias entre todos os pares de indivíduos constituídos por elementos dos dois grupos.

   De certo modo, esta metodologia parece intermédia das duas primeiras descritas. Enquanto no "menor distância" e no "maior distância", a inclusão de um novo indivíduo num grupo depende de um único valor de semelhança, o menor ou o maior respetivamente, a estratégia da média tem a vantagem de evitar valores extremos e de considerar toda a informação dos grupos. Estes passam a ser definidos como um conjunto de indivíduos no qual cada um tem mais semelhanças, em média, com todos os membros do mesmo grupo do que com todos os elementos de qualquer outro grupo. Este método é chamado de média entre os grupos. No entanto, existe outra metodologia envolvendo a média: a distância média dentro dos *clusters* (*Average linkage within groups*). Neste caso, os *clusters* são combinados de modo a que a distância média entre todos os pares possíveis de indivíduos dentro do grupo daí resultante seja mínima.

4. **Critério de centróide**: nesta estratégia, a distância entre dois grupos é definida como a distância entre os seus centróides, pontos definidos pelas médias das variáveis caracterizadoras dos indivíduos de cada grupo. Isto é, o método de centróide calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis. Uma desvantagem é que, se os dois grupos forem muito diferentes em termos de dimensão, o centróide do novo agrupamento estará mais próximo daquele que for maior e as características do grupo menor tenderão a perder-se. De facto, com este método, o centróide do novo grupo é uma combinação ponderada dos centróides dos dois grupos separados, sendo as ponderações proporcionais ao tamanho destes grupos.

5. **Critério de Ward *(1963)***: baseia-se na perda de informação resultante do agrupamento dos indivíduos e medida através da soma dos quadrados dos desvios das observações individuais relativamente às médias dos grupos em que são classificadas. Este método pode ser resumido nas seguintes etapas:

- primeiro são calculadas as médias das variáveis para cada grupo;
- em seguida, é calculado o quadrado da Distância Euclideana entre essas médias e os valores das variáveis para cada indivíduo;
- depois somam-se as distâncias para todos os indivíduos;
- por último, pretende-se minimizar a variância dentro dos grupos. A função objetivo que se pretende minimizar é também chamada soma dos quadrados dos erros (ESS) ou soma dos quadrados dentro dos grupos (WSS).

No início do processo de agrupamento, cada indivíduo constitui um grupo e $\mathrm{WSS} = 0$. Em seguida, são agrupados os dois indivíduos que provocam um aumento mínimo no valor da soma dos quadrados dos erros, passando a existir $n-1$ grupos; estes $n-1$ grupos são então reexaminados e transformados em $n-2$ grupos, mas de tal modo que o aumento na função objetivo seja minimizado; o processo continua de forma sistemática até todos os indivíduos formarem um grupo apenas. Este método tem como desvantagem a tendência de criar grupos de tamanho semelhante e a de encontrar soluções que podem ser ordenadas a partir de perfis relativos às variáveis iniciais.

Outra desvantagem adicional, partilhada também pelos critérios Distância Média e Maior Distância, é a tendência para encontrar grupos esféricos, mesmo quando a representação gráfica dos dados revela grupos com diferentes formas (figura 6). Isto poderá significar que o método, em vez de extrair a estrutura existente nos dados, impõe uma estrutura que lhe é alheia.

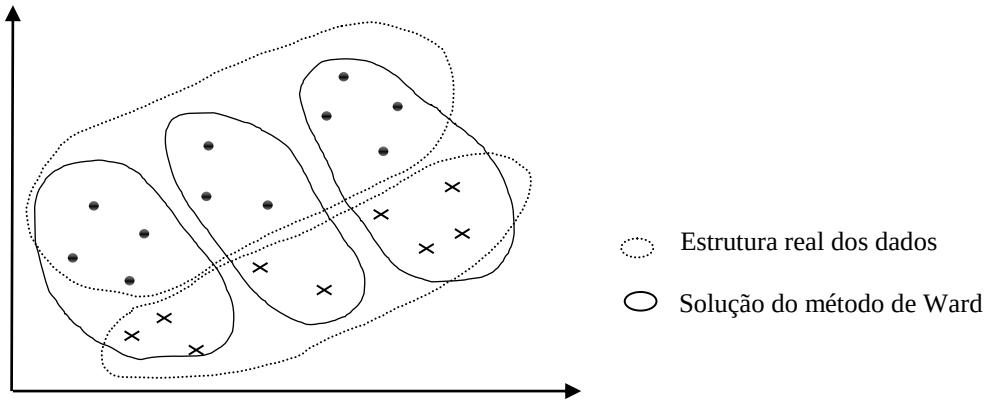

Figura 6. Comparação entre a estrutura real dos dados e a solução do critério de Ward [adaptado de Reis (2001)]

3.1.9 Validação dos resultados obtidos

Como o objetivo da análise de *clusters* é criar grupos homogéneos, surge o problema da escolha do número adequado de *clusters* ou grupos. A aplicação de métodos hierárquicos permite a apresentação dos resultados sob a forma de *dendograma* ou de uma *árvore de agrupamento*.

No dendograma, podemos analisar todas as fases do processo de agrupamento, desde a separação total dos indivíduos até à sua inclusão num só grupo. No entanto, como o objetivo passa por determinar o número ideal de grupos, surge a dúvida por onde cortar o dendograma. Infelizmente, este passo fundamental da análise de *clusters* não está ainda completamente resolvido (Reis, 2001).

Na figura seguinte, o corte do dendograma a uma distância de aproximadamente 3 revela a existência de dois grupos: $(2,5,3,4)$ e $(1,6,7)$.

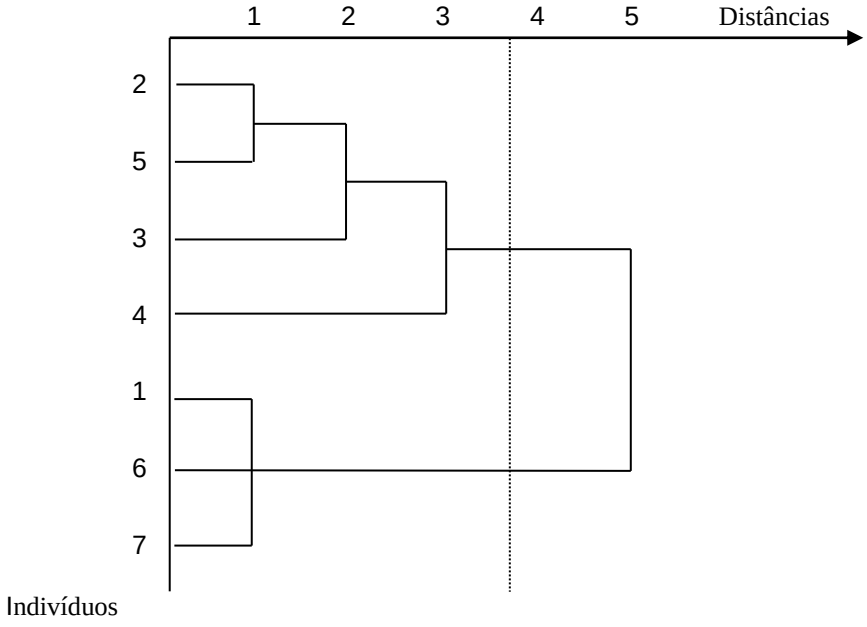

Figura 7. Dendograma [adaptado de Reis (2001)]

Outro método que nos permite visualizar, ao longo do processo, a subdivisão dos grupos e o correspondente número de indivíduos é a árvore de agrupamento.

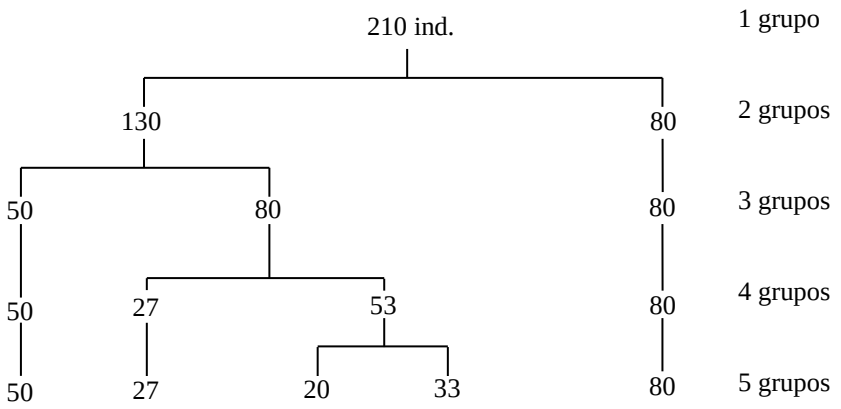

Figura 8. Árvore de agrupamento [adaptado de Reis (2001)]

Muitas vezes, o investigador tem conhecimento prévio do número aproximado de grupos em que a população em estudo deverá dividir-se. Contudo, este é um método muito subjetivo e não pode ser considerado satisfatório por se tornar enviesado pela necessidade de opiniões prévias quanto à correta estrutura dos dados.

De modo a contornarmos este problema, podemos proceder à comparação gráfica do número de *clusters* com o coeficiente de fusão, isto é, o valor numérico (distância ou semelhança) para o qual vários casos se unem para formar um grupo.

Poder-se-á considerar que estamos perante uma partição ótima quando a divisão de um novo grupo não introduz alterações significativas no coeficiente de fusão. Na figura 9, o exemplo indicado sugere que, a partir de 3 grupos, a curva se torna quase paralela a um dos eixos, isto é, os "saltos" mais significativos no coeficiente de fusão dão-se quando se passa de 1 para 2 grupos. Conclui-se então que o agrupamento ótimo se verificará na formação de 2 grupos.

Contudo, gera-se um problema quando a representação gráfica mostra apenas pequenos saltos e não existe nenhuma maneira de avaliar, através da visualização gráfica, qual o melhor número de grupos.

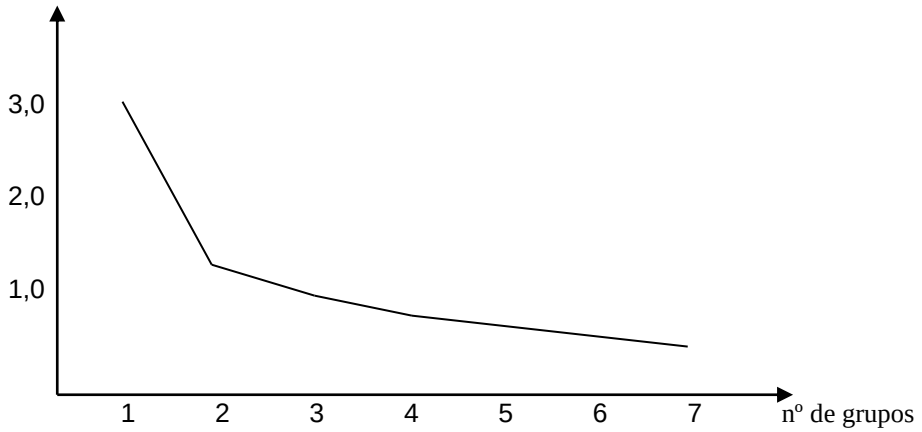

Figura 9. Coeficientes de fusão [adaptado de Reis (2001)]

Sabendo que para uma mesma base de dados, aplicando diferentes técnicas de análise de *clusters*, podemos obter resultados diferentes, surge a dúvida de qual critério de agrupamento é o mais apropriado. Muitos investigadores já se debruçaram sobre a melhor escolha no que respeita ao método de agrupamento e ao número de grupos. No entanto, os resultados a que chegaram não são consensuais.

Segundo Reis (2001), os autores Sokal & Rohlf (1962) definiram o coeficiente de correlação "cofenética" (r_c) que ainda hoje é a medida de validação mais utilizada pelos taxonomistas numéricos. Esta medida dá-nos a relação entre cada valor da matriz de semelhanças e um valor obtido a partir do dendograma significando, em última instância, a medida em que o dendograma resultante da aplicação de um método hierárquico representa os valores da matriz de semelhanças/distâncias. Mais precisamente, a correlação cofenética é a correlação entre os elementos da matriz de distância (ou semelhanças) e os correspondentes coeficientes de fusão, ou seja, as distâncias (ou semelhanças) a que os indivíduos se juntam pela primeira vez para formar grupos. Embora este método de validação seja apropriado – sobretudo quando se utiliza um método hierárquico aglomerativo, foi criticado por Farris (1969) que referiu a sua sensibilidade ao tamanho dos grupos como razão suficiente para não ser aceite como justificação direta e final da técnica utilizada.

Outro procedimento comum é o da utilização de vários critérios de agrupamento e a comparação posterior dos resultados obtidos. Se os resultados forem semelhantes, é possível concluir que, em princípio, qualquer método será de confiança e existem, de facto, grupos entre os indivíduos observados.

Para aferir o grau de convergência dos vários critérios de agregação, pode utilizar-se uma tabela de contingência indicando o número de indivíduos que se agrupam no mesmo *cluster*.

Na mesma tabela, poderá constar o nível de confiança de um teste para a independência dos resultados de cada par de critérios de agregação.

	Método de Ward	***Complete linkage***	***Single linkage***
Complete linkage	75% (0,02)	-	-
Single linkage	80% (0,01)	53% (0,09)	-
Centróide	72% (0,03)	60% (0,08)	70% (0,05)

(Nota: entre parêntesis apresenta-se o nível de significância do teste de independência do χ^2)

Tabela 4. Comparação dos resultados de aplicação de diferentes métodos de agregação para uma partição em 5 grupos. [adaptado de Reis (2001)]

3.1.10 Métodos partitivos iterativos

Os métodos partitivos iterativos aplicam-se diretamente sobre os dados originais, ao contrário dos métodos hierárquicos aglomerativos que requerem o cálculo de uma matriz de semelhanças/distâncias. Com este método, é assim possível a aplicação da análise de *clusters* a bases de dados de maior dimensão.

Segundo Reis (2001), são estes os passos utilizados por este tipo de métodos divisivos:

1. começam por uma partição inicial dos indivíduos por um número de clusters, predefinido pelo analista; calculam, para cada *cluster*, o respetivo centróide;

2. calculam as distâncias entre cada indivíduo e os centróides dos vários grupos; transferem cada indivíduo para o *cluster* relativamente ao qual se encontra a uma menor distância (por exemplo, a distância euclideana);

3. calculam os novos centróides de cada *cluster*;

4. repetem os passos 2 e 3 até que todos os indivíduos se encontrem em *clusters* estabilizados e não seja possível efetuar mais transferências de indivíduos de um *cluster* para outro.

Segundo Reis (2001), o processo de transferência de um indivíduo de um *cluster* para outro pode ser feito através de dois critérios diferentes:

- ***K-means* ou *nearest centroid sorting***: consiste essencialmente na transferência de um indivíduo para o *cluster* cujo centróide se encontra a uma menor distância. Este critério pode ser combinatório ou não, inclusivo ou exclusivo. É combinatório quando

se recalcula o centróide dos *clusters* sempre que há uma alteração na sua composição; é não-combinatório quando os centróides dos grupos são recalculados apenas depois de ser retificada a posição de todos os indivíduos. Este critério é exclusivo quando, ao tratar determinado indivíduo, este é excluído do cálculo do centróide de *cluster* a que pertence.

Este método inclui indivíduo no *cluster* que apresenta uma distância menor entre o indivíduo e o centróide do cluster. Para tal, ou se conhecem os centróides de cada grupo ou terão de ser calculados a partir dos dados originais.

Uma vez identificada uma partição inicial, são calculadas as distâncias euclideanas entre cada indivíduo e cada *cluster*, e todos os indivíduos são transferidos para os clusters para os quais apresentam menor distância. Por exemplo, pretendemos dividir um conjunto de indivíduos por 5 grupos, encontram-se para o indivíduo 32, os seguintes resultados:

	Distâncias aos Centróides				
	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
Resultados Indivíduo 32	2,0	2,7	3,0	1,5	4,2

Tabela 5. Distâncias do indivíduo 32 aos centróides dos grupos

O indivíduo 32 deverá, nesta fase de partição, ser transferido para o *cluster* 4. Depois deste processo ser repetido para todos os casos, poderá recalcular-se o centróide de cada grupo e transferir novamente aqueles casos que não satisfaçam o critério anterior.

Quando os centróides de cada *cluster* inicial não são identificados pelo investigador, o *software* estatístico disponível toma os primeiros K indivíduos da base de dados como os centróides iniciais. No entanto, nem sempre estes indivíduos são suficientemente diferentes para se obter uma boa partição. Uma melhor estratégica consiste na escolha prévia dos K indivíduos da base de dados que apresentem valores relativamente diferentes para as variáveis incluídas.

- ***Hill climbing**:* com este critério a transferência de indivíduos de um *cluster* para outro é feita com base num critério estatístico de otimização, tal como acontece com os métodos de agrupamento não-hierárquicos. A função a otimizar poderá ser o traço da matriz **W**, o traço da matriz ($\mathbf{W^{-1}B}$), o determinante da matriz **W**, ou o maior valor-

-próprio da matriz ($\mathbf{W^{-1}B}$), sendo **W** e **B** as matrizes de somas de quadrados e produtos cruzados dentro e entre os grupos, respetivamente.

Tanto o critério de *K-means* como o de *Hill Climbing* visam identificar grupos homogéneos num espaço multivariado. Embora o primeiro não utilize, de modo explícito, uma função de otimização, implicitamente acaba por otimizar o critério da matriz **W**, minimizando a variância dentro dos grupos. Mais uma vez, se aplicados a um mesmo conjunto de dados, estes dois critérios poderão produzir resultados diferentes.

Estes métodos interativos apresentam uma desvantagem pelo facto de não garantirem que a solução final seja um ótimo global e não um ótimo local, uma vez que é impossível tratar todas as partições possíveis do conjunto de dados nos k *clusters* predefinidos. Este é um problema de difícil resolução causado por uma má partição dos dados iniciais. Uma solução proposta, embora não consensual, é a de encontrar a partição inicial através de um método hierárquico aglomerativo e, a partir daí, aplicar um método interativo.

Concluindo, o método de análise de *clusters* vem dar resposta a um grande leque de problemas que envolvam uma classificação de indivíduos ou casos desde que tenhamos bastante informação.

3.2 Cartas de Controlo

3.2.1 Introdução

Uma carta de controlo é definida por um gráfico que mostra a evolução de uma determinada característica da qualidade ao longo do tempo, onde, para além dos valores da estatística, também constam os valores da linha central (LC) e dos limites inferior (LCI) e superior (LCS) de controlo. Assim, uma Carta de Controlo visa a sinalização dos possíveis desvios relativamente ao funcionamento normal do processo, ao longo do tempo. As Cartas de Controlo são igualmente importantes na análise do processo, isto é, na investigação dos fatores causais-chave que estão na base de uma certa característica da qualidade, como no controlo do processo; isto é, monitorizar uma característica da qualidade e agir sobre os fatores do processo sempre que se verifiquem situações de elevada dispersão (Montgomery, 2005).

As Cartas de Controlo surgiram com o intuito de detetar alterações suscetíveis de ocorrer e interferir num processo. No domínio dos processos industriais, as Cartas de Controlo têm vindo a ser utilizadas com sucesso na monitorização do processo e da qualidade. Nos últimos anos, diversas modificações nas cartas de Shewhart foram propostas na literatura (Runger e Montgomery, 1993; Montgomery, 2005). As modificações produzidas intentam a adaptação das cartas de controlo à monitorização de processos especiais (Tsung, 2000); algumas das modificações situam-se, por exemplo, ao nível das cartas adaptativas de controlo (Michel e Fogliatto, 2002).

Na atualidade, a sua importância expandiu-se aos mais variados ramos do conhecimento, com especial relevo para as áreas dos serviços e biomédica. Em termos de saúde pública, sobretudo a nível hospitalar, as Cartas de Controlo têm possibilitado a melhoria quer do processo de gestão económico-administrativo quer do bem-estar dos utentes, assim como a melhoria do quadro clínico das pessoas hospitalizadas e a prevenção do surgimento de focos infeto-contagiosos (Woodall, 2008).

3.2.2 Desenvolvimento

De uma forma geral, alguns dos pressupostos no acompanhamento do processo de produção são:

- seleção de uma característica da qualidade (por exemplo, o n.º de defeitos);
- seleção do(s) parâmetro(s) a controlar (por exemplo, a variância e o valor esperado);
- recolher regularmente amostras para verificação (por exemplo, o turno da manhã ou da tarde);
- registo sequencial dos valores observados de uma determinada medida estatística (por exemplo: média, desvio-padrão, percentagens observadas de peças com defeito);
- apresentação gráfica dos resultados atendendo aos limites apropriados;
- apresentação de uma carta de controlo.

A apresentação de esquemas/cartas de controlo foi, tal como mencionado anteriormente, proposta em 1924 por Shewhart com o propósito de fiscalizar a produção e reduzir o processo de variabilidade.

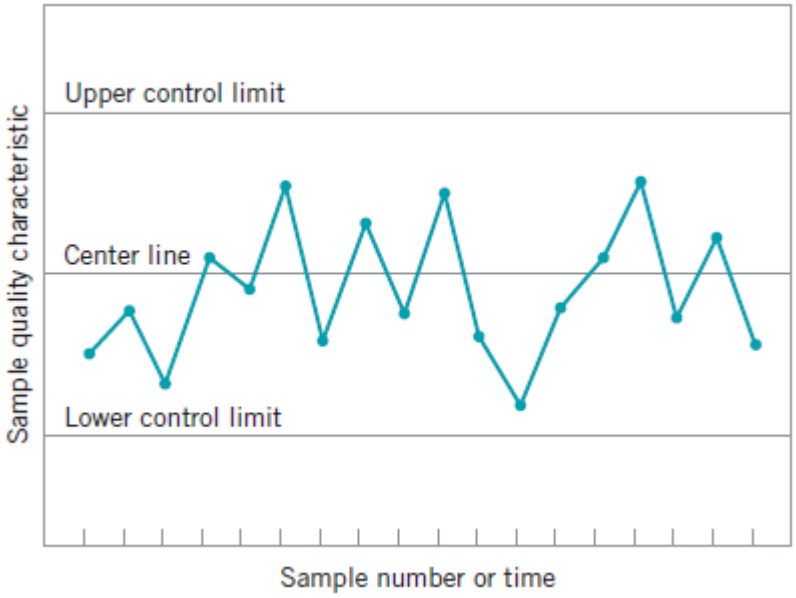

Figura 9. Exemplo de carta de controlo [adaptada de Montgomery, 2009]

Para Shewhart, a variabilidade da qualidade dos processos de produção pode ficar a dever-se as duas ordens de razão:

1. Causas aleatórias (*chance causes*), também designadas de causas comuns, cujo efeito resulta em variações negligenciáveis, incontroláveis e contidas (intrínsecas) na própria natureza aleatória da característica de qualidade (*background noise*). Neste tipo de causas, os valores individuais acerca de uma certa característica são diferentes mas, no seu todo, segue um determinado padrão que pode ser descrito por uma distribuição de probabilidade, com forma e parâmetros de localização e de dispersão (Pereira e Requeijo, 2008). Segundo Montegomery (2005), estas causas são inevitáveis, são inerentes ao próprio processo, e como não é possível serem totalmente eliminadas, se identificadas, deverão ser minimizadas ao máximo. De uma forma geral, estas causas estão relacionadas com 5 fatores: mão de obra, matéria-prima, meio ambiente, maquinaria e métodos.
2. Causas assinaláveis (*assignable causes*), também conhecidas por causas especiais ou fatores particulares de processo, que não estão inseridas na distribuição seguida por uma característica quando o processo está sob controlo estatístico, ou seja, são o resultado de alterações inaceitáveis da qualidade do produto. Na base da ocorrência deste tipo de causas, estão sobretudo os erros de operadores, as alterações ocasionais das condições de trabalho, o desajuste da maquinaria, a matéria-prima inadequada, entre outros (Pereira e Requeijo, 2008). Uma vez mais, Montgomery (2005) informa

que os esquemas de controlo de qualidade têm como objetivo a deteção de causas assinaláveis, na medida em que ao apontarem a existência de um desvio do parâmetro face ao seu valor alvo, traduzir-se-á numa perda da qualidade do produto. Estas causas podem ocorrer por:

a. *Shifts*, ou seja, alterações repentinas de um ou mais parâmetros da distribuição da característica de qualidade, desde o nível desejado para um nível diferente.
b. *Drifts*, ou seja, alterações graduais do valor do(s) parâmetro(s). Eventualmente, também podem ocorrer alterações momentâneas do(s) parâmetro(s), retornando depois à normalidade produtiva.

Segundo Montgomery (2005), existe uma relação íntima entre as cartas de Controlo e os Testes de Hipóteses. Genericamente, as Cartas de Controlo testam as hipóteses nula e alternativa, H_0 vs H_1.

H_0: o processo está sob controlo estatístico (IN, *in control*) quando se está na presença exclusivamente de causas de natureza aleatória.

H_1: o processo está fora do controlo estatístico (OUT, *out of control*) quando, para além das causas aleatórias, ainda se verifica a presença de causas assinaláveis.

Novamente, temos presente a probabilidade de ocorrer um erro do tipo I (α), também designado de risco-α, que se verifica quando se opta pelo estado OUT quando na realidade é IN (isto é, rejeita-se a H_0 e esta é verdadeira); ou pela ocorrência do erro do tipo II (β), ou risco-β, que se verifica quando se opta pelo estado IN quando na realidade é OUT (isto é, aceita-se a H_0 e esta é falsa).

Os testes de hipóteses baseiam-se no pressuposto dos valores do parâmetro da distribuição de probabilidade.

3.2.3 Tipos de cartas de Controlo de Shewhart

Da imensa panóplia de esquemas de controlo de qualidade, emergem os de Walter A. Shewhart, também designados por esquemas de Shewhart.

Um esquema típico de Shewhart para um parâmetro (por exemplo, o valor esperado μ) apresenta as seguintes características:

- No eixo das abcissas, está representado o número da amostra N (ou o momento da recolha dos valores);

- No eixo das ordenadas, está representado o valor observado de uma estatística.

Na base da classificação das cartas de controlo está o tipo de característica da qualidade que elas pretendem controlar. Desta forma, identificam-se:

- **Cartas de Controlo de Variáveis**, cujas características se expressam como valores numa escala de intervalo ou de razão (por exemplo: dimensões, peso, temperatura, pressão arterial, prémios de seguro, consumo de combustível). Estas cartas são utilizadas para analisar e ajustar a variação de um processo estatístico ao longo do tempo. Porque não fornecem apenas mais informação acerca do processo de produção como também são mais eficientes do que as cartas por atributos.
- **Cartas de Controlo de Atributos**, cujas características não se expressam numa escala contínua e as medidas resultam da contagem de dados expressos em escalas nominais ou ordinais (por exemplo: peças com defeito e sem defeito, habilitações académicas dos funcionários de uma empresa).

Todavia, em função do número de características da qualidade controladas num processo estatístico, as cartas de controlo também podem ser classificadas como:

- Cartas de Controlo Univariadas, que são aquelas onde se pretende monitorizar apenas uma característica da qualidade.
- Cartas de Controlo Multivariadas, que são aquelas onde se pretende monitorizar mais do que uma característica da qualidade num mesmo gráfico.

3.2.4 Principais Cartas de Controlo de Variáveis

As cartas de controlo de variáveis mais conhecidas são: Shewhart, CUSUM (*Cumulative Sum*) e EWMA (*Exponentially Weighted Moving Average*).

Relativamente às cartas de Shewhart, as mais utilizadas são:

- Cartas da média (para controlar a média do processo de produção) e da amplitude (para controlar a variabilidade do processo de produção) $\left(\overline{X} - R\right)$ ou do desvio padrão $\left(\overline{X} - S\right)$, onde é obrigatória a recolha de amostras com $n \geq 2$ observações.
- Cartas de valores individuais e amplitudes móveis $\left(X - \overline{MR}\right)$. Estas cartas apenas podem ser aplicadas a variáveis independentes e com distribuição normal. No caso da distribuição ser não-normal, os dados devem ser ajustados de forma a adotar os procedimentos típicos das cartas de controlo de variáveis.

Nota:

Os gráficos de Shewhart são normalmente utilizados em pares. Os gráficos R e S controlam a variação de um processo, enquanto os gráficos $\overline{X}$ monitoriza a média do processo. O gráfico que monitoriza a variabilidade deve ser verificado sempre em primeiro lugar, pois no caso de ele indicar a existência de uma condição fora do controlo (OUT), a interpretação do gráfico para a média será enganosa.

Convenções:

$n =$ tamanho da amostra;

$k =$ número (quantidade) de amostras;

$\overline{\overline{X}} =$ média das médias das amostras (média global);

$\overline{S} =$ desvio-padrão amostral médio;

$\overline{R} =$ amplitude amostral média;

A_2, A_3, d_2, D_3, D_4, etc., são fatores de correção (tabelados).

3.2.5 Escolha da carta de controlo

Com base na revisão da literatura e na proposta apresentada por Gomes et al. (2010), os principais passos a seguir na escolha de cartas de controlo, são:

1. Escolher as características que se pretende que sejam representadas em gráfico. Apesar da existência de um certo caráter subjetivo, pode concentrar-se a atenção nas características de qualidade que têm vindo a surgir com defeito. Esta situação pode ocorrer tanto na fase intermédia como no produto final.
2. Escolher o tipo de carta de controlo. Existem vários tipos de Cartas de Controlo em função da característica da qualidade que se pretende (se é do tipo variável ou atributo). Todavia, qualquer que seja a Carta de Controlo, esta consiste num gráfico onde se observam os limites do controlo que definem as fronteiras de dispersão admissível. As mais comuns são:

Variável (dados contínuos)	$\overline{X} - R; \overline{X} - S; \tilde{X} - R$
Atributo (dados discretos)	np (nº de unidades não conformes)
	p (% de unidades conformes)
	c (nº de não conformidades)
	u (nº de não conformidades por unidade)

3. Escolher o subgrupo racional.
4. Fornecer o sistema para a recolha de dados.
5. Identificar a linha central e os limites de controlo.
6. Calcular os limites de controlo e fornecer informações específicas acerca da interpretação dos resultados e das ações a adotar.

3.2.6 Fases de Elaboração de uma Carta de controlo

Fase I - Recolha de dados: o processo está a decorrer e os dados das características em estudo são reunidos segundo um plano criterioso e registados de forma que possa ser traçado um gráfico. Estes dados podem ser relativos a uma dimensão medida numa peça maquinada, a tempos de passagem de máquina, ao número de erros detetados, etc.

Fase II - Controlo e análise: os limites de controlo são calculados com base nos dados anteriormente recolhidos: eles refletem a variação que é previsível devido somente à presença de causas comuns. Eles são desenhados num gráfico que irá servir de guia para a análise do processo. Os limites de controlo não são os mesmos que os limites de especificação ou os objetivos mas são indicadores da variabilidade natural do processo.

Os dados vão sendo comparados com os limites de controlo para ver se a variação continua estável e resulta unicamente de causas comuns. Se causas especiais estiverem presentes, o processo é estudado com o sentido de descobrir o que o está a afetar. Ações corretivas deverão ser tomadas, de modo geral, localmente.

Novos dados devem ser recolhidos; os limites de controlo são recalculados e quaisquer causas especiais adicionais presentes são estudadas e eliminadas.

Fase III - Melhoria da capacidade: depois de todas as causas especiais terem sido eliminadas e o processo estar sob controlo estatístico, a capacidade do processo pode ser calculada.

Se a variação, devido a causas comuns, for excessiva, o processo não pode produzir produtos que de uma forma consistente cumpram as necessidades dos clientes. O próprio processo deve ser investigado e ações de gestão devem ser tomadas para melhorar o sistema.
Para que a melhoria contínua dos processos tenha efetivamente lugar, estas 3 fases devem ser indefinidamente repetidas. Recolher mais dados, trabalhar para reduzir a variação do processo operando-o num estado de controlo estatístico e melhorar sempre a sua capacidade.

Na maioria dos casos, o processo deve estar centrado no valor nominal especificado. Os processos otimizados produzem produtos que apresentam um pequeno grau de variação devido a causas comuns.

3.2.7 Procedimentos metodológicos na construção de uma carta de controlo

De uma forma genérica, os procedimentos metodológicos de base de uma carta de Controlo assentam em três passos fundamentais:

1. Selecionar uma amostra de um processo de produção ao longo do tempo e escolher uma medida estatística, como por exemplo a média, para representá-lo graficamente.
2. Em função da medida estatística escolhida, apoiando-se numa metodologia estatística apropriada e, eventualmente atendendo aos dados recolhidos, determinar uma Linha Central (LC) (*central line*) e duas linhas designadas de Limite de Controlo Superior (LCS) (*upper control limit*) e Limite de Controlo Inferior (LCI) (*lower control limit*).
3. Efetuar a análise estatística dos dados sendo que a análise preliminar de dados ao nível do controlo de qualidade é essencialmente gráfica.

A Figura abaixo mostra um exemplo de uma Carta de Controlo.

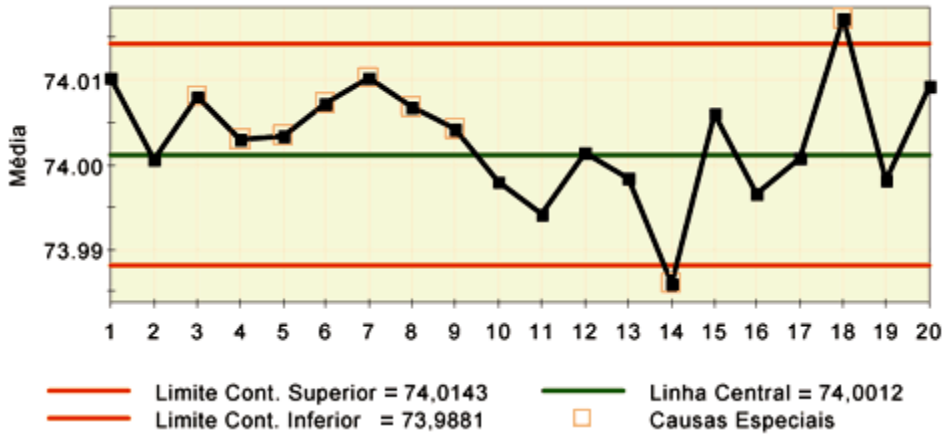

Figura 10. Exemplo de carta de controlo $\overline{X}$ [adaptada de Michel e Fogliatto (2002)]

Quando as medições se situarem dentro dos limites de controlo (LCS e LCI), em princípio, não é realizada nenhuma ação sobre o processo de produção pelo que, nesta situação, a produção prossegue, ou seja, o processo está na situação de funcionamento normal pelo que o processo está sob controlo. Se a medição ultrapassar o limite de controlo, neste caso, procurar-se-á diagnosticar qual a causa determinística que esteve na origem desta ocorrência, bem como escolher o melhor procedimento para a eliminar. Na escolha dos

limites de controlo, deve ser tida em consideração a distribuição amostral da estatística utilizada, de modo a reduzir (ou mesmo eliminar) a probabilidade da estatística assumir valores que saiam fora dos limites de controlo, quando o processo de produção está IN.

Todavia, a carta de controlo deverá emitir sinais de alerta, o mais rapidamente possível, sempre que o processo de produção esteja OUT.

Nota:

Quando apenas se verifica a intervenção de causas aleatórias, diz-se que o processo está sob controlo estatístico (estado IN), e nesta situação o processo de produção pode continuar (Gomes et al., 2010).

Por outro lado, se apenas se verifica a presença de causas determinísticas, o processo não está sob controlo estatístico (estado OUT), procedendo-se à deteção e à eliminação das fontes causadoras do problema (Gomes et al., 2010).

3.2.8 Cartas de Controlo de Shewhart para medidas individuais

Em muitas situações práticas, torna-se inviável a recolha de mais que uma observação para formar a amostra, pelo que a análise se baseia numa única observação $(n=1)$, o que dificulta a construção dos gráficos $\overline{X}-R; \overline{X}-S$ ou $\tilde{X}-R$. Por outro lado, em muitas situações, as amostras que irão ser utilizadas na construção gráfica possuem tamanhos unitários $(n=1)$.

Alguns exemplos onde situações deste tipo ocorrem:

- A taxa de produção é tão baixa que tornar-se-ia inconveniente permitir a acumulação superior a uma única observação para que a análise dos resultados possa ser efetuada.
- As medidas repetidas do processo diferem apenas devido a erros laboratoriais ou de análise, como por exemplo, aqueles que são suscetíveis de se verificarem em muitos processos químicos.
- Emprego da inspeção automatizada na qual toda a unidade produzida é alvo de avaliação.

Nestas situações, o gráfico de controlo surge como uma ferramenta útil. Para construção de gráficos para medidas individuais, o desvio-padrão pode ser estimado da mesma forma que é utilizada no gráfico $\overline{X} - R$: $\sigma = \frac{\overline{R}}{d_2}$.

Por sua vez, a amplitude $\overline{R}$ é estimada a partir do cálculo da amplitude móvel (*moving range*) MR, obtidas em $m-1$ amostras: $MR_i = |x_i - x_{i-1}|$ para $i = 2,3,...,m$

$$\overline{MR} = \frac{\sum MR_i}{m-1}$$

Para cartas de controlo para medidas individuais, os limites são:

- Limites de controlo para a média (gráfico *X*)

$$LCS = \overline{X} + 3\frac{\overline{MR}}{d_2} \qquad LC = \overline{X} \qquad LCI = \overline{X} - 3\frac{\overline{MR}}{d_2}$$

- Limites de controlo para a amplitude móvel (gráfico MR)

$$LCS = D_4\overline{MR} \qquad LC = \overline{MR} \qquad LCI = D_3\overline{MR}$$

Assim, $\overline{X}$ é a média aritmética dos valores individuais; MR é a amplitude móvel média; d_2, D_3 e D_4 são valores constantes e tabelados (Ford Motor Company, 2002).

3.2.9 Cartas de Controlo de Shewhart para média e amplitude – carta $\overline{X}$ e R

Este género de gráfico é utilizado no acompanhamento, controlo e na análise de processos com valores em escala contínua da qualidade de um produto. Como por exemplo comprimento, peso, altura, dimensão, IMC e temperatura. Um gráfico deste tipo é composto por dois gráficos que devem ser analisados em conjunto:

- $\overline{X}$, que monitoriza as médias dos subgrupos;
- R, que monitoriza a amplitude entre os subgrupos.

O tamanho de cada subgrupo (n) deve ser constante, embora se tenha que ter 2 ou mais subgrupos.

Deste modo, através de técnicas de monitorização com cartas de controlo, é possível observar tais processos e verificar uma possível existência de qualquer anomalia no mesmo;

isto é, podem detetar-se variações na centralização e/ou na dispersão do processo. Fica claro, no entanto, que o objetivo é identificar essas anormalidades da maneira mais rápida possível e também descobrir quando tal anomalia passou a influenciar o processo, para que se identifique a causa da mudança e se corrija o erro mais rapidamente. De salientar que as cartas podem também ser usadas para identificar melhorias. Desta forma, são um instrumento de grande importância na melhoria contínua de processos.

As Cartas de Controlo, para a média e amplitude, utilizam-se para analisar e controlar processos cuja característica da qualidade se exprime através de uma variável quantitativa. As cartas de médias $\left(\overline{X}\right)$, para controlar a média do processo de produção e as cartas de amplitude $\left(R\right)$, para controlar a variabilidade do processo de produção.

De uma forma geral, este tipo de cartas permite saber se o processo está ou não controlado, ou seja, se representa ou não variações do tipo causal.

Desde que o processo esteja sob controlo estatístico elas permitem:

- prever de forma adequada o comportamento do processo ajudando a garantir que este tenha consistência em termos de custo e qualidade;
- melhorar, com base na informação disponível nas cartas, os processos no sentido de reduzir a variabilidade, fornecendo um instrumento para verificação da eficácia das ações de melhoria.

Ao distinguirem entre as causas comuns e as causas especiais que afetam os processos, os gráficos de controlo facilitam indicações precisas sobre a oportunidade e possibilidade de ações corretivas, no próprio local de trabalho ou através de decisões da direção da empresa.

Para cartas de média e amplitude, os parâmetros são:

- Gráfico da média $\overline{X}$:

$$LCS = \overline{\overline{X}} + A_2 * \overline{R} \qquad LC = \overline{\overline{X}} \qquad LCI = \overline{\overline{X}} - A_2 * \overline{R}$$

- Gráfico da Amplitude R :

$$LCS = D_4 * \overline{R} \qquad LC = \overline{R} \qquad LCI = D_3 * \overline{R}$$

Notas:

- Para construção de gráficos de controlo baseados em $\overline{X}$ e R, consideram-se m amostras preliminares, onde é usual considerar m como sendo, pelo menos, 20 ou 25. Cada amostra contém n observações acerca da característica da qualidade que se pretende avaliar. Cada amostra, também conhecida por subgrupos racionais, deverão ser extraídas sempre que

se acredita que o processo está IN e mantidas as condições de operação de modo uniforme (tanto quanto possível).

- É usual a construção de gráficos $\overline{X}$ e R sempre que a dimensão da amostra é $n < 6$, caso contrário é preferível a utilização de gráficos $\overline{X}$ e S.
- Depois do gráfico de controlo estar construído, é necessário avaliar o comportamento dos pontos no interior do gráfico, de forma a verificar se o processo está IN ou OUT, para a eventualidade de ser necessário recalcular os limites do gráfico motivado pela presença de pontos que extrapolaram os limites (superior e/ou inferior). Este procedimento deve ser repetido até o processo estar IN.

3.2.10 Cartas de Controlo de Shewhart para média e desvio-padrão – carta $\overline{X}$ e S

Estes gráficos são semelhantes aos gráficos $\overline{X}$ e R, ainda que o cálculo do desvio padrão da amostra (S) seja mais difícil do que o da amplitude (R). O gráfico $\overline{X}$ é utilizado com o propósito de controlar a média do processo, ao passo que o gráfico S é utilizado no controlo da variabilidade do processo. Contudo, os dois gráficos devem ser empregues em simultâneo. Do ponto de vista estatístico, a utilização do desvio padrão amostral assume-se como um procedimento muito mais eficiente.

A utilização dos gráficos de controlo $\overline{X}$ e S deve verificar-se sempre que uma característica da qualidade observada é expressa em unidades reais (como por exemplo: peso em quilogramas, altura em centímetros, performance na corrida de velocidade em segundos e centésimos de segundo, temperatura em graus celsius).

Este gráfico é composto por dois gráficos que devem ser analisados em conjunto:

- $\overline{X}$, que monitoriza as médias dos subgrupos;
- S, que monitoriza a variabilidade entre os subgrupos.

O tamanho de cada subgrupo (n) deve ser constante, embora tenhamos que ter 2 ou mais subgrupos. Sempre que forem adotadas amostras de maior dimensão $(n > 10)$, a amplitude já não é eficiente para avaliar a variabilidade do processo, pelo que devemos usar S.

Para as cartas de média e desvio padrão, os limites são:

- Gráfico da média $\bar{x}$:

$$LCS = \overline{\overline{X}} + A_3 * \overline{S} \qquad LC = \overline{\overline{X}} \qquad LCI = \overline{\overline{X}} - A_3 * \overline{S}$$

- Gráfico do desvio-padrão s:

$$LCS = B_4 * \overline{S} \qquad LC = \overline{S} \qquad LCI = B_3 * \overline{S}$$

Em que:

$\overline{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n} \overline{X_i}$,

$\overline{S} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2}$, onde $\overline{X}$ é a média da I-ésima amostra para $i = 1,2,...,n$

3.2.11 Cartas de Controlo de Shewhart para mediana e amplitude – carta $\tilde{X}$ e R

Perante situações em que se torna importante a localização do ponto central da distribuição ordenada., é interessante o recurso aos gráficos $\tilde{X}$ e R.

Este gráfico é composto por dois gráficos que devem ser analisados em conjunto:

- $\tilde{X}$, que monitoriza a mediana dos subgrupos;
- R, que monitoriza a amplitude entre os subgrupos.

Para as cartas da mediana e amplitude, os limites são:

a. Gráfico da mediana $\tilde{X}$:

$$LCS = \tilde{\overline{X}} + A_2 * \overline{R} \qquad LC = \tilde{\overline{X}} \qquad LCI = \tilde{\overline{X}} - A_2 * \overline{R}$$

b. Gráfico da amplitude R:

$$LCS = D_4 * \overline{R} \qquad LC = \overline{R} \qquad LCI = D_3 * \overline{R}$$

3.2.12 Vantagens e Desvantagens das Cartas de Controlo de Shewhart

Como vimos, as cartas e controlo apresentam inúmeras vantagens mas não devem ser desvalorizadas as desvantagens.

- Cartas $\overline{X}$ e R: a principal vantagem reside na facilidade aquando da elaboração dos cálculos, sendo que a principal desvantagem se situa ao nível da menor segurança e na variabilidade do processo.
- Cartas $\overline{X}$ e S: como principal vantagem, encontra-se a menor variabilidade do processo e como desvantagem a maior dificuldade computacional.
- Cartas $\tilde{X}$ e R: a principal vantagem está na maior facilidade no controlo contínuo, onde não há necessidade de cálculo, sendo que a mais importante desvantagem se situa no fato da mediana ser um estimador mais fraco do que a média aritmética.
- Cartas X e MR: como principal vantagem a adequação para casos em que as medições são demoradas e dispendiosas, e como principal desvantagem a fraca sensibilidade às alterações do processo, comparativamente a outras cartas.

Nota final acerca das principais desvantagens das cartas de Shewhart:

As principais desvantagens das cartas de Shewhart residem na insensibilidade a pequenas variações suscetíveis de ocorrer nos processos bem como no facto de utilizarem apenas a informação do último ponto. Como alternativa, sugere-se a utilização de cartas CUSUM (*Cumulative Sum*) e EWMA (*Exponentially Weighed Moving Average*) que não foram alvo de abordagem neste trabalho.

3.3 Regressão Logística

3.3.1 Introdução

Em muitos cenários de análise de regressão, a variável dependente é qualitativa e assume apenas valores de classes discretas e mutuamente exclusivas. Nestes casos, a regressão categorial é a técnica de análise de regressão a utilizar. É de salientar que a regressão categorial serve os mesmos propósitos da regressão linear, nomeadamente os inferenciais e os de estimação. As diferenças entre os dois modelos residem nos pressupostos de aplicação e no método de obtenção das estimativas dos coeficientes do modelo. Enquanto na regressão linear a variável dependente é do tipo quantitativo, na regressão categorial a variável dependente é qualitativa, e as variáveis independentes ou preditoras, também designadas de covariáveis, podem ser quantitativas ou qualitativas. Contudo, a regressão categorial toma designações

diferentes consoante o tipo de variável dependente qualitativa que se pretende modelar. Quando a variável dependente é nominal dicotómica, a regressão categorial designa-se por regressão logística binária (Maroco, 2007).

A regressão logística é amplamente usada em ciências médicas e sociais. Por exemplo: em medicina, permite determinar os fatores que caracterizam um grupo de indivíduos doentes em relação a indivíduos sãos; no domínio dos seguros, permite encontrar frações da clientela que sejam sensíveis a determinada política securitária em relação a um dado risco particular; em instituições financeiras, pode detetar os grupos de risco para a subscrição de um crédito; em econometria, permite explicar as intenções de voto em atos eleitorais.

O êxito da regressão logística assenta sobretudo nas numerosas ferramentas que permitem interpretar, de modo aprofundado, os resultados obtidos.

3.3.2 Modelo de regressão logística binária univariado

Os modelos de regressão são utilizados na análise de dados com o intuito de descrever a relação entre uma ou mais variáveis independentes e uma variável resposta (Martins, 2008).

A análise apresentada neste capítulo baseia-se essencialmente no trabalho de Hosmer e Lemeshow (1989).

Qualquer problema de regressão passa por estimar o valor esperado da variável resposta, Y, dado o valor das variáveis independentes, x.

Na regressão linear assume-se que este valor esperado pode ser expresso como uma equação linear em função de x.

Considerando o modelo de regressão linear simples, tem-se

$$E[Y \mid x] = \beta_0 + \beta_1 x .$$

Tendo em conta a expressão anterior, verifica-se, que $E[Y \mid x]$ pode tomar qualquer valor compreendido no intervalo $]-\infty;+\infty[$.

Na regressão linear, a observação da variável resultado pode ser expressa como

$$y = E[Y \mid x] + \varepsilon ,$$

Sendo ε o erro associado.

De acordo com $y = E[Y \mid x] + \varepsilon$, ε dá o desvio de uma observação em relação à média condicional. A hipótese mais comum é que este ε segue uma distribuição normal com média

zero e variância constante, ao longo dos níveis da variável independente. Daqui resulta que esta distribuição condicional da variável, resultado dado o valor da variável ε, segue uma distribuição normal, com média $E[Y \mid x]$ e variância constante.

Contudo, isto não se verifica quando se tem uma variável resultado dicotómica. Assim, nesta situação deve-se expressar o valor da variável como

$$y = \pi(x) + \varepsilon,$$

considerando-se $\pi(x) = E[Y \mid x]$.

De acordo com Hosmer e Lemeshow (1989), quando se trabalha com dados dicotómicos, a média deverá assumir valores entre 0 e 1. A variação de $E[Y \mid x]$ em função de x é menor, consoante a aproximação da média condicional de 0 ou 1. Assim, a curva resultante tem uma forma em S, sendo semelhante ao gráfico de uma distribuição cumulativa de uma variável aleatória. Neste caso, usa-se o modelo de regressão logística.

Foram propostas muitas funções para análise de variáveis dicotómicas, Cox, em 1970, citado por Hosmer e Lemeshow (1989), apresentou várias razões para a escolha da distribuição logística para a análise de dados, destacando-se:

1. o ponto de vista matemático, como sendo uma função extremamente flexível e muito usada;
2. por si mesma, conduz a uma fácil interpretação dos resultados em termos biológicos.

A resposta esperada é dada pela expressão

$$E[Y \mid x] = \beta_0 + \beta_1 x,$$

sendo Y uma variável aleatória que segue uma distribuição de *Bernoulli*, com a seguinte lei de probabilidade:

$$\begin{cases} Y = 1 \Rightarrow P(Y = 1) = \pi(x) \; sucesso \\ Y = 0 \Rightarrow P(Y = 0) = 1 - \pi(x) \; insucesso \end{cases}$$

Aplicando a definição de valor esperado, obtém-se:

$$E[Y \mid x] = \pi(x).$$

Logo,

$$E[Y \mid x] = \beta_0 + \beta_1 x = \pi(x).$$

Considere-se uma amostra de n observações independentes com o par (x_i, y_i), onde x_i e y_i representam o valor da variável independente e o valor da variável resposta, respetivamente, sendo i o i-ésimo elemento.

A função de regressão logística univariada é dada pela esperança de Y dado x, ou seja,

$$\pi(x) = E[Y \mid x] = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}.$$

Os parâmetros considerados são estimados pelo método de máxima verosimilhança que consiste em determinar os valores dos parâmetros que maximizem a probabilidade de obter o conjunto de valores observados.

Uma propriedade interessante, que a função logística possui, é que pode ser linearizada. Assim, fazendo essa transformação vem

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right),$$

Obtendo-se

$$g(x) = \beta_0 + \beta_1 x,$$

onde $-\infty \le g(x) \le +\infty$; $-\infty \le x \le +\infty$.

Esta transformação é chamada de transformação *logit* de probabilidade $\pi(x)$. A razão $\pi(x)/[1-\pi(x)]$, na transformação *logit* é a chamada *odds* ou "*chance*".

A importância desta transformação é que $g(x)$ tem muitas propriedades desejáveis dos modelos de regressão linear. A função *logit*, $g(x)$, é linear nos seus parâmetros, podendo ser contínua, e variar entre valores de $]-\infty;+\infty[$, dependendo do domínio de variação de x (Martins, 2008).

3.3.2.1 Função de verosimilhança

O método geral de estimação alternativo ao da função dos mínimos quadrados, para o modelo de regressão linear, é o método de máxima verosimilhança (MMV). Este método dá a base para a aproximação de estimação com o modelo de regressão logística (Braga,1994).

Atendendo a Hosmer e Lemeshow (1989), o MMV permite obter valores para os parâmetros desconhecidos, que maximizam a probabilidade de obter o conjunto de observações.

A função de verosimilhança expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os estimadores de máxima verosimilhança destes parâmetros são escolhidos de modo a ser aqueles que maximizam a função de verosimilhança.

Neste caso, em que se tem apenas dois resultados possíveis (sucesso $Y=1$ e o insucesso $Y=0$), e desde que as observações sejam independentes, a função de verosimilhança é dada por:

$$l(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}\cdot(1-\pi(x_i))^{1-y_i}\ .$$

Em que $\pi(x_i)$ representa a $P[Y=1\mid x]$, ou seja, a probabilidade de sucesso.

O princípio de máxima verosimilhança usa para estimativa de β, os valores que maximizam a expressão obtida anteriormente. Contudo, é mais fácil trabalhar com a expressão dos logaritmos da verosimilhança, sendo

$$L(\beta)=\ln(l(\beta))=\sum_{i=1}^{n}\{y_i\cdot\ln[\pi(x)]+(1-y_i)\cdot\ln[1-\pi(x)]\}$$

e para se obter o valor de β , que maximiza $L(\beta)$, deriva-se esta em ordem a cada parâmetro e igualam-se as equações de verosimilhança a zero.

Para regressão logística envolvendo duas variáveis, as equações de verosimilhança são não lineares em β , o que vai requerer métodos especiais para a sua resolução, sendo o método de resolução de equações não lineares, usualmente aplicado, o método de *Newton-Raphson* (Martins, 2008).

3.3.2.2 Teste de *Wald*

Em regressão logística, tem-se variáveis resultado e uma ou mais variáveis explicativas. Para cada variável explicativa do modelo, haverá um parâmetro associado.

O teste de *Wald*, citado por Romero (2007), é uma das possíveis formas de testar se os parâmetros associados com um grupo de variáveis explicativas tomam o valor zero.

Segundo Romero (2007), este teste é utilizado para avaliar se o parâmetro é estatisticamente significativo. A estatística teste que se utiliza é obtida através da razão do coeficiente pelo seu respetivo erro padrão. Esta estatística de teste segue uma distribuição normal $N(0,1)$. A estatística de teste, para avaliar se o parâmetro β é igual a zero é:

$$W=\frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}}\ .$$

3.3.3 Modelo de regressão logística binária multivariado

A regressão logística pode ser utilizada, fazendo as necessárias adaptações, para modelar situações com mais do que uma variável independente.

Considere-se n observações independentes do par (x_i, y_i), em que x_i é um vetor de m variáveis independentes e y_i uma variável dicotómica. A função logística que se usa para modelar esta situação é semelhante à usada para o modelo univariado, apresentado anteriormente, envolvendo as m variáveis independentes:

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_m x_m}}{1-e^{\beta_0+\beta_1 x_1+\ldots+\beta_m x_m}}$$

Os $m+1$ parâmetros desconhecidos são estimados pelo método da máxima verosimilhança, aplicando processos iterativos, onde as equações de verosimilhança são dadas por:

$$\begin{cases} \dfrac{\partial L}{\partial \beta_0} = 0 \\ \dfrac{\partial L}{\partial \beta_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \\ \displaystyle\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0 \end{cases} , j = 1,\ldots,m$$

Independentemente do número de variáveis usadas para definir o modelo de regressão logística, pretende-se distinguir dois grupos distintos de indivíduos, consoante apresentem ou não determinada característica.

Salienta-se que neste estudo, é importante reduzir o número de variáveis a serem incluídas no modelo. Esta redução constitui uma mais-valia em termos estatísticos pois o aumento do número de variáveis incluídas tende a aumentar o risco de sobreajuste do modelo, principalmente em amostras de pequena dimensão (Hosmer e Lemeshow, 1989).

Assim, regra geral, esta situação traduz-se em valores extremamente elevados das estimativas dos coeficientes e/ou dos erros padrão.

Com o objetivo de verificar se as variáveis independentes possibilitam identificar corretamente os elementos que pertencem a cada grupo, constrói-se o modelo de regressão logística que inclui todas as variáveis e, posteriormente, avalia-se a qualidade do seu ajuste. Assim, os valores preditos são então comparados com os valores da variável resposta, que toma dois valores possíveis, 0 ou 1. Os indivíduos são bem classificados se o valor absoluto da diferença entre o valor predito e o da variável resposta for menor que 0,5. Se a maior percentagem de indivíduos for bem classificada, é conveniente que se tente encontrar um

novo modelo com menos variáveis que nos permita separar os elementos de dois grupos (Martins, 2008).

3.3.3.1 Teste de significância estatística

Para testar a significância do ajustamento modelo completo é necessário formular as hipóteses:

$H_0 =$ O modelo ajusta-se aos dados. *vs* $H_1 =$ O modelo não se ajusta aos dados.

Dado o interesse em se utilizar um teste estatístico de forma a avaliar a razão de verosimilhança, será usado o seu logaritmo, o qual multiplicado por menos dois, resulta numa distribuição conhecida. Este valor é designado por *D*, sendo o teste utilizado o da razão de verosimilhança.

Assim, a estatística *D* tem como objetivo comparar o modelo em análise e o modelo saturado) ou seja,

$$D = -2\ln\left[\frac{L_C}{L_S}\right],$$

onde L_C é a verosimilhança do modelo reduzido (completo ou não) e L_S é a verosimilhança do modelo saturado (modelo com todas as variáveis independentes e suas interações).

Tem-se assim o seguinte teste para testar a significância em que as m variáveis são independentes. Para este teste temos as seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0 \qquad vs \qquad H_1 : \beta_j \neq 0 \ \exists_{j=0,...,m \text{ e } m=1,...,k}$$

sendo o Teste da Razão de Verosimilhança, o qual se pode definir do seguinte modo:

$$G = D(\textit{verosimilhança} \text{ sem as } m \text{ variáveis}) - D(\textit{verosimilhança} \text{ com as } m \text{ variáveis}) =$$

$$= -2\ln\left[\frac{\textit{verosimilhança}\ (\text{modelo sem as } m \text{ variáveis})}{\textit{verosimilhança}\ (\text{modelo com as } m \text{ variáveis})}\right]$$

O teste G segue a distribuição de Qui-Quadrado com m graus de liberdade, sob a validade da hipótese nula.

Assim, ao rejeitar H_0, pode-se concluir que pelo menos um, ou até os m coeficientes poderão ser diferentes de zero (Braga, 1994).

3.3.3.2 Métodos de seleção de variáveis

A inclusão ou a exclusão de uma variável no modelo pode variar conforme o problema a considerar ou até mesmo a área científica em análise (Braga, 1994).

Quando se minimiza o número de variáveis a incluir no modelo, obtém-se um modelo numericamente mais estável e mais generalizado. As variáveis que não estão corretamente incluídas no modelo podem provocar o aumento dos erros padrão estimados assim como uma maior dependência do modelo que se traduz nos dados observados (Braga,1994).

Indo ao encontro de Hosmer e Lemeshow (1989), seguem-se alguns passos que podem ajudar quando se tem que selecionar as variáveis a serem incluídas no modelo de regressão logística. Este processo é semelhante ao utilizado na construção do modelo de regressão linear.

Deste modo o processo pode ser descrito tendo em conta os seguintes passos:

(A) Deve-se iniciar o processo por uma análise univariada e individual de cada uma das variáveis. Hosmer e Lemeshow (1989) sugeriram que variáveis nominais, ordinais e contínuas com alguns valores inteiros poderão ser tratadas recorrendo-se a tabelas de contingência dos p níveis da variável dependente *versus* os k níveis da variável independente.

Quando se trata de variáveis independentes e contínuas, é desejável que a análise univariada envolva o ajuste de um modelo de regressão logística como o objetivo de se obter estimativas dos coeficientes, estimativas de erro padrão, o teste de razão de verosimilhança para a significância dos coeficientes e estatísticas de *Wald* univariada.

Pode-se ainda usar como alternativa o *teste-t* para duas amostras. O *teste-t* para duas amostras independentes usa-se quando se pretende comparar as médias de uma variável quantitativa em dois grupos diferentes de indivíduos e se desconhecem as respetivas variâncias populacionais (Pestana e Gageiro, 2005).

A análise baseada neste teste poderá ser útil na determinação da inclusão ou exclusão da variável no modelo.

(B) Quando a análise univariada estiver concluída, passa-se para uma análise multivariada. Após sujeitas a um teste univariado, seleccionam-se as variáveis que apresentarem um valor prova inferior a 0.25, sendo essas variáveis tomadas como candidatas ao modelo multivariado (pode ainda incluir-se, no mesmo modelo, variáveis consideradas importantes no contexto do estudo ou análise).

A escolha do valor 0.25 como critério de seleção foi feita tendo em conta os trabalhos realizados em regressão linear e regressão logística de Bendel e Afifi, e, Mickey e Greenland, citados por Braga (1994).

Segundo estes autores, o valor de 0.05, por vezes, falha para algumas das variáveis em análise. Por outro lado, quando se consideram níveis elevados, podem-se incluir no modelo variáveis com interesse questionável.

Geralmente, a decisão começa por ter em conta um modelo multivariado com todas as variáveis possíveis dependendo da dimensão e número de elementos que constituem cada grupo de variáveis candidatas ao modelo.

Assim, quando se têm dados adequados para suportar a análise, será conveniente começar o modelo multivariado nesse ponto. Caso contrário, esta aproximação pode conduzir a um modelo numericamente instável. Neste último caso, a estatística de *Wald* não deverá ser usada para a seleção das variáveis. Dever-se-á recorrer a uma aproximação para seleção de variáveis baseada no método passo a passo, no qual as variáveis seleccionadas, quer por inclusão quer por exclusão segundo uma ordem sequencial baseada unicamente num critério estatístico (Braga, 1994; Hosmer e Lemeshow, 1989).

(C) Com o modelo multivariado construído, tem que se verificar a importância de cada variável a ser incluída neste. Para tal, deve aplicar-se o teste de *Wald* para cada variável e comparar o valor de cada coeficiente estimado com o seu valor no modelo univariado contendo somente essa variável.

As variáveis, que não contribuam para explicar corretamente o modelo, deverão ser eliminadas e ajustar-se um novo modelo. Este novo modelo deverá ser comparado com o antigo, aplicando-se o teste da razão de verosimilhança.

O processo de retirar, reajustar e verificar deve continuar até parecer que as variáveis explicativas do modelo estejam todas incluídas e em oposição às pouco importantes excluídas do modelo.

Se, no fim do processo da análise univariada, se tiver um número elevado de variáveis candidatas a explicativas ao modelo, será aconselhável utilizar-se a técnica passo a passo (Braga, 1994; Hosmer e Lemeshow, 1989).

(D) Por fim, e após se ter obtido um modelo que pareça conter as variáveis importantes, deve fazer-se uma reanálise de forma a se considerar a necessidade da inclusão de interação entre variáveis (Braga, 1994; Hosmer e Lemeshow, 1989).

3.3.3.2.1 Seleção automática

Qualquer procedimento para adição ou remoção de variáveis num modelo é baseado num algoritmo que verifica a importância das variáveis, incluindo ou excluindo-as do modelo, baseando-se na regra de decisão.

O critério para adição ou remoção de variáveis, em regressão linear, é geralmente baseado na estatística F, comparando os modelos com e sem as variáveis em análise. Em regressão logística, os erros seguem uma distribuição binomial sendo baseado no teste de razão de verosimilhança.

Existem métodos automáticos que podem ser utilizados na decisão de inserir e remover variáveis.

Seguidamente, descrevem-se os métodos implementados no SPSS.

- ***Enter***: é um procedimento para a selecção de variáveis em que todas elas, em bloco, entram no processo uma única vez; (SPSS Inc, 2007).

- ***Forward***: Método de selecção *Stepwise*. Este procedimento inicia-se com um modelo que não contenha variáveis explicativas. A ideia do método é adicionar uma variável de cada vez, selecionando em primeiro lugar aquela que apresentar um valor de correlação mais elevado, em módulo, com a variável resposta, e assim, consecutivamente, até que o processo pára quando o aumento do coeficiente de determinação, devido à inclusão de uma nova variável explicativa no modelo, não é mais importante (http://portalaction.com.br).

  - *Forward* (condicional): baseado na significância da estatística de pontuação e testes de remoção com base na probabilidade de uma estatística de razão de verosimilhança, com base em estimativas de parâmetros condicionais; (SPSS Inc., 2007).
  - *Forward (Likelihood Ratio):* baseado na significância da estatística de pontuação e testes de remoção com base na probabilidade de uma estatística de razão de verosimilhança baseada na máxima verosimilhança parcial das estimativas; (SPSS Inc., 2007).
  - *Forward (Wald)*: Método de selecção *Stepwise* baseado na significância da estatística de pontuação e testes de remoção com base na probabilidade da estatística de *Wald*; (SPSS Inc., 2007).

- ***Backward****: Enquanto o método Forward começa sem nenhuma variável no modelo e adiciona variáveis a cada passo, o método Backward faz o oposto. Este incorpora inicialmente todas as variáveis, e ao longo do processo cada uma pode ou não ser eliminada.*

A primeira variável a ser removida é aquela que apresenta um menor coeficiente de correlação parcial com a variável resposta (http://portalaction.com.br).

- *Eliminação Backward (Condicional): baseada na estatística de razão de verosimilhança de probabilidade das estimativas condicionais dos parâmetros (SPSS Inc., 2007).*
- *Backward Elimination (Likelihood Ratio): baseado na probabilidade da estatística de razão de verosimilhança apoiado nas estimativas de probabilidades parciais (SPSS Inc., 2007).*
- *Backward Elimination (Wald): baseado nas probabilidades da estatística de Wald (SPSS Inc., 2007).*

Stepwise é um dos métodos mais utilizados e consiste na combinação dos dois métodos anteriores (*Forward* e *Backward*). Este inicia com uma variável (a que apresentar maior correlação com a variável resposta) e, a cada passo do *Forward*, depois de incluir uma variável, aplica o *Backward* para ver se será descartada alguma variável. Continua-se o processo até este não incluir ou excluir nenhuma variável (http://portalaction.com.br).

3.3.4 Razão de possibilidades (*odds ratio*)

Atualmente, muitos investigadores optam por analisar a relação entre duas variáveis de escala nominal através do rácio de produtos cruzados – *razão de possibilidade* – pois tem uma interpretação mais fácil do que o teste de *Qui Quadrado* (Bessa, 2007).

De acordo com Bessa (2007), a razão de possibilidade é uma medida antiga tendo sido usada por *Snow* no seu trabalho clássico de identificação do factor risco da propagação da cólera em Londres (1853). Sendo utilizado como medida de associação em estudos de "caso-controle" e em estudos transversais controlados.

A razão de possibilidade é a razão entre duas *odds*, onde as *odds* são calculadas da seguinte forma:

$$odds = \frac{\text{" probabilidade de um acontecimento ocorrer"}}{\text{" probabilidade de um acontecimento não ocorrer"}}$$

Assim, a razão de possibilidade é uma forma de se comparar se a probabilidade de um determinado evento é a mesma para dois grupos.

Considerando-se a seguinte tabela 2 por 2:

	X^-	X^+	
Y^-	a	b	$a+b$
Y^+	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Tabela 6. Razão de possibilidade [adaptado de Pestana & Gageiro (2005)]

Daqui tira-se que:

$$razão\ de\ possibilidades = \frac{a \times d}{b \times c}.$$

e ainda que:

- $razão\ de\ possibilidades = 1$, implica que o evento é igualmente provável em ambos os grupos;
- $razão\ de\ possibilidades > 1$, significa que o evento é mais provável no 1º grupo;
- $razão\ de\ possibilidades < 1$, implica que o evento é menos provável no 1º grupo.

Conclui-se que o significado da razão de probabilidade é semelhante ao risco relativo obtido em estudos de *coorte* e expressa a força de associação entre o evento e o grupo (Pestana e Gageiro, 2005).

- Assim, segundo o que foi referido em Pestana e Gageiro (2005), uma medida mais direta comparando as probabilidades em dois grupos é o risco relativo, que também é conhecida como a relação de risco. O risco relativo é simplesmente a razão de duas probabilidades condicionais.

3.3.5 Avaliar o ajuste do modelo

Quando se fala na qualidade do ajuste de um modelo de regressão logística tem que se ter em atenção a análise de medidas das diferenças entre os seus valores observados da variável resposta, y, e os resíduos.

Sendo o objetivo avaliar o "bom" ajuste do modelo construído através da regressão logística, pode-se fazê-lo usando representações gráficas dos valores dos resíduos. Este caso permite comparar os resíduos dos vários elementos. Pode ainda aplicar-se testes baseados em estatísticas desses valores, fundamentados no valor da estatística de teste e avaliando a qualidade do ajuste do modelo de uma forma global (Martins, 2008).

Após aplicação de um teste de análise de resíduos e quando a qualidade do modelo não é validada por todos esses elementos, o ideal será verificar a existência de elementos com valores de resíduos elevados (em módulo) comparando-os com os resíduos dos restantes elementos (Martins, 2008).

Relativamente às medidas das diferenças dos valores observados e preditos, usados em regressão logística, destacam-se os resíduos de *Pearson* e os *Deviance residuals*, denotados por r e d, respetivamente.

Ou seja:

$$r_j = r\left(y_j, \hat{\pi}_j\right) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}}$$

$$\textbf{e } d_j = d\left(y_j, x_j\right) = \pm\sqrt{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + \left(m_j - y_j\right)\ln\left(\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}\right)\right]}$$

onde $j = 1,2,...,J$, sendo J o número de valores diferentes de x, $x = \left(x_1, x_2, ..., x_m\right)$, e m_j o número de indivíduos com $x = x_j$.

Sob a validade do modelo ser adequado, as estatísticas acima têm aproximadamente uma distribuição $\chi^2_{J-(m+1)}$.

Devendo-se rejeitar a hipótese nula para valores elevados da estatística de teste, essa aproximação só é valida se os valores de m_j forem também elevados (*Kuss* (2002), citado por Martins, (2008)).

Em 1989, Hosmer e Lemeshow propuseram uma estatística de qualidade de ajuste para um modelo de Regressão Logística em que os dados devem ser agrupados em g grupos com as respetivas probabilidades estimadas.

Denote-se:

n_g : o número de indivíduos;

c_g : o número de valores diferentes do conjunto das p variáveis independentes;

o_g : soma dos valores da variável resposta, com $o_g = \sum_{j=1}^{c_g} y_i$;

$\bar{\pi}_g$: média das probabilidades estimadas para o grupo k, com $\bar{\pi}_g = \sum_{j=1}^{c_g} \frac{m_j\hat{\pi}_j}{n_g}$.

Assim, a estatística de Hosmer-Lemeshow segue uma distribuição aproximadamente de um *Qui-quadrado* com $g-2$ graus de liberdade, segundo a hipótese do modelo ser o adequado.

Rejeitando-se a hipótese nula para valores elevados de estatística de teste, C, e podendo expressar-se da seguinte forma:

$$C = \sum_{k=1}^{g} \frac{o_g - n_g \overline{\pi}_g}{n_g \overline{\pi}_g \left(1 - \overline{\pi}_g\right)}$$

Note-se que este resultado depende dos grupos que são escolhidos (Martins, 2008).

3.3.6 Curva ROC

3.3.6.1 Perspetiva Histórica

Uma prática comum, na área relacionada com a medicina, é a forma de se descrever como e quanto uma variável contínua ou categórica ordinal é capaz de classificar materiais ou indivíduos em grupos definidos.

A análise **ROC** (***R**eceiver **O**perating **C**haracteristc*) é uma ferramenta que permite medir e especificar problemas no desempenho do diagnóstico em medicina.

A curva ROC foi usada pela primeira vez durante a segunda Guerra Mundial, aplicada à análise de radar antes de ter sido empregue na teoria de deteção de sinais (*Green e Sweets*, citado Braga (2000)). Depois do ataque a *Pearl Harbor*, em 1941, o exército dos Estados Unidos focou-se na investigação vocacionada a aumentar a previsão de detetar corretamente aviões Japoneses através dos sinais de radar.

Nas décadas de 60 e 70, as curvas ROC foram utilizadas na psicologia experimental e em ramos da biomédica, respetivamente. Nesta ultima, o objetivo principal passou basicamente por classificar os indivíduos em "doentes" ou "não doentes". (Braga, 2000).

3.3.6.2 Conceitos Básicos

A análise da curva ROC pode ser feita por meio de um gráfico simples e robusto, que nos permite estudar a variação da sensibilidade e especificidade, para diferentes valores de corte.

Este conceito aplica-se a várias áreas e permite ter perspectivas diferentes de diversas situações.

Assim, na área da Medicina, a *sensibilidade* (*Sens*.) é definida como a probabilidade do teste fornecer um resultado positivo, dado que o indivíduo é realmente portador da "doença" enquanto que a *especificidade* (*Esp.*) é definida como a probabilidade do teste fornecer um resultado negativo quando o indivíduo não é portador da "doença" (Margotto).

Outra aplicação será a de que as curvas ROC foram desenvolvidas no ramo das comunicações como uma forma de demonstrar as relações entre sinal-ruído. Neste sentido, podemos interpretar o sinal como os verdadeiros positivos (*sensibilidade*) e o ruído como os falsos positivos (*1- especificidade*) (Braga, 2000).

3.3.6.3 Gráfico da curva ROC

Segundo Braga (2000), a curva ROC é um gráfico de Sensibilidade (ou taxa de verdadeiros positivos) versus taxa de falsos positivos, ou seja, representa-nos a Sensibilidade (ordenadas) e 1 – Especificidade (abcissas), resultantes da variação de um valor de corte ao longo do eixo de decisão x. Assim, a representação da curva ROC, permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da e Especificidade correspondente ao ponto que se encontra mais próximo do canto superior esquerdo do diagrama, uma vez que o indício de verdadeiro positivo é 1 e o de falso positivo 0.

Graficamente tem-se:

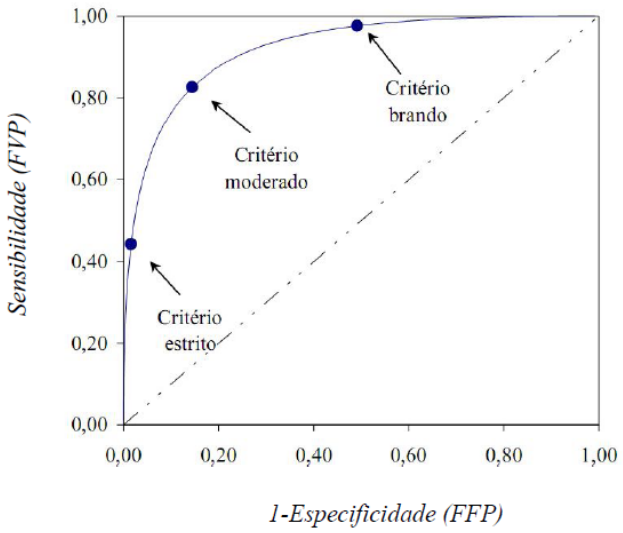

Figura 11. Curva ROC, para uma dada capacidade de discriminação, com a variação do critério de decisão [adaptado de Braga (2000)]

A curva ROC discrimina entre dois estados, onde cada ponto da curva representa um compromisso diferente entre a *Sensibilidade* e o falso positivo que pode ser definido pela adoção de um valor diferente do ponto de corte de anormalidade. Um critério restrito é aquele que traduz uma pequena fração de falsos positivos assim como uma pequena fração de verdadeiros positivos (Braga, 2000).

O valor do ponto de corte é definido com um valor que pode ser selecionado arbitrariamente pelo pesquisador entre os valores possíveis para a variável de decisão, acima da qual o paciente é classificado positivo e abaixo do qual é classificado como negativo.

De acordo com Braga (2000), para cada ponto de corte são calculados valores de *Sensibilidade* e *Especificidade*, estes valores podem assim ser dispostos no gráfico. Um classificador perfeito corresponderia a uma linha horizontal no topo do gráfico, o que é bastante difícil de se obter. Na prática, curvas consideradas boas estarão entre a linha diagonal e a linha perfeita, onde quanto maior a distância da linha diagonal, melhor o sistema. A linha diagonal indica uma classificação aleatória, ou seja, um sistema que aleatoriamente seleciona saídas como positivas ou negativas. Finalmente, a partir de uma curva ROC, devemos poder selecionar o melhor limiar de corte para obtermos o melhor desempenho possível.

Se o objetivo for verificar diferenças entre duas ou mais Curvas ROC, a avaliação é feita através da determinação da área abaixo da curva, usando uma modificação do teste da soma de ordens de *Wilcoxon* para esta comparação. Assim é possível quantificar a exatidão de um teste diagnóstico (proporcional à área abaixo da curva), além da possibilidade de comparar testes diagnósticos.

3.3.6.4 Área abaixo da curva ROC

Quando apresentam a curva ROC, alguns autores optam por apresentar, para o eixo das abcissas, a *Especificidade* em alternativa a *1-Especificidade*, isto não altera a estimativa da área abaixo da curva. Se a curva é ajustada, utilizando-se a teoria pertinente à distribuição normal, a área e o seu desvio padrão podem ser obtidos por recurso aos estimadores de máxima verosimilhança (*Begg*, 1987 citado em Martinez, Neto-Louzada, & Pereira, (2003)).

Analiticamente, a área abaixo da curva ROC pode ser determinada através de:

- Métodos de resolução numérica como, por exemplo, a regra do trapézio;

- Métodos estatísticos: relação com a estatística de Wi*lcoxon-Mann-Witney* (*Hanley*, 1988, citado por Braga (2000)); e estimativa de Máxima Verosimilhança (*Hanley e McNeil*, 1982, citado por Braga (2000)).

3.3.6.5 Comparação de modelos com recurso ao teste da área abaixo da curva ROC

Numa escala comum, os gráficos que representam duas ou mais curvas ROC associadas a diferentes testes diagnósticos contínuos permitem uma imediata comparação de desempenhos (Martinez, Neto-Louzada, e Pereira, 2003).

Salienta-se que, quando se está a comprar duas curvas ROC, pode encontrar-se duas situações distintas:

a) As curvas ROC empíricas são diferentes e não se cruzam, sendo o teste diagnóstico com maior área abaixo da curva aquele que apresenta melhor desempenho;
b) As curvas ROC cruzam-se, as áreas abaixo da curva são próximas mas os testes diagnósticos apresentam desempenhos diferentes.

Um método, para testar se as diferenças entre duas áreas abaixo das curvas ROC provenientes de amostras independentes são significativas, consiste na utilização da razão crítica z, definida por Hanley e McNeil (1983):

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2}} \sim N(0,1)$$

onde A_1 e A_2 correspondem as áreas e SE_1 e SE_2 correspondem aos erros estimados para a curva ROC, respetivamente para os testes diagnósticos 1 e 2. As áreas e os respetivos erros padrão são obtidos através da aproximação à estatística de *Wilcoxon-Mann-Whitney* (Braga, 2000).

Quando os valores da área abaixo da curva ROC são superiores a $0{,}5$, os erros padrão associados às áreas podem ser obtidos através da seguinte expressão:

$$SE(A) = \sqrt{\frac{A(1-A) + (n_A - 1)(Q_1 - A^2) + (n_N - 1)(Q_2 - A^2)}{n_A n_N}}$$

onde Q_1 é referente à probabilidade de duas observações anormais, aleatoriamente escolhidas serem classificadas com maior desconfiança do que uma observação normal, aleatoriamente

escolhida, e Q_2 corresponde à probabilidade de uma observação anormal, aleatoriamente escolhida, ser classificada com maior desconfiança do que duas observações normais aleatoriamente escolhidas. E n_A e n_N correspondem, respetivamente, à dimensão dos pacientes anormais e normais (Braga, 2000).

Capítulo 4

4 Um modelo de análise de dados com base em Análise de *Clusters*, Cartas de Controlo e Regressão Logística para monitorizar o ensino *online*.

4.1 Introdução

Apesar do ensino *online* oferecer benefícios para os estudantes e para os professores, também apresenta alguns desafios importantes. Segundo Levy (2007), Sweet (1986), Tyler-Smith (2005), Xeno et al. (2002), citado por Juan et al. (2009), qualquer curso de ensino à distância apresenta taxas de abandono mais elevadas que os cursos de ensino tradicional. O ensino *online* pode criar a sensação de isolamento dos estudantes. O trabalho de um professor, neste tipo de ensino, não é fácil e absorve muito do seu tempo. Para além de acompanharem minuciosamente todas as atividades realizadas por cada estudante, ainda têm de perceber o tipo de interações que existem entre os estudantes e/ou grupos de estudantes. Nomeadamente, identificarem os estudantes que se baseiam nos trabalhos dos outros colegas para elaborarem os seus, detetarem os estudantes que estão em risco de abandono do curso, perceberem possíveis conflitos dentro dos grupos de trabalho, etc.

É necessário que os professores estejam munidos de ferramentas que lhes possibilitem orientar/apoiar o mais rápido possível os estudantes.

Tal como no ensino tradicional, neste tipo de ensino, um professor deve procurar fazer uma análise descritiva dos dados dos estudantes de forma a conhecê-los melhor. Por exemplo, saber se a turma é formada por mais homens ou mulheres, ter uma noção sobre a faixa etária dos discentes, a sua profissão e a habilitação académica que mais prevalece, etc. Estes conhecimentos permitem ao professor orientar as suas práticas pedagógicas.

Os trabalhos em grupo são uma metodologia muito utilizada neste tipo de ensino. Assim, se os professores agruparem os estudantes mediante algumas características só trará vantagens para eles e para os estudantes. Como por exemplo, formarem grupos de trabalho mediante a performance dos estudantes na plataforma, permitirá aos professores aplicar tarefas diferenciadas e até mesmo enviar-lhes um *email* de alerta para os estudantes que revelem pouca atividade na plataforma.

As cartas de controlo são de uso muito simples mas extremamente eficazes para conseguir o estado de controlo estatístico. Estas fornecem aos professores, informações úteis e de confiança acerca da altura própria para introduzir ações de correção e também quando não devem ser introduzidas alterações.

Quando o processo está sob controlo estatístico, o seu desempenho é previsível. Então ambos, o professor e o estudante, podem confiar que obterão resultados com níveis de qualidade consistente.

Depois de um processo estar sob controlo estatístico, o seu desempenho pode ser continuamente melhorado no sentido de diminuir a variação. Os efeitos esperados para melhorias propostas para o sistema, podem ser antecipadas e os efeitos reais de pequenas alterações podem ser identificadas através dos dados recolhidos pela carta de controlo.

Desta forma, é pertinente analisar as cartas de controlo para a média e para o desvio-padrão, para as variáveis n.º de *posts* no fórum/n.º médio horas *online* por semana e n.º médio horas *online* por semana/nº de leituras no fórum por semana.

Tendo em conta a associação indicada na literatura entre a idade, as habilitações académicas, a nacionalidade e o nº médio de horas que um estudante passa por semana *online*, acrescentando estes pressupostos a outros fatores, surge o interesse do professor conhecer que variáveis apresentam significado preditivo para a conclusão, do estudante, com sucesso de uma determinada unidade curricular do seu curso.

4.2 Objetivos

O objetivo principal é desenvolver um modelo que permita aos professores de cursos de ensino *online* supervisionar de forma eficiente a atividade e o desempenho dos seus estudantes. Este modelo pretende ajudar os professores a:

- conhecer melhor os estudantes que frequentam a sua disciplina;
- classificar os estudantes de acordo com o seu desempenho;
- agrupar os estudantes de acordo com o seu desempenho na plataforma, de modo a permitir uma pedagogia diferenciada;
- acompanhar o desempenho dos estudantes nas tarefas propostas;
- avaliar a significância de algumas características sobre a probabilidade de concluirem o curso com sucesso total ou parcial.

4.3 Funcionamento do modelo

A ferramenta utilizada neste modelo não depende do sistema de gestão de aprendizagem utilizada pelas instituições/universidades. Apenas faz uso dos ficheiros de registo do servidor e/ou de uma base de dados dos registos académicos provenientes do sistema de gestão de aprendizagem, de modo a gerarem relatórios, que podem ser em forma de gráficos ou de mensagem simples. De seguida, esses relatórios aparecerão automaticamente nos perfis dos alunos e/ou dos professores, de acordo com as regras de permissão estipuladas pela universidade, e seguirão via *email* para os endereços dos estudantes. Pois, de acordo com Juan et al. (2009), a geração automática de *emails* para os alunos é uma estratégia importante, uma vez que aumenta a divulgação de informações, permitindo aos estudantes, com pouca ou quase nenhuma assiduidade na plataforma, receber o mais rapidamente possível.

De seguida, apresenta-se um esquema ilustrativo do funcionamento do modelo proposto:

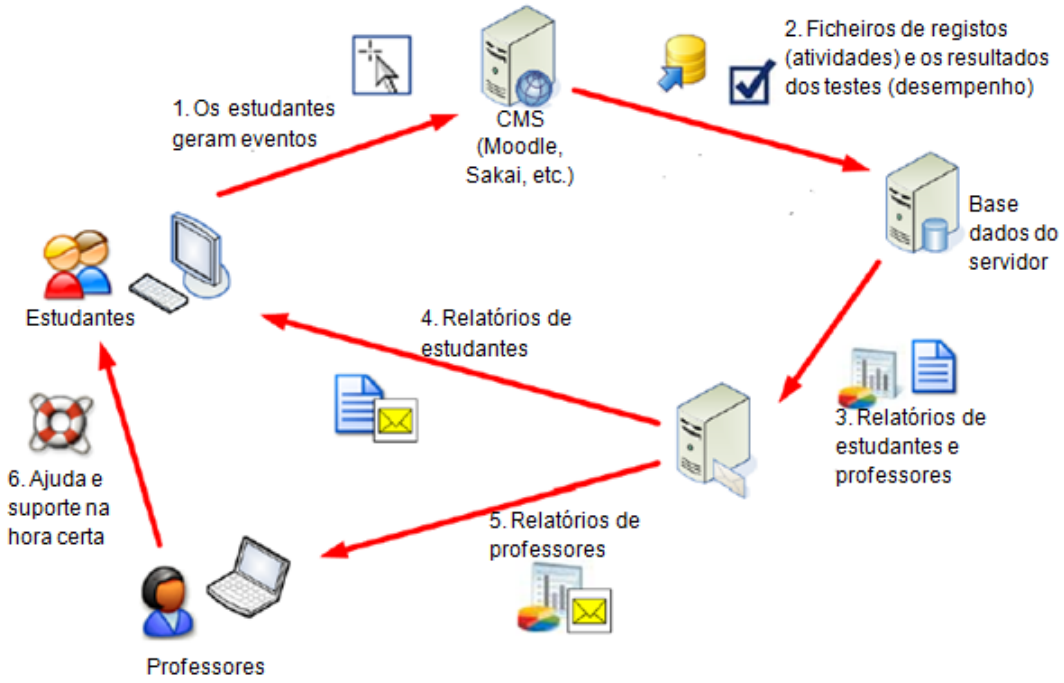

Figura 12. Esquema do funcionamento do modelo proposto [adaptado de Juan et al. (2009)]

4.4 Gráficos de controlo para supervisionar níveis de atividade e desempenho

Partindo do princípio que qualquer professor, mesmo que não tenha grandes conhecimentos sobre análise de dados, poderá usar estes relatórios, Juan et al. (2009) tiveram em conta os seguintes princípios na elaboração dos gráficos:

- devem ser de fácil compreensão;

- devem fornecer, através de uma simples visualização, a informação desejada;

- devem estabelecer uma ligação entre a informação individual e a global e possibilitar o contacto com o estudante ou com um grupo de estudantes com características semelhantes.

Para além destes princípios, na elaboração dos relatórios, deve ter-se alguma atenção à forma como é redigido o relatório para os professores, onde constam a constituição dos vários grupos mediante o desempenho dos estudantes e as características que influenciam a aprovação no curso.

Tipo de Informação	Objetivo	Tipo de Gráfico	Mensagem	Destinatário	Periodicidade
Diagnóstica	Conhecer os dados sociodemográficos dos estudantes.	Gráfico circular; gráfico de barras e diagrama de extremos e quartis.		Professor	No início de cada curso
Atividade: formação de grupos de trabalho	Formar grupos homogéneos de estudantes.		Clustering	Professor	Ao longo do curso
Atividade: classificação dos estudantes	Identificar os estudantes que são susceptíveis de estar "em risco" de abandonar o curso.	Gráfico de dispersão entre o n.º de eventos por aluno durante a semana e o n.º de eventos por aluno durante uma semana média.		Professor	Por semana
Atividade: acompanha-mento individual dos estudantes	Monitorizar os níveis de atividades de cada aluno ao longo do curso.	Carta de controlo de atividade para cada aluno.		Professor	Por semana
Atividade: aompanha-mento dos estudantes na plataforma.	Monitorização dos níveis de desempenho dos estudantes na plataforma.	Carta de controlo do desempenho dos estudantes na plataforma.		Professor e estudante	Ao longo do curso
Atividade: monitorizar o nível de participação	Monitorizar a percentagem de estudantes que terminou o teste.	Gráfico de linhas.		Professor e estudantes	No final de cada teste
Desempenho: distribuição da pontuação de cada teste	Para cada teste que se realize ao longo do curso, obter distribuição estatística das notas dos estudantes.	Histograma, gráfico de barras, gráfico circular e diagrama de extremos e quartis.		Professor e estudantes	No final de cada teste
Desempenho: classificação dos estudantes	Para identificar os estudantes que estão em risco de mau desempenho.	Gráfico de dispersão entre a pontuação do estudante no teste e a pontuação média do estudante.		Professor	No final de cada teste
Desempenho: acompanhamento individual dos estudantes	Monitorização dos níveis de desempenho de cada estudante ao longo do curso.	Carta de controlo de desempenho de cada estudante.		Professor e estudante	No final de cada teste
Desempenho: acompanha-mento dos estudantes	Avaliar a significância de algumas características dos estudantes sobre a probabilidade de terminarem o curso com sucesso.		Regressão Logística Binária	Professor	Ao longo do curso

Tabela 7. Informações importantes contidas no modelo proposto [adaptado de Juan et al. (2009)]

Alguns dos gráficos que constam na tabela 5 já se encontram na maioria dos sistemas de gestão de aprendizagem ou são de fácil implementação. Outros fazem parte da ferramenta de *Educational Data Mining*: SAMOS. Desta forma, seguem os gráficos/relatórios que representam uma contribuição original do modelo proposto.

1. Análise de *Clusters*

A análise de *clusters* sobre as distâncias euclidianas quadradas entre sujeitos com o método de agregação *Ward* produziu o dendograma da figura 13. De acordo com o critério R-squared, foram retidos 3 clusters que explicam 82.1% $(R\text{-}sq = 0.8209)$ da variância total.

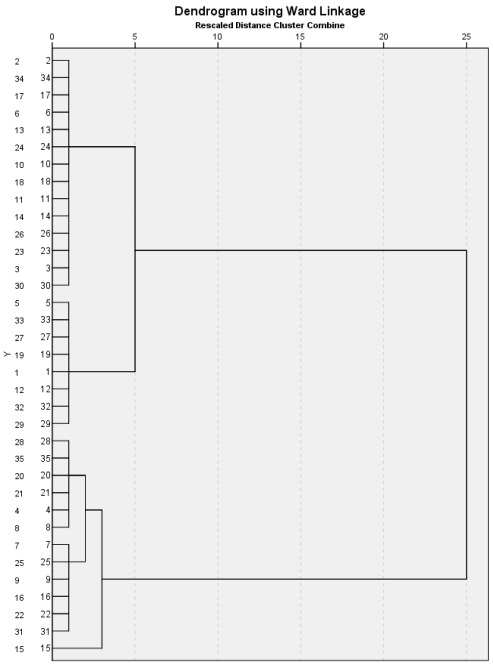

Figura 13. Dendograma Ward

A classificação de cada um dos 35 estudantes na solução refinada com o método k--Means com $k = 3$ encontra-se na tabela 6.

Cluster	N.º Estudante
1	6
	10
	12
	13
	17
	18
	24
	32

Cluster	N.º Estudante
2	4
	7
	8
	11
	15
	20
	21
	22
	25
	31

Cluster	N.º Estudante
3	1
	2
	3
	5
	9
	14
	16
	19
	23
	26
	27
	28
	29
	30
	33
	34
	35

Tabela 8. Classificação dos 35 estudantes em 3 *clusters* pelo método *k-means* com $k = 3$

Na tabela 7, apresentam-se a estatística F para cada variável.

Variável	F
Idade	13.620
N.º *Posts* no Fórum por semana	12.254
N.º Leituras no Fórum por semana	15.030
N.º médio horas *online* por semana	17.333

Tabela 9. Estatística F para cada variável

A variável que, aparentemente, permite diferenciar mais os clusters é a "N.º médio horas online por semana" $(F = 17.3)$, seguida pela "N.º Leituras no Fórum por semana" $(F = 15.0)$ e pela "Idade" $(F = 13.6)$. Finalmente a variável "N.º Posts no Fórum por semana" $(F = 12.3)$ é a variável que diferencia menos os três clusters.

No primeiro *cluster*, o N.º de *Posts* colocados no fórum por semana e o n.º de leituras efetuadas no fórum pelos estudantes por semana e menos importância para a idade. No segundo *cluster*, destaca-se a idade dos estudantes e pouca importância se dá ao N.º de *Posts* colocados no fórum pelos estudantes por semana. No terceiro *cluster*, atribui-se maior importância à idade e menos importância ao N.º de leituras efetuadas no fórum pelos estudantes.

2. Cartas de Controlo

As cartas de controlo são ferramentas importantes que sevem para acompanhar o comportamento/desempenho dos estudantes, servindo para que os professores os monitorizem.

- N.º de *posts* no fórum por semana e nº médio horas *online* por semana

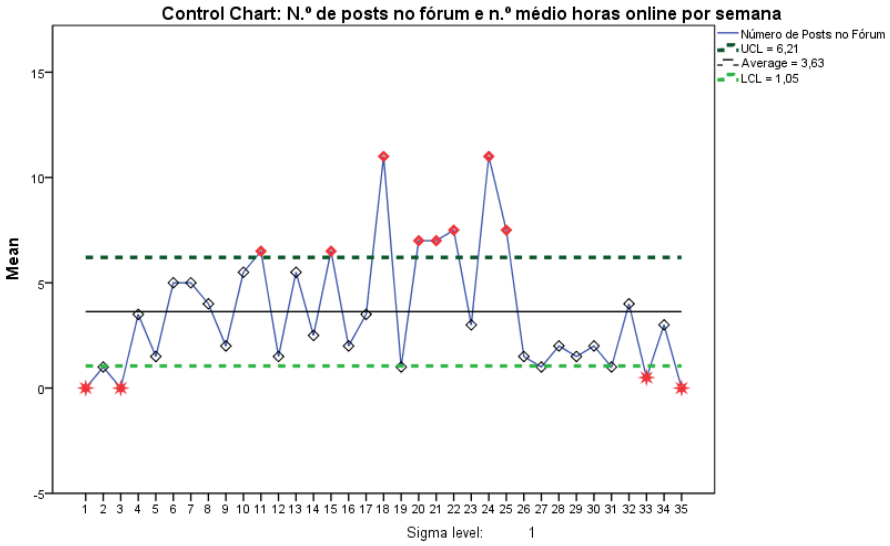

Figura 14. Carta de Controlo para a média – N.º de *posts* no fórum e n.º médio horas *online* por semana

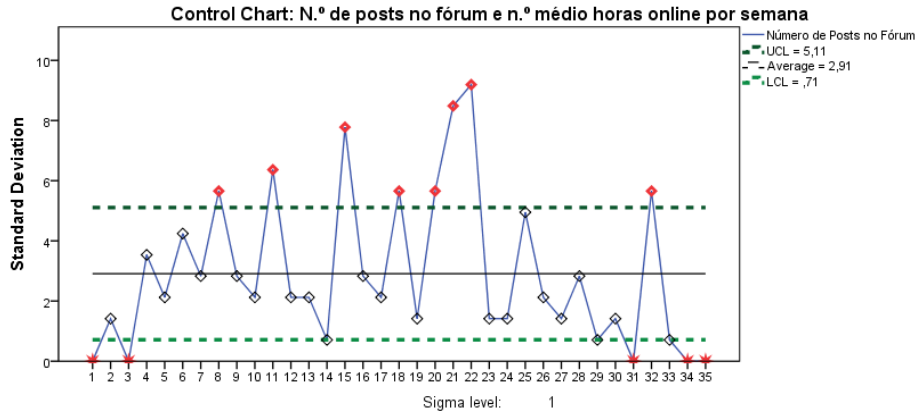

Figura 15. Carta de Controlo para o desvio padrão – N.º de *posts* no fórum e n.º médio horas *online* por semana

A carta de controlo para a média do n.º de posts colocados no fórum e do n.º médio de horas que os estudantes estão online por semana mostra que 23 dos 35 estudantes estão sob controlo estatístico ou seja, acompanhando de forma regular as atividades propostas. No entanto, os estudantes n.ºs 1, 3, 33 e 34 extrapolaram o limite inferior de controlo, pelo que se trata de uma situação preocupante para o professor, uma vez que têm um número de participações no fórum e um número de horas online por semana aquém do desejado. Os estudantes n.ºs 11, 15, 18, 20, 21, 22, 24 e 25 extrapolaram o limite superior de controlo, pelo que se trata de uma situação confortável para o professor, pois significa que estão empenhados nas suas atividades quer pelo número de participações no fórum quer pelo tempo dedicado à realização das mais diversas tarefas.

- N.º médio horas *online* e n.º leituras no fórum por semana

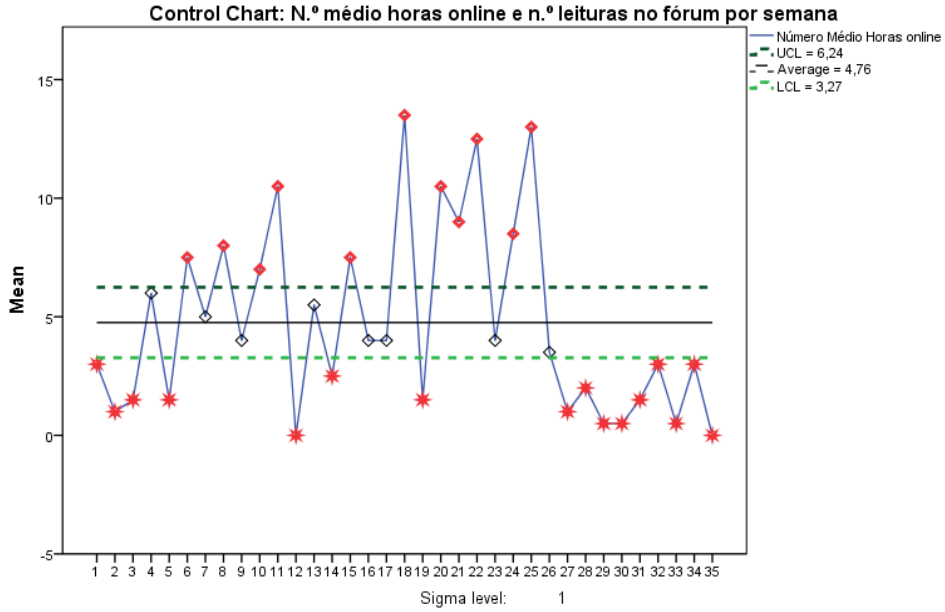

Figura 16. Carta de Controlo para a média – N.º médio horas *online* por semana e n.º leituras no fórum por semana

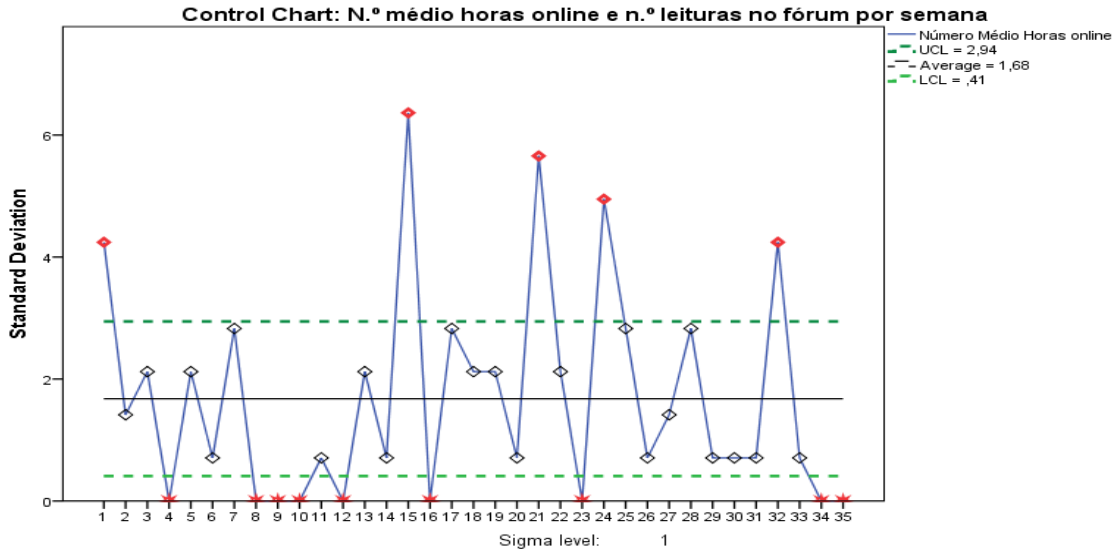

Figura 17. Carta de Controlo para o desvio padrão – N.º médio horas *online* por semana e n.º leituras no fórum por semana

A carta de controlo para a média do n.º médio horas online por semana e n.º leituras no fórum por semana mostra que 9 dos 35 estudantes estão sob controlo estatístico, evidenciando um normal empenho na exploração dos materiais disponibilizados pelo professor na plataforma. No entanto, os estudantes n.ºs 1, 2, 3, 5, 12, 14, 27, 28, 29, 30, 31, 32, 33, 34 e 35 extrapolaram o limite inferior de controlo, pelo que se trata de uma situação preocupante para o professor, já que indica um número insuficiente de consulta de documentos e ainda de um tempo insuficiente de horas dispendidas nessa tarefa. Os estudantes n.ºs 6, 8, 10, 11, 15, 18, 20, 21, 22, 24 e 25 extrapolaram o limite superior de controlo, pelo que se trata de uma situação confortável para o professor, pois significa que estão empenhados nas suas atividades quer ao nível do tempo dispendido, quer ao nível dos materiais consultados.

3. Regressão Logística Binária

A Regressão Logística pelo método "Enter" revelou que as variáveis nacionalidade $\left(b_{nacionalidade} = 1.649; X^2_{Wald}(1) = 0.867,\ p = 0.352, OR = 5.203\right)$ e habilitação académica $\left(b_{habilitação} = -1.766; X^2_{Wald}(1) = 0.958,\ p = 0.328, OR = 0.171\right)$ não apresentam um efeito estatisticamente significativo sobre o Logit da probabilidade de não terminar o curso. Pelo contrário, a variável idade $\left(b_{idade} = -0.413; X^2_{Wald}(1) = 4.828,\ p = 0.028, OR = 0.662\right)$ e o número médio de horas online por semana $\left(b_{Nmédiohorason} = -4.400;\ X^2_{Wald}(1) = 4.666;\ p = 0.031; OR = 0.012\right)$ apresentam um efeito

estatisticamente significativo sobre o Logit da probabilidade de não terminar o curso de acordo com o modelo Logit ajustado $\left(G^2(4)=34.358,\ p<0.001,\ X^2_{HL}=5.129,\ \ p=0.578,\ R^2_N=0.836,\ R^2_{CS}=0.625\right)$. A tabela 8 resume os coeficientes de regressão logística e a sua significância no modelo.

Variável	B	S.E.	X^2_{Wald}	d.f.	p-value	Exp(B)	I.C. a 95% para Exp (B)
Idade	-0.413	0.188	4.828	1	0.028	0.662	]0.458;0.956[
Nacionalidade	1.649	1.771	0.867	1	0.352	5.203	]0.162; 167.431[
Habilitação	-1.766	1.804	0.958	1	0.328	0.171	]0.005; 5.870[
NmédiohorasOn	-4.400	2.037	4.666	1	0.031	0.012	]0.000; 0.665[
Constant	16.735	7.583	4.871	1	0.027	1,853E8	

Tabela 10. Coeficientes *Logit* do modelo de regressão logística da variável "Sucesso" em função da idade, da nacionalidade, das habilitações académicas e do número de horas que os estudantes passam *online* por semana.

Assim, recorrendo ao método *Forward:LR* ajustou-se um novo modelo, estatisticamente significativo $\left(G^2(2)=32{,}756;\ p<0.001;\ H^2_{HL}(6)=0{,}718;\ p=0.994;\ R^2_{CS}=0.608;\ R^2_N=0.812;\right.$ $\left.R^2_{MF}=0.679\right)$ com apenas as variáveis idade $\left(b_{Idade}=-0.409;\ X^2_{Wald}(1)=5.832;\ p=0.016;\ OR=0.664\right)$ e número médio de horas online por semana $\left(b_{NmédiohorasOn}=-3.938;\ X^2_{Wald}(1)=6.019;\ p=0.014;\ \ OR=0.019\right)$.

As funções de probabilidade de não terminar o curso em função da idade e do número médio de horas online por semana são ilustradas na figura 18:

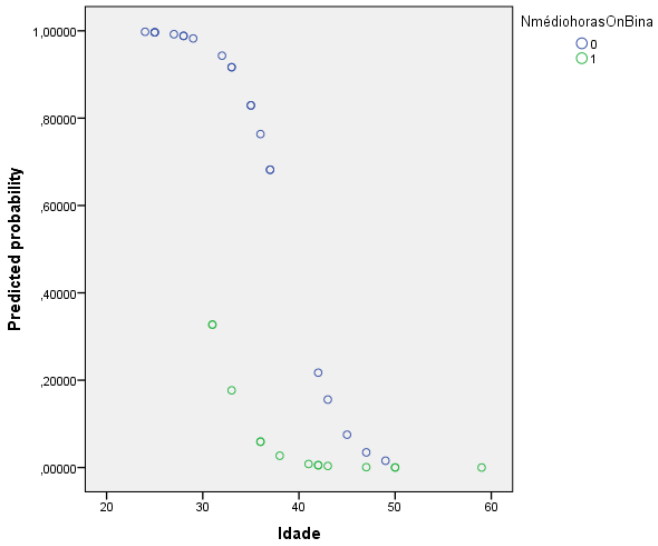

Figura 18. Probabilidade de não terminar o curso em função da idade e o número médio de horas *online* por semana

$$\hat{\pi} = \frac{1}{1+e^{-[15.895-3.938 NmédiohorasON-0.409 idade]}}$$

$\hat{\pi} = \frac{1}{1+e^{-[15.895-3.938-0.409 idade]}}$ para os estudantes que passam mais horas online

e

$\hat{\pi} = \frac{1}{1+e^{-[15.895-0.409 idade]}}$ para os estudantes que passam menos horas online.

A figura 18 apresenta a probabilidade predita para a idade entre os estudantes cujo número médio de horas online por semana de ser menor ou igual a cinco e superior a cinco. Nela, é claramente visível que, para dois estudantes com a mesma idade, a probabilidade de um estudante que passa menos horas online ter sucesso – não terminar o curso - é muito maior do que um estudante que passa mais horas online. Esta diferença é muito significativa principalmente quando os estudantes têm entre os 30 e 45 anos de idade.

O modelo de regressão logística ajustado foi também usado para classificar os estudantes amostrados, tendo-se observado uma percentagem de classificação de 88.6%. O modelo ajustado apresenta ainda elevada sensibilidade (87.5%) e especificidade (89.5%) bem como uma capacidade discriminante excelente $\left(ROC\ c = 0.972; p < 0.001\right)$.

Apresenta-se, de seguida, um quadro-resumo das três técnicas que complementam a ferramenta de *Education Data Mining*: Análise de *Clusters* e Cartas de Controlo e Regressão logística. Apresentam-se ainda algumas alertas que os profissionais da educação *online* poderão ter em conta a fim de melhorar o desempenho dos seus educandos.

Regressão Logística Binária		
Variáveis iniciais	**Variáveis com influência na conclusão do curso**	**Alerta para o professor:**
- Idade; - Nacionalidade; - Habilitação; - N.º médio de horas *online* por semana	- Idade; - N.º médio de horas *online* por semana	Se compararmos dois estudantes da mesma idade, a probabilidade de um estudante que passa menos horas *online* não terminar o curso é muito maior que a de um estudante que passa mais horas *online*.

Análise de *clusters*	**Estratégia**	**Cartas de Controlo**	
Grupo/ Nível de desempenho	**Pedagogia diferenciada:**	- para a média e desvio padrão para o n.º *posts* no fórum/n.º médio de horas *online* por semana	- para a média e desvio padrão para o n.º de leituras no fórum por semana/n.º médio de horas *online* por semana
Grupo 1 Desempenho elevado/risco baixo	- aumentar o grau de dificuldade das atividades, exigir mais, etc. de modo a que os estudantes não desmotivem e que seja possível a aumentar o patamar de exigência.	**6 10 12 13 17 18 24 32**	**6 10 12 13 17 18 24 32**
Grupo 2 Desempenho médio	O professor deve estar atento de forma a que nenhum estudante deste grupo passe para o grupo 3.	**4 7 8 11 15 20 21 22 25 31**	**4 7 8 11 15 20 21 22 25 31**
Grupo 3 Desempenho baixo/ risco elevado	- diminuir o grau de dificuldade das atividades de modo a não levar os estudantes ao abandono.	**1 2 3 5 9 14 16 19 23 26 27 28 29 30 33 34 35**	**1 2 3 5 9 14 16 19 23 26 27 28 29 30 33 34 35**
Nota:	Os n.ºs dos estudantes a **verde** são aqueles que estão fora de controlo estatístico por excesso de desempenho – o professor não se deve preocupar: Os estudantes a vermelho estão fora de controlo estatístico por baixo desempenho – o professor tem de estar em alerta e enviar um *email* de motivação.		
Conclusão:	O professor deve entrar em contacto imediato com aos estudantes n.º 1, 3, 33 e 35, pois correm um elevado risco de não terminarem o curso. Os alunos n.º 2, 5, 14 19, 27, 28, 29, 30 e 34 correm o risco moderado de não terminarem o curso. Daí que o professor deve propor que eles interajam mais entre si, que contactem mais frequentemente o professor e que consultem de forma mais regular os materais propostos para análise.		

Tabela 11. Resumo

Capítulo 5

5. Considerações Finais e Perspetivas Futuras

Neste trabalho, tornou-se evidente que a análise de *clusters*, as cartas de controlo e a regressão logística são três técnicas de *Educational Data Mining* que ajudam a melhorar o ensino *online*, uma vez que constituem uma ferramenta de trabalho para o professor que, em qualquer momento do processo de ensino e aprendizagem, pode analisar as perspectivas de sucesso e insucesso que um estudante tem, tendo para seu benefício uma menor margem de erro minimizando, assim, o número de casos de abandono escolar.

O modelo apresentado poderá ser melhorado. Deste modo, a pesquisa apresentada nesta dissertação deve ser desenvolvida e ampliada ao longo das seguintes linhas:

- fundamentar a Análise de *Clusters* aplicada no modelo com outras análises, como por exemplo, a Análise Discriminante, pois esta última permite calcular as probabilidades de erro associadas às conclusões obtidas;

- aplicar a Análise de *Clusters* com o intuito de formar grupos de estudantes que tenham proximidade geográfica, de modo a permitir-lhes um maior número de contactos presenciais para, desta forma, possibilitar trocas de ideias geradoras de saber; proporcionar ainda um contacto verdadeiramente pessoal onde os discentes poderão ir além da barreira virtual que o próprio computador impõe e viabilizar a troca de desabafos e maior capacidade de se motivarem mutuamente.

- utilizar técnicas de *Learning Analytics* que terão o mesmo objetivo que é o de melhorar o ensino *online* nas suas mais diversas vertentes.

6. Bibliografia

[1] Ayers, E., Nugent, R., Dean, N. (2009). A comparison of student skill knowledge estimates. Em: *International Conference on Educational Data Mining*. Cordoba, Espanha. 1–10.

[2] Arnold, K. E. (2010). Signals: Applying Academic Analytics. *EDUCAUSE Quarterly*. 33 (1). Acedido em 8 de agosto de 2013, em: http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/SignalsApplyingAcademicAnalyti/199385.

[3] Avouris, N., Komis, V., Fiotakis, G., Margaritis, M., Voyiatzaki, E. (2005). Why logging of fingertip actions is not enough for analysis of learning activities. Em: *Workshop on Usage Analysis in Learning Systems, AIED Conference*. Amesterdão, 1–8.

[4] Baker, R., Corbett, A., Roll, I., Koedinger, K. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, **18**(3):287–314.

[5] Baker, R. S. J. D., Yacef, K. (2009). The State of Educational Data Mining in 2009: A review and future visions. *Journal of educational Data Mining*. **1**: 3-17.

[6] Baker, R. (2010). Data Mining for Education. Em: McGaw, B., Peterson, P., Baker, E. (eds.), *International Encyclopedia of Education* (3rd edition), vol. 7. Oxford, Reino Unido: Elsevier. 112-118.

[7] Baker, R., Isotani, S., Carvalho, A. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*. **19** (2): 2-13.

[8] Baker, R. S. J. d., Costa, E., Amorim, L., Magalhães, J., Marinho, T. (2012). Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. *Jornada de Atualização em Informática na Educação*. **1**: 1- 29.

[9] Braga, A. (1994). *Acidente Vascular Cerebral e seus Factores de Risco. Estudo de ocorrência de quatro tipos de AVC.* Tese de Mestrado, Universidade do Minho.

[10] Braga, A. (2000). *Curva ROC: Aspectos funcionais e Aplicações.* Tese de Doutoramento, Universidade do Minho.

[11] Bellaachia, A., Vommina, E. (2006). MINEL: a framework for mining e-learning logs. Em: *Fifth IASTED International Conference on Web-based Education*. México. 259–263.

[12] Bienkowski, M., Feng, M., Means, B. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief*. Washington, D.C.: Office of Educational Technology, U.S. Department of Education Department of Education. EUA.

[13] Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers. *Educational Technology & Society*, 15 (3), 58–76.

[14] D'Mello, S., Craig, S., Witherspoon, A., Mcdaniel, B., Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, **18**(1-2):45–80.

[15] EDUCAUSE (2010) *Next Generation Learning Challenges: Learner Analytics Premises*. Acedido em 10 de julho de 2013, em: http://net.educause.edu/ir/library/pdf/NGLC003.pdf

[16] Elias, T. (2011) *Learning Analytics: Definitions, Processes and Potential*. Acedido em 10 de julho de 2013, em: http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf.

[17] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. Em: *AAAI*, **17**: 37-54.

[18] Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6): 304–317. Acedido em 8 de agosto de 2013, em: http://oro.open.ac.uk/36374/1/IJTEL40501_Ferguson%20Jan%202013.pdf

[19] Frias-Martinez, E., Chen, S., Liu, X. (2006). Survey of data mining approaches to user modeling for adaptive hypermedia. Em: *IEEE Trans Syst Man Cybern.* Parte C, vol. **36**, nº 6, 734–749.

[20] Garcia, E., Romero, C., Ventura, S., Castro, C. (2009). Collaborative data mining tool for education. Em: *International Conference on Educational Data Mining*. Cordoba, Espanha, 299–306.

[21] Gomes, M.I., Figueiredo, F., Barão, M.I. (2010). *Controlo Estatístico da Qualidade*. Edições Sociedade Portuguesa de Estatística. 2ª edição.

[22] Graf, S., Ives, C., Rahman, N., Ferri, A. (2011). AAT: a tool for accessing and analysing students' behaviour data in learning systems. Em: *First International Conference on Learning Analytics and Knowledge. Banff, Alberta, Canada, 174–179.*

[23] Grobelnik, M., Mladenic, D., Jermol, M., (2002). Exploiting text mining in publishing and education. Em: *Proceedings of the ICML Workshop on Data Mining Lessons Learned*. Sydney, Australia. 34–39.

[24] Hershkovitz, A., Nachmias, R. (2008) Developing a log-based motivation measuring tool. Em: *First International Conference on Educational Data Mining*. Montreal, Canada, 226–233.

[25] Horizon Project Shortlist (2012) *NMC Horizon Project Short List*, 2012 Higher Education Ed. Acedido em 10 de julho de 2013, em: http://www.nmc.org/news/download-horizon-project-2012-higher-ed-short-list.

[26] Hosmer, D. J., Lemeshow, S. (1989). *Applied Logistic Regression.* Copyright by John Wiley & Sons, Inc.

[27] Johnson, L., R. Smith, H. Willis, A. Levine, Haywood, K. (2011). Learning Analytics. *The 2011 Horizon Report.* Austin, Texas: The New Media Consortium. Acedido em 7 de agosto de 2013, em: http://net.educause.edu/ir/library/pdf/HR2011.pdf.

[28] Johnson, M., Barnes, T. (2009). EDM Visualization Tool: Watching Students Learn. Em: *Third International Conference on Educational Data Mining*. Pittsburgh, PA; 297-298.

[29] Jovanovic, J., Gasevic, D., Brooks, C., Devedzic, V., Hatala, M. (2007). LOCO-Analyist: a tool for raising teacher's awareness in online learning environments. Em: *European Conference on Technology-Enhanced Learning*. Creta, Grécia; 112–126.

[30] Juan, A., Daradoumis, T., Faulin, J., Xhafa, F. (2009). SAMOS: a model for monitoring students' and groups' activities in collaborative e-learning. Em: *Int. J. Learn Technol* , **4**: 53–72.

[31] Juan, A., Daradoumis, T., Faulin, J., Xhafa, F. (2009). A data analysis model based on control charts to monitor online learning processes. Em: *Int. J. Business Intelligence and Data Mining*, 4: 159-174.

[32] Juan, A., Minguillón, J., Huertas, A., Sancho, T. (2011). Computer-supported statistics courses in online environments: adding e-repositories to the equation. Em: *Int. J. Teaching and Casa Studies*, 3: 16-34.

[33] Kay, J., Maisonneuve, N., Yacef, K., Zaiane, O.R. (2006). Mining Patterns of Events in Students' Teamwork Data. Em: *Proceedings of Educational Data Mining Workshop*. Taiwan. 1-8.

[34] Kim, J., Chern, G., Feng, D., Shaw, E., Hovy, E., (2006). Mining and Assessing Discussions on the Web through Speech Act Analysis. Em: *Proceedings of the AAAI Workshop on Web Content Mining with Human Language Technologies*. Atenas. 1-8.

[35] Koedinger, K., Cunningham, K., Skogsholm, A., Leber, B. (2008). An open repository and analysis tools for finegrained, longitudinal learner data. Em: *First International Conference on Educational Data Mining*. Montreal, Canada, 157–166.

[36] Lera-López, F., Faulin, J., Juan, A., Cavaller, V. (2010). Monitoring Students´Activity and Performance in Online Higher Education: A European Perspective. Em: Juan, A.,

Daradoumis, T., Xhafa, F., Caballe, S., Faulin, J. (eds.), *Monitoring and Assessement in Online Collaborative Environments: Emergent Computational Technologies for E-Learning Support*, Information Science Reference. Hershey.

[37] Lu, J. (2004). Personalized e-learning material recommender system. Em: *International conference on information technology for application*. Utah, USA. 374–379.

[38] Markellou, P., Mousourouli, I., Spiros, S., & Tsakalidis, A. (2005). Using semantic web mining technologies for personalized e-learning experiences. Em: *Proceedings of the web-based education*. Grindelwald, Suíça. 461–826.

[39] Maroco, J. (2007). *Análise Estatística com utilização do SPSS*. 3ª edição, Edições Sílabo. Lisboa.

[40] Martins, P. S. (2008). *Análise estatística de performance de um conjunto de testes auditivos.* Tese de Mestrado, Universidade de Aveiro.

[41] Mazza R., Milani C. (2004). GISMO: a graphical interactive student monitoring tool for course management systems. Em: *International Conference on Technology Enhanced Learning*. Milan, Italia. 1–8.

[42] Merceron A., Yacef K. (2005). Educational data mining: a case study. Em: *International Conference on Artificial Intelligence in Education*. Amesterdão. 1–8.

[43] Misso, F., Jacobi, L. F. (2007). *Variáveis dummy:especificações de modelos com parâmetros variáveis.* Brasil: Ciências e Natura, UFSM.

[44] Montgomery, D.C. (2005). *Introduction to Statistical Quality Control*. 5ª edição. John Wiley & Sons, Inc. USA. pp. 194–311; 645–706.

[45] Montgomery, D.C. (2009). *Introduction to Statistical Quality Control*. 6ª edição. John Wiley & Sons, Inc. USA.

[46] Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., Heiner, C. (2005) An educational data mining tool to browse tutor-student interactions: time will tell! Em: *Proceedings of the Workshop on Educational Data Mining*. Amesterdão. 15–22.

[47] Pahl, C., Donnellan, C., (2003). Data mining technology for the evaluation of web-based teaching and learning systems. Em: *Proc.Congress E-learning.* Montreal, Canada. 1-7.

[48] Pavlik, P., Cen, H., Koedinger, K. (2009). Learning factors transfer analysis: using learning curve analysis to automatically generate domain models. Em: Barners, T., Desmarais, M., Romero, C. & Ventura, S. (Eds.), Em: *2nd Internacional Conference Education Data Mining.* Cordoba, Espanha.121–130.

[49] Pedraza-Perez, R., Romero, C., Ventura, S. (2011). A Java desktop tool for mining Moodle data. Em: *International Conference on Educational Data Mining*. Eindhoven, Holanda, 319–320.

[50] Pestana, M., Gageiro, J. (2005). *Análise de dados para Ciências Sociais- A Complementaridade do SPSS.* 4ª Edição, Edições Sílabo. Lisboa.

[51] Rabbany, R., Takaffoli, M., Zaiane, O. (2011). Analyzing participation of students in online courses using social network analysis techniques. Em: *International Conference on Educational Data Mining*. Eindhoven, Holanda, 21–30.

[52] Reis, E. (2001). *Estatística Multivariada Aplicada*. 2ª edição, Edições Sílabo. Lisboa.

[53] Romero C, Ventura S, De Bra P. (2004) Knowledge discovery with genetic programming for providing feedback to courseware author. *User Model User-Adapted Interac,*. **14**:425–464.

[54] Romero, C., Ventura, S., Garcia, E. (2008). Data Mining in course management systems: Moodle case study and tutorial. *Computers & Education*, **51**: 368-384.

[55] Romero, C., Ventura, S., Zafra, A., De Bra, P. (2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computer & Education*, **53**: 828–840.

[56] Romero, C., Ventura, S. (2011). Preface to the special issue on data mining for personalised educational systems. *User Model User-Adapted Interact*, **21:**1–3.

[57] Romero, C., Ventura, S. (2013). Data Mining in Education. *Data Mining Know Discov*, **3**: 12-27

[58] Rumel, D. (1986). Odds ratio-algumas considerações. Em: *Revista de Saúde Publica*, **20** (3): 253-258. Brasil: Departamento de Epidemiologia da Faculdade de Saúde Publica da Universidade de São Paulo.

[59] Runger, G.C., Montgomery, D.C. (1993) Adaptive Sampling Enhancements for Shewhart Control Charts. Em: *IIE Transactions in Quality and Reliability Engineering*, 25: 41–51.

[60] Scheur, O., McLaren, BM. (2011). Educational data mining. Em: *The Encyclopedia of the Sciences of Learning*, Springer. Nova Iorque.

[61] Siemens, G., Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE review*, **46** (5).

[62] Siemens, G. (2012). Learning Analytics: Envisioning a Research Discipline and a Domain of Practice. Athabasca University.

[63] Siemens, G., Baker, RSJd. (2012). Learning analytics and educational data mining: towards communication and collaboration. Em: *Proceeding of the 2nd Internacional Conference on Learning Analytics and Knowlwdge.* Vancouver, British Columbia, Canada. 1-3.

[64] SPSS Inc. (2007). *SPSS Regression 17.0.* Acedido em 10 janeiro de 2014, em http://www.helsinki.fi/~komulain/Tilastokirjat/IBM-SPSS-Spec-Regression.pdf.

[65] Tang, T. & McCalla, G. (2005). Smart Recommendation for an Evolving E-Learning System: Architecture and Experiment. *International Journal on E-Learning*, *4*(1): 105-129. Norfolk, VA: AACE. Acedido em 13 de agosto de 2013, em: http://www.editlib.org/p/5822.

[66] Tsung, F. (2000). Statistical Monitoring and Diagnosis of Automatic Controlled Processes Using Dynamic PCA. *International Journal of Production Research*, **38** (3): 625-637.

[67] Ueno, M. (2004). Data mining and text mining technologies for collaborative learning in an ILMS ”Samurai”. Em: *IEEE International Conference on Advanced Learning Technologies*. Joensuu, Finlândia. 1052-1053.

[68] Woodall, W.H. (2008). The Use of Control Charts in Health-Care and Public-Health Surveillance. *Journal of Quality Technology*, **38**: 89-104.

[69] Zaïane, O., & Luo, J. (2001). Web usage mining for a better web-based learning environment. Em: *Proceedings of conference on advanced technology for education.* Banff, Alberta. 60–64.

[70] Zaïane, O. (2002). Building a recommender agent for e-learning systems. Em: *Proceedings of the International Conference in Education*. Auckland, Nova Zelândia. **1**: 55-59.

http://mineracaodedados.wordpress.com/tag/ferramentas/

http://isi.cbs.nl/glossary/bloken00.htm

http://portalaction.com.br

7 Anexos

A - Questionário

Educational Data Mining e *Learning Analitics* na melhoria do ensino *online*

Muito obrigada desde já pela sua colaboração!

Se está a ler este texto é porque decidiu participar neste estudo.

O presente questionário tem como objetivo conhecer a atividade académica dos estudantes que frequentam o ensino *online*. Além deste objetivo, pretende-se relacionar essa atividade com características sociodemográficas.

A sua participação tomará apenas alguns minutos do seu tempo e consiste no preenchimento deste breve questionário. As respostas são totalmente confidenciais e em nenhum momento do questionário precisa de se identificar.

Caso o questionário lhe suscite alguma dúvida poderá contactar-me através do telemóvel 962907286 e do *email* susanafaria76@gmail.com (não precisando de se identificar).

Por favor, responda a este questionário apenas uma vez.

Grata pela sua atenção e participação!

Código: _____

1. DADOS SOCIODEMOGRÁFICOS

1.1 SEXO:

◯ Feminino

◯ Masculino

1.2 IDADE:

☐ (anos)

1.3 NACIONALIDADE:

◯ Portuguesa

◯ Brasileira

◯ Angolana

◯ Outra

1.4 ESTADO CIVIL / SITUAÇÃO CONJUGAL:

◯ Solteiro

◯ Casado

◯ União de Facto

◯ Divorciado

◯ Separado

◯ Viúvo

1.5 HABILITAÇÕES ACADÉMICA (NÍVEL DE ENSINO MAIS ELEVADO QUE COMPLETOU)

◯ Ensino Secundário

◯ Ensino pós secundário *(cursos de especialização tecnológica, nível IV)*

◯ Bacharelato

◯ Licenciatura

◯ Mestrado

◯ Doutoramento

1.6 SITUAÇÃO PROFISSIONAL

◯ Estudante

◯ Reformado

◯ Desempregado

◯ Empregado. Profissão: ☐

1.7 RENDIMENTO MENSAL LÍQUIDO INDIVIDUAL (DO RESPONDENTE): _____€

2. ATIVIDADE ACADÉMICA

2.1 NÚMERO DE *POSTS* EFETUADOS NO FÓRUM POR SEMANA: _____

2.2 NÚMERO DE LEITURAS EFETUADAS NO FÓRUM POR SEMANA: _____

2.3 NÚMERO MÉDIO DE *UPLOADS* EFETUADOS POR SEMANA: _____

2.4 NÚMERO MÉDIO DE *DOWNLOADS* EFETUADOS POR SEMANA: _____

2.5 NÚMERO MÉDIO DE HORAS *ONLINE* POR SEMANA: _____

2.6 NÚMERO MÉDIO DE HORAS QUE ESTÁ NA PLATAFORMA POR SEMANA: _____

O seu questionário terminou! Muito obrigada pela sua colaboração!

B - Análise Estatística

Os dados foram tratados com recurso ao programa estatístico IBM – SPSS 20 (IBM Corporation, New York, USA). A análise descritiva dos dados é apresentada através de tabelas de frequências, gráficos de barras e gráficos circulares. Para as variáveis quantitativas é determinado o valor máximo, valor mínimo, desvio padrão e variância. Inicialmente, foi realizada uma filtragem nos dados para detetar valores perdidos e anormais que pudessem influenciar os resultados. Esta análise será evidenciada na forma descritiva dos resultados.

Análise de Clusters

Para agrupar os estudantes em grupos homogéneos relativamente à idade, n.º médio de horas *online*, n.º de *posts* e n.º de leituras no fórum por semana, recorreu-se à análise de *clusters*. Visto não ser possível designar um melhor critério de agregação dos casos em análise (Reis, 2001), proceder-se-á a uma conjugação de critérios. A análise começará pelo procedimento de agregação hierárquico através do método *Ward* e os grupos serão comparados de forma a verificar a existência de diferenças significativas através do R^2. No final, será aplicado o agrupamento não hierárquico *k-means*. Esta ordem exploratória dos dados está de acordo com o descrito por Maroco (2007).

Cartas de Controlo

A fim de se verificar se o processo está sob controlo estatístico recorreu-se à elaboração de cartas de controlo para a média e desvio padrão para:
- o n.º de *posts* no fórum e n.º médio horas *online* por semana;
- o n.º médio horas *online* e n.º leituras no fórum por semana.

Regressão Logística Binária

Para avaliar a significância da idade, da nacionalidade, das habilitações académicas e do número de horas que os estudantes passam *online* por semana sobre a probabilidade de obterem sucesso, foi aplicada regressão logística pelo método Forward: LR (Maroco, 2007). Os pressupostos foram analisados através da análise gráfica dos resíduos, análise de diagnóstico dos casos influentes e testando a área da curva ROC.

C - Resultados

C1 - Análise Descritiva e Preparação dos Dados

Começa-se por analisar os dados que terão de ser considerados como perdidos para o problema em causa, assim como assimetria nas distribuições e observações atípicas. Os pressupostos de aplicabilidade já foram abordados no método.

Os dados utilizados neste estudo foram obtidos através das respostas ao questionário que se encontra no Anexo A. A este questionário responderam 35 estudantes com idades compreendidas entre os 25 e 59 anos (média = 37 (desvio padrão = 9)), sendo 16 do sexo feminino (45,7%) e 19 do sexo masculino (54,3%).

- **Sexo:**

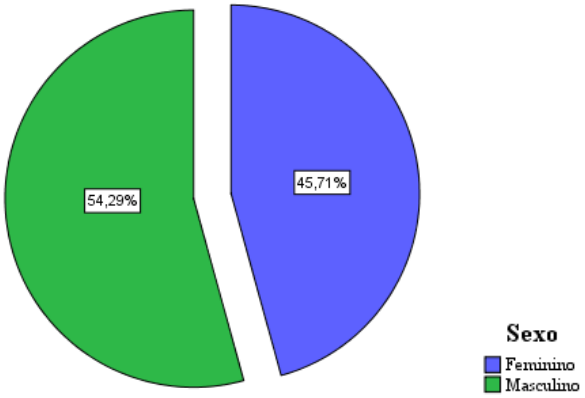

Figura 19. Gráfico circular da variável Sexo

- **Idade:**

Tabela 12. Descriptive Statistics

	N	Min	Max	Mean	Std. Dev.	Variance
Idade	35	24	59	36,97	8,535	72,852

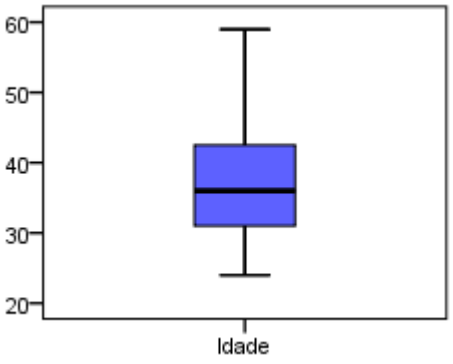

Figura 21. Diagrama de Extremos e Quartis das Idades

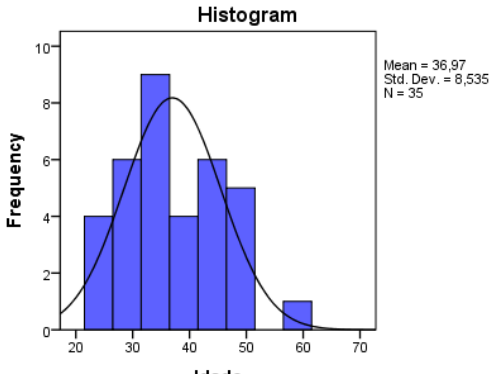

Figura 20. Histograma das Idades

- **Localidade:**

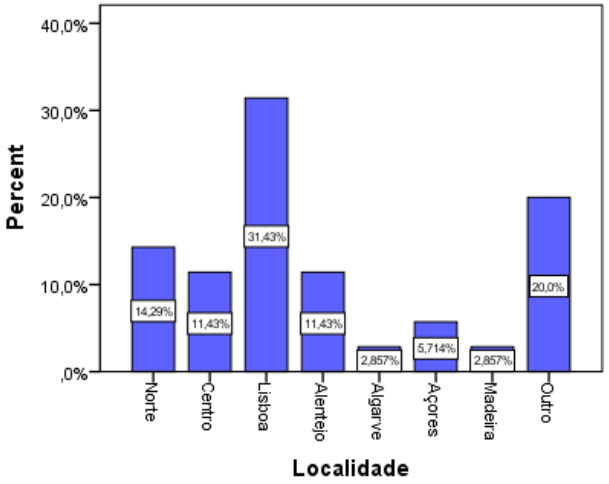

Figura 22. Gráfico de Barras da Localidade

- **Habilitação:**

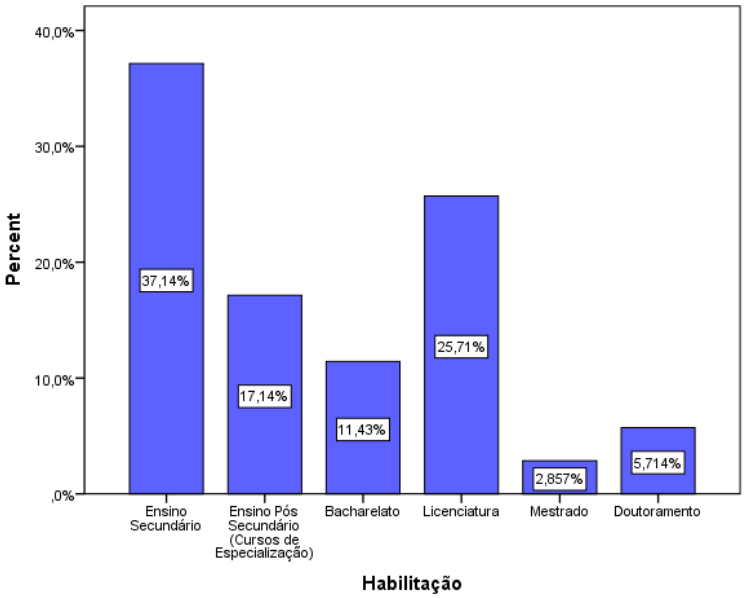

Figura 23. Gráfico de Barras da Habilitação

C2 - Análise de Clusters

Para agrupar os estudantes em grupos homogéneos relativamente à Idade, n.º médio de horas *online*, n.º de *posts* e n.º de leituras no fórum por semana recorreu-se à análise de clusters. Visto não ser possível designar um melhor critério de agregação dos casos em análise (Reis, 2001), proceder-se-á a uma conjugação de critérios. A análise comecerá pelo procedimento de agregação hierárquico através do método Ward e os grupos serão comparados de forma a verificar a existência de diferenças significativas através do R^2. No final, será aplicado o agrupamento não hierárquico *k-means*. Esta ordem exploratória dos dados está de acordo com o descrito por Maroco (2007).

Método Hierárquico

No método hierárquico será aplicado o critério de ***Ward*** porque os clusters são calculados de forma a minimizar a soma dos quadrados dos erros, sendo retidos, ao longo dos passos realizados, dentro dos possíveis, aqueles que apresentam menor soma do quadrado dos erros (Maroco, 2007). Face aos dados, outro critério poderia alterar significativamente a distribuição de indivíduos ente grupos (tal foi verificado) havendo um grupo com um número muito elevado de estudantes e os outros com quase nenhuma inclusão. O critério Ward é baseado na perda de informação que resulta do agrupamento de estudantes, sendo medida a soma dos quadrados dos desvios das observações individuais em relação às médias dos grupos (Reis, 2001).

Será colocada no SPSS, como medida de distância, a euclidiana, por ser uma das mais utilizadas e fazer parte do procedimento a realizar após o cálculo das médias. Na opção para clusters naturais será colocado o número de sete.

Colocando no SPSS, vamos obter o seguinte esquema de aglomeração expresso pela tabela 11.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	34	,500	0	0	14
2	5	33	1,000	0	0	13
3	14	26	1,500	0	0	5
4	1	12	2,000	0	0	16
5	14	23	2,833	3	0	22
6	28	35	3,833	0	0	26
7	7	25	4,833	0	0	28
8	20	21	5,833	0	0	9
9	4	20	6,833	0	8	18
10	10	18	7,833	0	0	19
11	6	13	8,833	0	0	20
12	3	30	10,333	0	0	22
13	5	27	11,833	2	0	17
14	2	17	13,333	1	0	25
15	9	16	15,333	0	0	24
16	1	32	17,500	4	0	21
17	5	19	19,750	13	0	29
18	4	8	22,250	9	0	26
19	10	11	25,250	10	0	27
20	6	24	28,917	11	0	25
21	1	29	32,750	16	0	29
22	3	14	36,717	12	5	27
23	22	31	41,217	0	0	24
24	9	22	47,467	15	23	28
25	2	6	55,133	14	20	30
26	4	28	65,633	18	6	31
27	3	10	76,708	22	19	30
28	7	9	90,458	7	24	31
29	1	5	116,333	21	17	33
30	2	3	173,911	25	27	33
31	4	7	276,161	26	28	32
32	4	15	456,411	31	0	34
33	1	2	799,909	29	30	34
34	1	4	2548,800	33	32	0

Tabela 13. Agglomeration Schedule

Na primeira linha temos a informação que o primeiro *cluster* a ser formado contém os estudantes n.º 2 e n.º 34. Na coluna dos coeficientes, temos a informação sobre as distâncias ou, como neste caso, sobre a soma dos quadrados do erro. No passo seguinte, os estudantes n.º 5 e n.º 33 foram agrupados e continuando até ao final. Por exemplo, no passo 13, o estudante n.º 27 foi agrupado ao par (5, 33). Na coluna *Next Stage*, é possível observar estas passagens de ligação entre os passos, assim como, nas colunas que a precedem, surge a indicação sobre o estágio onde aquela combinação apareceu pela primeira vez.

O dendograma representa o esquema de algomeração com as fases desse processo, mas com os coeficiente escalonados entre 0 e 25, ao invés de 0 a 2548.8 como na tabela 11. No dendograma, podemos aplicar uma linha de corte da distância ficando com os clusters abaixo dessa linha. O dendograma parece apontar para um número de 3 clusters *grosso modo* ou então para um número bem mais elevado. Resta saber se o número de 7 pode ser considerado

satisfatório, ou seja, se a divisão num novo grupo não introduz alterações significativas no coeficiente de fusão (Reis, 2001).

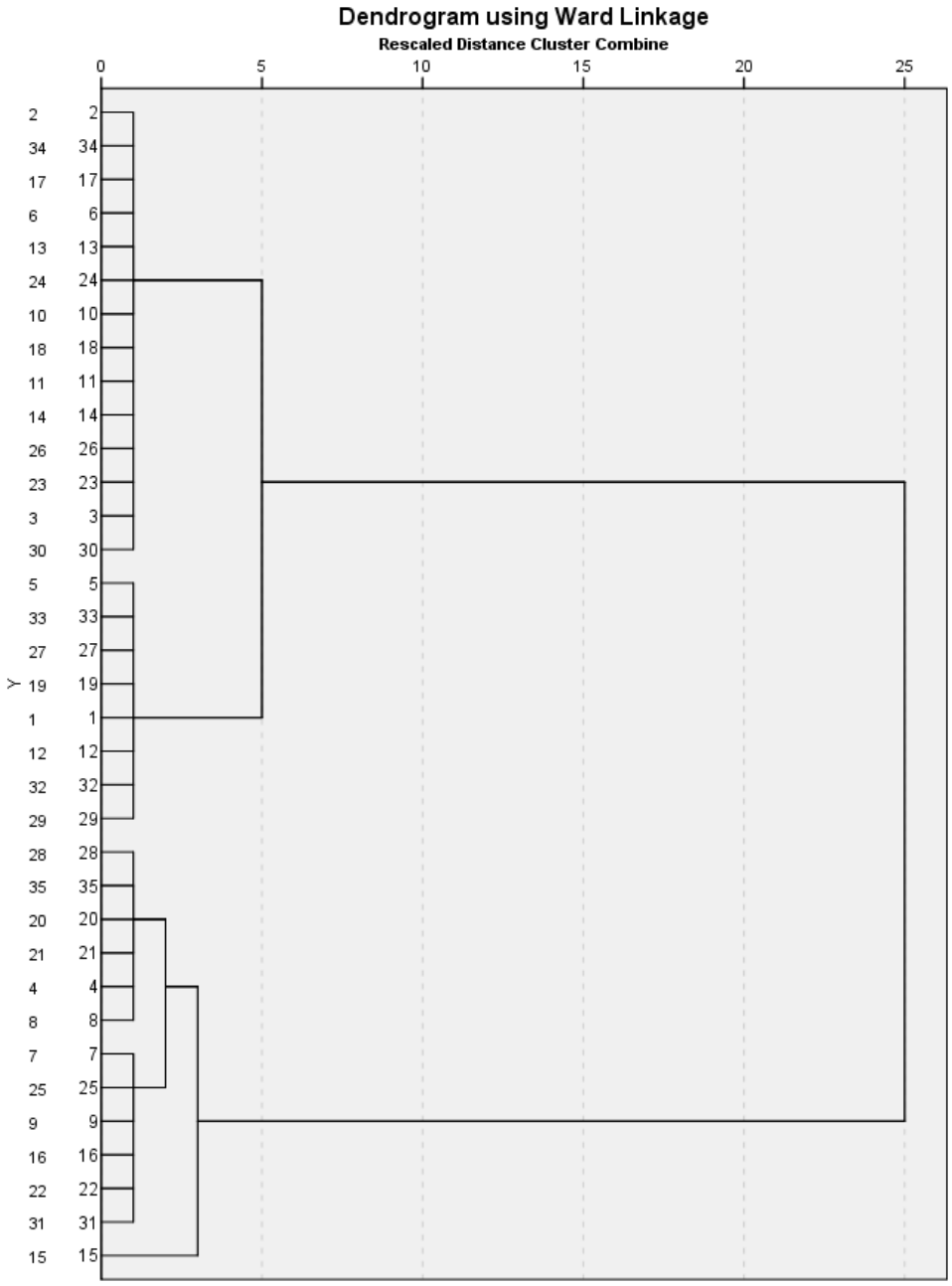

Figura 24. Dendograma Ward

Comparação critério R^2

O R^2 é uma medida de percentagem da variabilidade total retida em cada uma das soluções de *clusters* (Maroco, 2007). Pelo que é importante encontrar um número mínimo de

clusters com uma percentagem significativa da variabilidade total, recorrendo à ANOVA *one-way* para a variável CLUX_1, em que o X representa o número de *clusters* pretendido e vai variar entre 1 e 12. Será colocado o exemplo para CLU9_1 sendo os restantes procedimentos similares.

		Sum of Squares	df	Mean Square	F	Sig.
Idade	Between Groups	2451,888	8	306,486	317,686	,000
	Within Groups	25,083	26	,965		
	Total	2476,971	34			
N.º *Posts* no Fórum por semana	Between Groups	4,221	8	,528	1,726	,140
	Within Groups	7,950	26	,306		
	Total	12,171	34			
N.º Leituras no Fórum por semana	Between Groups	7,005	8	,876	1,902	,103
	Within Groups	11,967	26	,460		
	Total	18,971	34			
N.º Médio Horas *Online* por semana	Between Groups	20,052	8	2,507	3,158	,012
	Within Groups	20,633	26	,794		
	Total	40,686	34			

Tabela 14. ANOVA CUL9_1

Como $R^2 = \frac{SQC}{SQT}$, então $R_9^2 = \frac{2451.88 + 4.221 + 7.005 + 20.052}{2476.971 + 12.171 + 18.971 + 40.686} = 0.9742$.

Que na totalidade de CLU12_1, CLU11_1, CLU10_1, CLU9_1, CLU8_1, CLU7_1, CLU6_1, CLU5_1, CLU4_1, CLU3_1, CLU2_1, CLU1_1 resume-se ao seguinte quadro:

N.º Clusters	R^2
1	0
2	0.6864
3	0.8209
4	0.8917
5	0.9318
6	0.9544
7	0.9645
8	0.9699
9	0.9742
10	0.9784
11	0.9814
12	0.9838

Tabela 15. Número de clusters e R^2

Segundo Maroco (2007), interessa encontrar um número mínimo de *clusters* cuja percentagem seja significativa em relação à variabilidade total, dando como exemplo 80% (Maroco, 2007). Neste caso, através da tabela 13, podemos observar que este número apenas é

conseguido a partir do 3.º *cluster*. Na figura 25, verificamos que o ponto intersecional das curvas está perto dos 0.8 o que justificaria 3 *clusters*.

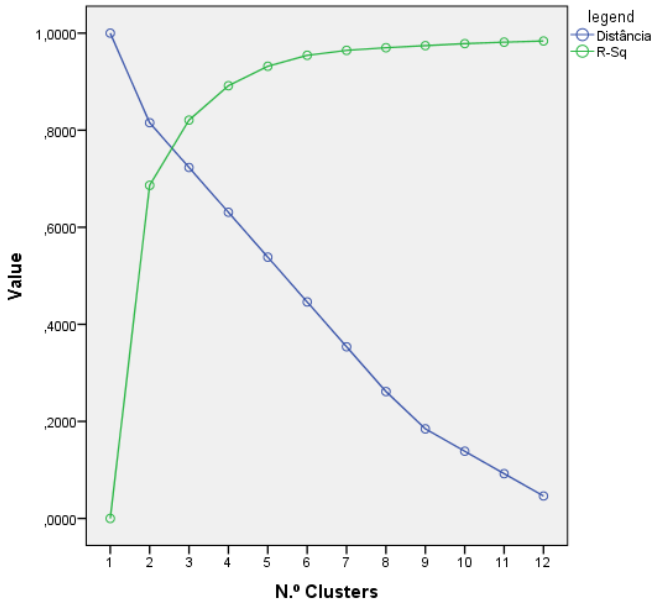

Figura 25. Relação R-Sq e distâncias

Após estudar como se distribuem as variáveis para a análise de *clusters*, vamos proceder à utilização do método não hierárquico *k-means*.

Método não hierárquico *k-means*

Agora procedemos à aplicação direta do número de grupos determinado *a priori*, mas já conhecendo como os indivíduos se relacionam em relação aos grupos. Parte-se do princípio que existe o interesse de realizar três grupos mas conhecendo as limitações desta aplicação. A tabela 14 apresenta as iterações que evidenciam a variação do centro dos *clusters*, terminando no passo 4.

Tabela 16. Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	1,891	1,812	1,265
2	,307	,202	,193
3	,154	,000	,071
4	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 4. The minimum distance between initial centers is 4,404.

Serão omissos os resultados da tabela "*Cluster Membership*" que nos indica quanto cada observação é semelhante ao centro e cada variável em relação a cada cluster,

respetivamente. Já a tabela 15 fornece informação sobre a distância entre os centróides e os *clusters*.

Tabela 17. Distances between Final Cluster Centers

Cluster	**1**	**2**	**3**
1		2,027	2,422
2	2,027		2,581
3	2,422	2,581	

Na tabela 16, temos a ANOVA *one-way* que nos fornece alguma informação já esperada, nomeadamente, em relação à variável *N.º Posts no Fórum por semana*. Na coluna *Cluster Mean Square*, aspeto fundamental neste caso da ANOVA, podemos verificar qual ou quais as variáveis que permitem a separação dos *clusters*. O *N.º Posts no Fórum por semana* nas ANOVAS por *cluster* na aglomeração hierárquica tinha sido o único a iniciar sem diferenças entre grupos e agora verifica-se que o seu C*luster Mean Square* é o mais baixo e o *Erro Mean Square* o mais elevado. Logo não é uma variável discriminativa.

Tabela 18. ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore: **Idade**	7,817	2	,574	32	13,620	,000
Zscore: **N.º Posts no Fórum por semana**	7,373	2	,602	32	12,254	,000
Zscore: **N.º Leituras no Fórum por semana**	8,234	2	,548	32	15,030	,000
Zscore: **N.º Médio Horas *Online* por semana**	8,840	2	,510	32	17,333	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Situação contrária observa-se para o *N.º Médio Horas Online por semana*, sendo a variável que apresenta a variabilidade entre *clusters* mais elevada (8.840) e sendo a menor distância dentro dos *clusters*. Desta forma, uma das possibilidades seria retirar a variável *N.º de Posts no fórum por semana* como discriminante na formação de grupos. Outro aspeto seria verificar a relevância no aumento do número de *clusters*. Na tabela 17 podemos ver a distribuição dos estudantes pelos grupos.

Tabela 19. Number of Cases in each Cluster

Cluster	1	8,000
	2	10,000
	3	17,000
Valid		35,000
Missing		,000

Em relação às características dos grupos, face ao sinal das médias apresentadas na tabela 18, podemos caracterizar os grupos em função do grau de importância de cada aspeto medido pelas variáveis em causa.

Tabela 20. Final Cluster Centers

	Cluster		
	1	2	3
Zscore: **Idade**	-,69961	1,02264	-,27232
Zscore: **N.º Posts no Fórum por semana**	1,00878	,21489	-,60113
Zscore: **N.º Leituras no Fórum por semana**	,77454	,57374	-,70198
Zscore: **N.º Médio Horas *Online* por semana**	,26772	,95333	-,68677

Destaca-se, no primeiro grupo, o N.º de *Posts* colocados no fórum por semana e o n.º de leituras efetuadas no fórum pelos estudantes por semana e menos importância para a idade. No segundo grupo, destaca-se a idade dos estudantes e pouca importância se dá ao N.º de *Posts* colocados no fórum pelos estudantes por semana. No terceiro grupo, atribui-se maior importância à idade e menos importância ao N.º de leituras efetuadas no fórum pelos estudantes.

O primeiro grupo é constituído pelos estudantes n.ºs: 6, 10, 12, 13, 17, 18, 24 e 32. O segundo grupo é formado pelos estudantes n.ºs: 4, 7, 8, 11, 15, 20, 21, 22, 25 e 31. O terceiro e último grupo é composto pelos estudantes n.ºs: 1, 2, 3, 5, 9, 14, 16, 19, 23, 26, 27, 28, 29, 30, 33, 34 e 35.

C3 – Regressão Logística Binária

- Para avaliar a significância da idade, da nacionalidade, das habilitações académicas e do número de horas que os estudantes passam *online* por semana sobre a probabilidade de obterem sucesso – não terminarem o curso - recorreu-se à regressão logística pelo método *Forward:LR* como descrito em Maroco (2007). Procedeu-se também à validação dos pressupostos por intermédio da análise gráfica dos resíduos e ao diagnóstico de casos influentes.

As 35 observações foram incluídas para análise (ver tabela 19). A possibilidade que se pretende modelar é a de um estudante não terminar o curso. De seguida, é possível visualizar a codificação das variáveis categoriais, assim como a frequência para as várias possibilidades, assim como, os *outputs* de base a estes resultados. Para este estudo, a possibilidade de "sucesso" é a de um estudante "não terminar o curso".

Tabela 21. Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	35	100,0
	Missing Cases	0	,0
	Total	35	100,0
Unselected Cases		0	,0
Total		35	100,0

a. If weight is in effect, see classification table for the total number of cases.

Pela Tabela 20, observa-se que caso não soubéssemos nada acerca das nossas variáveis, ou seja, caso arbitrariamente todos os sujeitos estivessem na condição de terminar o curso, então a taxa de acerto seria de 54,3%. A previsão de casos de não terminar o curso é de 45,7% $\left[(16\times100)/35\right]$. Há 19 sujeitos que terminam o curso e 16 que não terminam o curso.

Tabela 22. Classification Table[a,b]

	Observed		Predicted		
			Sucesso/Insucesso		Percentage Correct
			Termina o curso	Não termina o curso	
Step 0	Sucesso/Insucesso	Termina o curso	19	0	100,0
		Não termina o curso	16	0	,0
	Overall Percentage				54,3

a. Constant is included in the model.
b. The cut value is ,500

Na Tabela 21, apresenta os resultados apenas com a constante incluída no modelo, antes de qualquer coeficiente ser introduzido na equação. A regressão logística compara este modelo com o modelo que inclui todas as variáveis, a fim de determinar se o modelo posterior

é o mais apropriado. Se apenas existisse a constante o modelo não era significativo (*p*=0,613).

Tabela 23. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-,172	,339	,257	1	,613	,842

Pela Tabela 22, observam-se as variáveis não introduzidas na equação, a qual informa quais as variáveis capazes de contribuir para a melhoria do modelo. A resposta é SIM para as variáveis idade, HabilitacaoBina e NmédiohorasOnBina, as quais são estatisticamente significativas $(p < 0.05)$, pelo que a sua inclusão parece representar um poder preditivo aumentado para o modelo. A estatística *Overall Statistics* testa a hipótese nula se os coeficientes das variáveis que não constam na equação são iguais a zero. Pela Tabela 22, observa-se que o *Score*=22,322 e *p*<0,001 pelo que a adição de uma ou mais variáveis ao modelo irá afectar significativamente o poder preditivo do modelo.

Tabela 24. Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Idade	16,774	1	,000
		NacionClasses	1,281	1	,258
		HabilitacaoBina	4,609	1	,032
		NmédiohorasOnBina	13,988	1	,000
	Overall Statistics		22,322	4	,000

Nas tabelas 23, 24 e 25, podemos observar que o modelo pelo método "Enter" apresenta uma significância elevada $\left(G^2(4) = 34.358,\ p < 0.001,\ X^2_{HL} = 5.129,\right.$
$\left. p = 0.578,\ R^2_N = 0.836,\ R^2_{CS} = 0.625\right)$. Desta forma, é possível concluir que existe pelo menos uma variável independente no modelo com poder preditivo em relação ao Sucesso.

Pela Tabela 23, observa-se o modelo χ^2. A significância global é testada no SPSS pelo *Model Chi-square*, o qual é derivado para verosimilhança da observação dos dados actuais sob o pressuposto que o modelo foi considerado ajustado e preciso. Neste âmbito existem duas hipóteses:

$H_0 =$ o modelo final é um modelo com bom ajuste

vs

$H_1 =$ o modelo final não é um modelo com bom ajuste (i.e., os preditores têm um efeito significativo)

Tabela 25. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	34,358	4	,000
	Block	34,358	4	,000
	Model	34,358	4	,000

Pela Tabela 24, ainda que na regressão logística não exista nenhuma estatística análoga ao coeficiente de determinação R^2, o modelo fornece algumas aproximações. *Cox & Snell R Square* tenta copiar o R^2 múltiplo baseado na verosimilhança, mas como o seu valor máximo é geralmente inferior a 1.0, o que torna esta estatística de difícil interpretação. Neste estudo, é indicativo que 62,5% da variação da variável dependente é explicada pelo modelo logístico. A estatística modificada de *Nagelkerke* varia entre 0 e 1, sendo uma medida mais confiável de relação. A estatística *Nagelkerke R Square*, normalmente produz valores mais elevados do que *Cox & Snell R Square*. Neste estudo, *Nagelkerke* $R^2 = 0.836$ indicando a uma relação boa (83,6%) entre os preditores e a predição, ou seja, cerca de 83,6% do sucesso/insucesso é explicado pelo modelo. Por último, o valor $-2LL$ (-2 log likelihood) é de 13,905. Segundo Hair et al. (2010), ao contrário do R^2 de *Cox & Snell* e do R^2 de *Nagelkerke* onde a melhoria do modelo é traduzida por valores elevados, já um modelo $-2LL$ possui bom ajuste com valor pequeno, sendo zero o valor mínimo. Esta estatística ($-2LL$) é semelhante à soma dos erros ao quadrado, pelo que indica quanta informação não está explicada após o modelo ter sido ajustado. Valores baixos, como aquele que está registado neste estudo (13,905) significam um bom ajuste, uma vez que quanto menor for, maior será a percentagem de observações explicadas.

Tabela 26. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13,905[a]	,625	,836

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

Pela Tabela 25, surge o teste de *Hosmer & Lemeshow* (HL), o qual assume a adequação da amostra, com a *regra do polegar* "*rule of thumb*" tendo casos suficientes, de tal modo que 95% das células têm frequência esperada superior a 5. A estatística *Step* indica a melhoria do modelo em cada iteração. Neste estudo apenas resultou em uma etapa. Neste estudo, o HL, com uma probabilidade (p) determinada a partir da distribuição χ^2, com 7 graus de liberdade no teste de ajuste do modelo, tem valor $p = 0.644$ o que significa que não existem evidências de significância estatística $(p > 0.05)$ entre os valores observados e os previstos, tal como é pretendido para o *goodness-of-fit test statistic* (teste estatístico de um bom ajuste),

pelo que podemos considerar que modelo se ajusta bem aos dados. Com isto, percebe-se que o resultado desejável de não significância indica que o modelo de predição não difere significativamente do observado.

Tabela 27. Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5,129	7	,644

Tabela 28. Classification Table[a]

Observed			Predicted		
			Sucesso/Insucesso		Percentage Correct
			Termina o curso	Não termina o curso	
Step 1	Sucesso/Insucesso	Termina o curso	17	2	89,5
		Não termina o curso	3	13	81,3
	Overall Percentage				85,7

a. The cut value is ,500

Na tabela 26 temos a informação que a sensibilidade do modelo é de 81.3% (classifica corretamente 81.3% dos estudantes que não terminam o curso) e a sua especificidade é de 89.5% (classifica corretamente 85.7% dos estudantes que terminam o curso).

Através da tabela 27 podemos observar quais as variáveis que são significativas para a equação do modelo. Segundo o teste de *Wald* verificamos que as variáveis idade $\left(b_{idade} = -0.413; X^2_{Wald}(1) = 4.828, p = 0.028, OR = 0.662\right)$ e o número médio de horas *online* por semana $\left(b_{NmédiohorasOn} = -4.400;\ X^2_{Wald}(1) = 4.666;\ p = 0.031; OR = 0.012\right)$ são significativas na discriminação. Pelo contrário, podemos verificar que as variáveis nacionalidade $\left(b_{nacionalidade} = 1.649; X^2_{Wald}(1) = 0.867, p = 0.352, OR = 5.203\right)$ e habilitação académica $\left(b_{habilitação} = -1.766; X^2_{Wald}(1) = 0.958, p = 0.328, OR = 0.171\right)$ não são significativas para o modelo.

A Tabela 27 contém importantes dados. A estatística de *Wald* (com distribuição χ^2) e as probabilidades associadas fornecem um indicador da significância de cada preditor na equação, ou seja, informa acerca da significância de cada coeficiente (i.e., se cada coeficiente é significativamente diferente de zero). Desta forma, se cada coeficiente é claramente diferente de zero, então pode assumir-se que a variável contribui significativamente para a previsão de Y.

A forma de analisar a estatística de *Wald* é observar o nível de significância de cada coeficiente logístico (B), procurando-se identificar o quanto a variável independente participa

individualmente na explicação da variável dependente, sendo que para $\alpha = 0.05$ serão aceites as variáveis independentes com $p < 0.05$ rejeitando-se a hipótese nula dada a forte contribuição desta(s) para o modelo. Pelos resultados obtidos, para um erro tipo I, verifica-se que a idade $(p = 0.028)$, NmédiohorasOn $(p = 0.031)$ e a constante $(p = 0.027)$ contribuem significativamente para a predição; as restantes variáveis não têm contributo significativo $(p > 0.05)$.

A coluna Exp(B) apresenta a extensão para elevar a medida correspondente por uma unidade de influência dos *Odds ratio*. É interpretado em termos de mudança na probabilidade. Se o valor excede 1, qualquer aumento no preditor conduz a um aumento na probabilidade do resultado ocorrer. A coluna mais à direita apresenta o intervalo de confiança a 95% para a Exp(B).

A coluna S.E. refere-se ao erro padrão de cada coeficiente.

A coluna B refere-se aos coeficientes logísticos que podem ser usados para a equação de predição (semelhante à equação de regressão linear).

Probabilidade final estimada: $P(Y) = \frac{1}{1+e^{-(16{,}735-0{,}413*\beta_1+1{,}649*\beta_2-1{,}766*\beta_3-4{,}400*\beta_4)}}$

A equação do modelo de regressão logística é dado por:

$$Ln(\pi/1-\pi) = 16.735 + (-0.413 \times idade) + (1.649 \times nacionalidade) + (-1.766 \times habilitação) + \\ + (-4.4 \times NmédiohorasOn)$$

De notar que a não significância das variáveis Nacionalidade e Habilitação, podem ser justificação para as eliminar da equação, pois não parecem melhorar a previsão de Y.

Variável	B	S.E.	X^2_{Wald}	d.f.	p-value	Exp(B)	I.C. a 95% para Exp (B)
Idade	-0.413	0.188	4.828	1	0.028	0.662	]0.458;0.956[
Nacionalidade	1.649	1.771	0.867	1	0.352	5.203	]0.162; 167.431[
Habilitação	-1.766	1.804	0.958	1	0.328	0.171	]0.005; 5.870[
NmédiohorasOn	-4.400	2.037	4.666	1	0.031	0.012	]0.000; 0.665[
Constant	16.735	7.583	4.871	1	0.027	1,853E8	

Tabela 29. Coeficientes *Logit* do modelo de regressão logística da variável "Sucesso" em função da idade, da nacionalidade, das habilitações académicas e do número de horas que os estudantes passam *online* por semana segundo o modelo "Enter".

Assim, ao observarmos as tabelas 28, 29 e 30 do *output* para o método *Foward Stepwise (Likelihod Ratio)*, podemos verificar que foram realizados dois passos, ambos significativos, em que o segundo passo será o mais significativo e ajustado aos dados.

Tabela 30. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	23,002	1	,000
	Block	23,002	1	,000
	Model	23,002	1	,000
Step 2	Step	9,754	1	,002
	Block	32,756	2	,000
	Model	32,756	2	,000

Tabela 31. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	25,261[a]	,482	,644
2	15,507[b]	,608	,812

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.
b. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

Tabela 32. Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1,772	6	,939
2	,718	6	,994

$$\left[G^2(2) = 32{,}756;\ p < 0.001;\ H_{HL}^2(6) = 0{,}718;\ p = 0.994;\ R_{CS}^2 = 0.608;\ R_N^2 = 0.812;\ R_{MF}^2 = \right.$$
$$\left. = 1 - \frac{\ln(L_c)}{\ln(L_0)} = 1 - \frac{15.507}{15.507 + 32.756} = 1 - \frac{15.507}{48.263} = 0.679 \right]$$

Na tabela 31 podemos verificar as variáveis que entram no modelo de forma significativa, comparativamente com a análise anterior do método "Enter". Entre elas a idade $\left(b_{Idade} = -0.409;\ X_{Wald}^2(1) = 5.832;\ p = 0.016;\ OR = 0.664\right)$ e número médio de horas *online* por semana $\left(b_{NmédiohorasOn} = -3.938;\ X_{Wald}^2(1) = 6.019;\quad p = 0.014;\ OR = 0.019\right)$.

		B	S.E.	X^2_{Wald}	d.f.	p-value	Exp(B)	I.C. a 95% para Exp (B)
Passo 1	**Idade**	-0.339	0.114	8.823	1	0.003	0.712	]0.569;0.891[
	Constante	11.967	4.046	8.749	1	0.003	1.6E5	
Passo 2	**Idade**	-0.409	0.169	5.832	1	0.016	0.664	]0.477; 0.926[
	NmédiohorasOn	-3.938	1.605	6.019	1	0.014	0.019	]0.001; 0.453[
	Constant	15.895	6.371	6.225	1	0.013	8.0E6	

Tabela 33. Coeficientes *Logit* do modelo de regressão logística da variável "Sucesso" segundo o modelo "Forward Stepwise: LR".

Apesar destas variáveis apresentarem um poder discriminativo significativo, assim como o próprio modelo em relação ao anterior, a sensibilidade do modelo é de 87.5% e a sua especificidade é de 89.5%.

A figura 26 apresenta a probabilidade predita para a idade entre os estudantes cujo número médio de horas *online* por semana é menor ou igual a cinco e superior a cinco. Nela é claramente visível que, para dois estudantes com a mesma idade, a probabilidade de um estudante que passa menos horas *online* ter sucesso – não terminar o curso - é muito maior do que um estudante que passa mais horas *online*. Esta diferença é muito significativa principalmente quando os estudantes têm entre os 30 e 45 anos de idade.

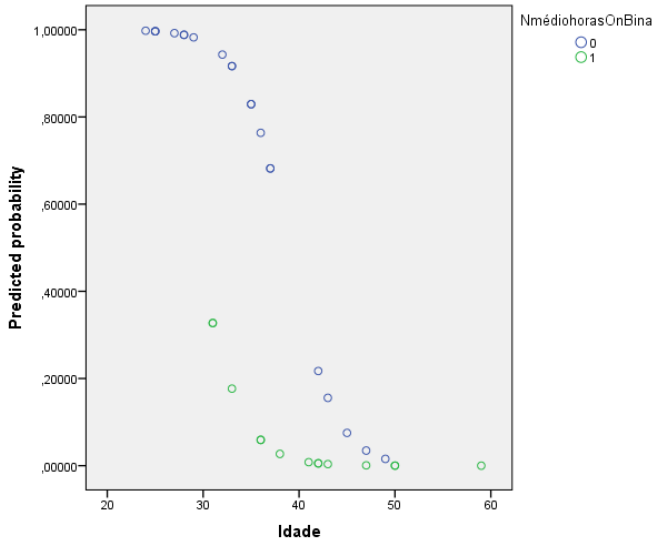

Figura 26. Probabilidade predita para a idade mediante o número médio de horas *online* por semana

Na figura 27 podemos observar a análise de resíduos e de observações influentes, para percebermos a existência de *outliers* que influenciem os resultados. É possível observar que não existem valores com $|r_j| > 2$ que sejam candidatos a *outliers*.

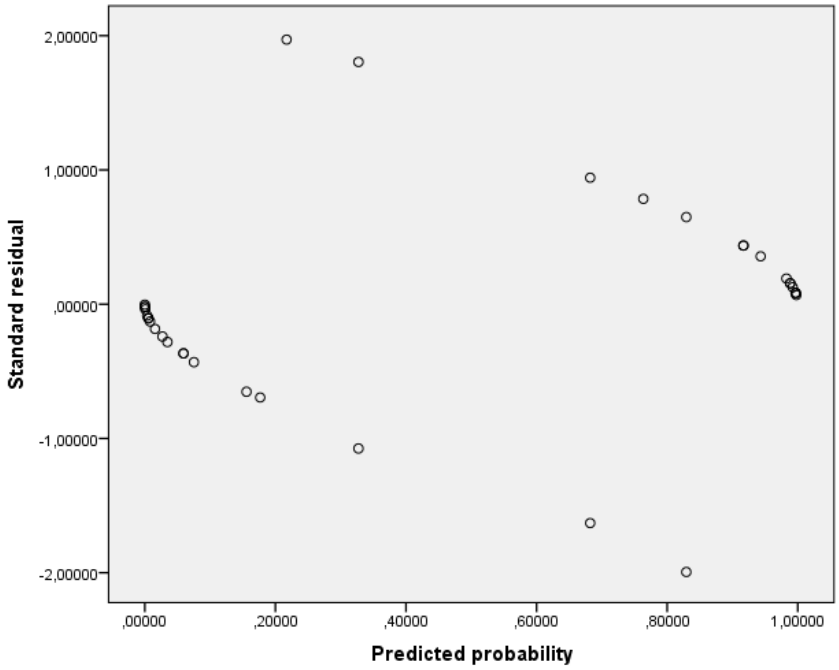

Figura 27. Probabilidade predita para a idade mediante o número médio de horas *online* por semana

Através da curva ROC (figura 28) podemos verificar que o modelo tem uma capacidade discriminante bastante elevada, pois quando testamos: $H_0 : c = 0.5$ vs $H_1 : c > 0.5$ e avaliando a probabilidade de ocorrência de não terminar o curso, vamos rejeitar a hipótese nula $(c = 0.972; p < 0.001)$ (conforme tabela 31).

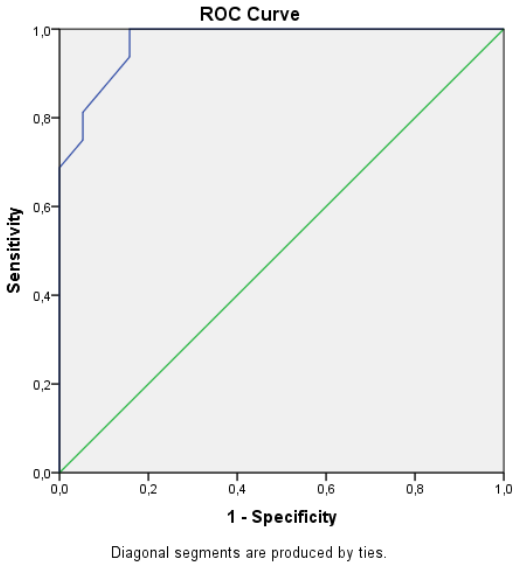

Figura 28. Curva ROC

Tabela 34. Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,972	,022	,000	,928	1,000

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

As funções de probabilidade encontradas são descritas da seguinte forma:

$$\hat{\pi} = \frac{1}{1+e^{-[15.895-3.938 NmédiohorasOn-0.409idade]}};$$

$\hat{\pi} = \frac{1}{1+e^{-[15.895-3.938-0.409idade]}}$ para os estudantes que passam mais horas *online*

e

$\hat{\pi} = \frac{1}{1+e^{-[15.895-0.409idade]}}$ para os estudantes que passam menos horas *online*.