Luis Pedro Albano Ramos

Licenciado em Eng. Informática

Detecção de Padrões de Desempenho Académico no Ensino Superior

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática

Orientador : João Moura Pires, Prof. Auxiliar, Universidade Nova de Lisboa

Júri:

Presidente: Presidente do Júri

Arguente: Primeiro Arguente

Vogal: Primeiro Vogal

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Março, 2012

Detecção de Padrões de Desempenho Académico no Ensino Superior

For Pony

Agradecimentos

Queria desde já agradecer à instituição que me formou, a Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa, por possibilitar a minha graduação nesta área que tanto gosto. Obrigado ao meu orientador, Prof. João Moura Pires, pelo débito de conhecimento e por me guiar neste longo caminho. Um obrigado grande também aos meus colegas de curso, que estiveram sempre lá para ouvir os meus desabafos. Agradeço também ao meu pai e à minha mãe pela força e exemplo que me permitiu chegar aqui. Obrigado a toda a equipa da Bliss Applications, em particular ao André Gil, pela nova perspectiva sem a qual este documento não existiria.
Em especial, um grande obrigado à Vanessa Tavares, por tudo.

Resumo

O paradigma de avaliação do ensino superior foi alterado em 2005 para ter em conta, para além do número de entradas, o número de alunos diplomados. Esta alteração pressiona as instituições académicas a melhorar o desempenho dos alunos. Um fenómeno perceptível ao analisar esse desempenho é que a performance registada não é nem uniforme nem constante ao longo da estadia do aluno no curso. Estas variações não estão a ser consideradas no esforço de melhorar o desempenho académico e surge motivação para detectar os diferentes perfis de desempenho e utilizar esse conhecimento para melhorar a o desempenho das instituições académicas.

Este documento descreve o trabalho realizado no sentido de propor uma metodologia para detectar padrões de desempenho académico, num curso do ensino superior. Como ferramenta de análise são usadas técnicas de data mining, mais precisamente algoritmos de agrupamento. O caso de estudo para este trabalho é a população estudantil da licenciatura em Eng. Informática da FCT-UNL.

Propõe-se dois modelos para o aluno, que servem de base para a análise. Um modelo analisa os alunos tendo em conta a sua performance num ano lectivo e o segundo analisa os alunos tendo em conta o seu percurso académico pelo curso, desde que entrou até se diplomar, transferir ou desistir. Esta análise é realizada recorrendo aos algoritmos de agrupamento: algoritmo aglomerativo hierárquico, *k-means*, SOM e SNN, entre outros.

Palavras-chave: Data-Mining, Clustering, Agrupamento, Análise ao Ensino Superior, Melhorar Desempenho Académico, k-Means, SOM, SNN, Algoritmo Agrupamento Hierárquico

Abstract

The paradigm behind the evaluation of higher education has changed in 2005 to take into account the number of graduate students, as well as the number of new students. This change pressures the higher education institutions to improve their student's performance. A noticeable phenomenon analyzing students is that said performance is neither uniform nor constant throughout the student's stay in the course. This variations are not being used in the effort to improve academic performance and motivation arises to detect the different performance profiles and use that knowledge to improve the performance of higher education institutions.

This document describes the work done to propose a methodology to detect patterns of academic performance in a higher education course. As a analysis tool, data-mining techniques are used, more precisely, clustering algorithms. The case study for this work is the student population of the degree in Computer Science of FCT-UNL.

Two student models are proposed, that serve has the base of the analysis. The first model takes into account the students' performance in a academic year. The second model takes into account their academic "route" trough the course, since they entered until they graduate, transfer or give up. This analysis makes use of the following clustering algorithms: hierarchical agglomerative clustering, k-means, SOM and SNN, among others.

Keywords: Data-Mining, Clustering, Higher Education Analisys, Improve Academic Performance, k-Means, SOM, SNN, Hierarchical Clustering Algorithm

Conteúdo

Lista de Figuras

Lista de Tabelas

Listagens

1

Introdução

Este capítulo apresenta a motivação para o presente trabalho e esclarece o contexto em que se insere. Os objectivos que se pretendem atingir são sumariamente apresentados na secção 1.2 e na secção 1.3 é apresentada, na generalidade, a estratégia seguida. A secção 1.4 descreve os dados que vão ser utilizados nesta análise. A última secção (1.5) apresenta a estrutura do documento.

1.1 Motivação e Contexto

Esta secção apresenta a motivação para esta dissertação, assim como o contexto em que se desenrola.

O ensino superior público alterou em 2005 o paradigma de avaliação do desempenho das instituições académicas do ensino superior para fins do cálculo do financiamento das instituições [Rep05]. O financiamento, que depende de valores como o número de novas entradas na universidade e da qualidade do corpo docente, passa a ter em conta o número de alunos que conclui os estudos [Rep06]. Esta alteração força as instituições académicas a ter em conta questões como a empregabilidade dos seus alunos e a sustentabilidade da própria instituição nos seus modelos de ensino, forçando também a uma re-avaliação do desempenho dos seus alunos, na perspectiva de o melhorar. Este desempenho é analisado através de indicadores tais como o tempo médio para obtenção do diploma, a taxa de abandono e a classificação média dos alunos, que oferecem uma visão global do desempenho da população estudantil. Qualquer esforço de melhorar o desempenho académico passa pelo estudo dessa população e pela análise da qualidade dos indicadores utilizados.

Os dados presentes na figura 1.1 mostram o número de entradas e de saídas do ensino superior, ou seja, o número de alunos que se matricularam pela 1ª vez e o número de alunos que terminaram o seu curso, desde o ano lectivo de 2006/2007 até 2011/2012[1]. Observa-se que o número de entradas e saídas no ensino superior tem aumentado, mas

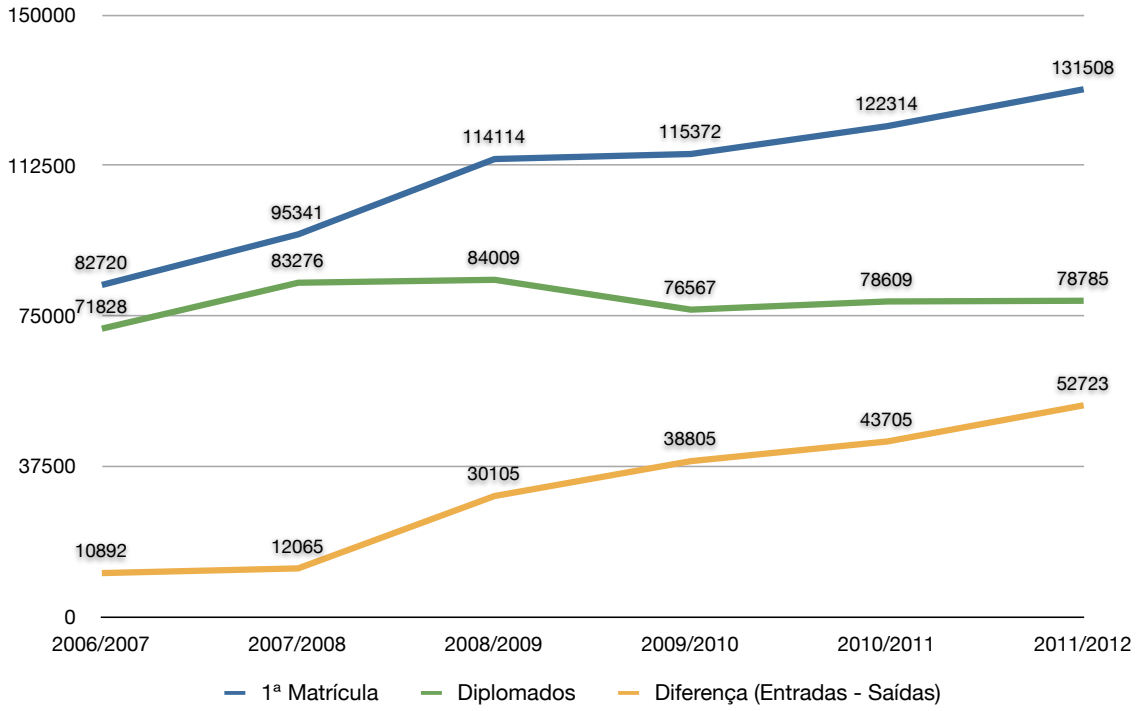


Figura 1.1: Número de Alunos com 1ª Matrícula e Diplomados no Ensino Superior. (PORDATA)

também é visível que a diferença entre o número de entradas e saídas tem vindo a aumentar e que esse valor é muito elevado. Isto indica que a população média das universidades aumenta de ano para ano o que, aliado à redução de financiamento, coloca uma grande pressão sobre as instituições.

Na figura 1.2 apresenta-se um gráfico equivalente mas desta vez para os alunos da Licenciatura de Eng. Informática (LEI - 1º ciclo) da Universidade Nova de Lisboa. O elevado número de diplomados nos anos lectivos de 2006/07 e 2007/08 pode ser explicado tendo em conta a transferência dos alunos da licenciatura antiga (de 5 anos) para a nova (1º ciclo de três anos). Com a transferência e subsequente atribuição de equivalências às unidades curriculares da nova licenciatura, muitos alunos reuniam as condições para obter o diploma. Comparando as diferenças de entradas e saídas entre o ensino superior e LEI, em média, nos últimos cinco anos, o número de diplomados do ensino superior é igual a 75,6% do número de entradas, face aos 68,4% de LEI, o que indica uma tendência para o aumento da população estudantil. Um aumento tão elevado causa problemas tanto de infra-estrutura como de gestão de recursos e é no interesse das instituições académicas melhorar o desempenho dos seus alunos para que o número de diplomados

[1]Dados retirados do site www.pordata.pt [12d].

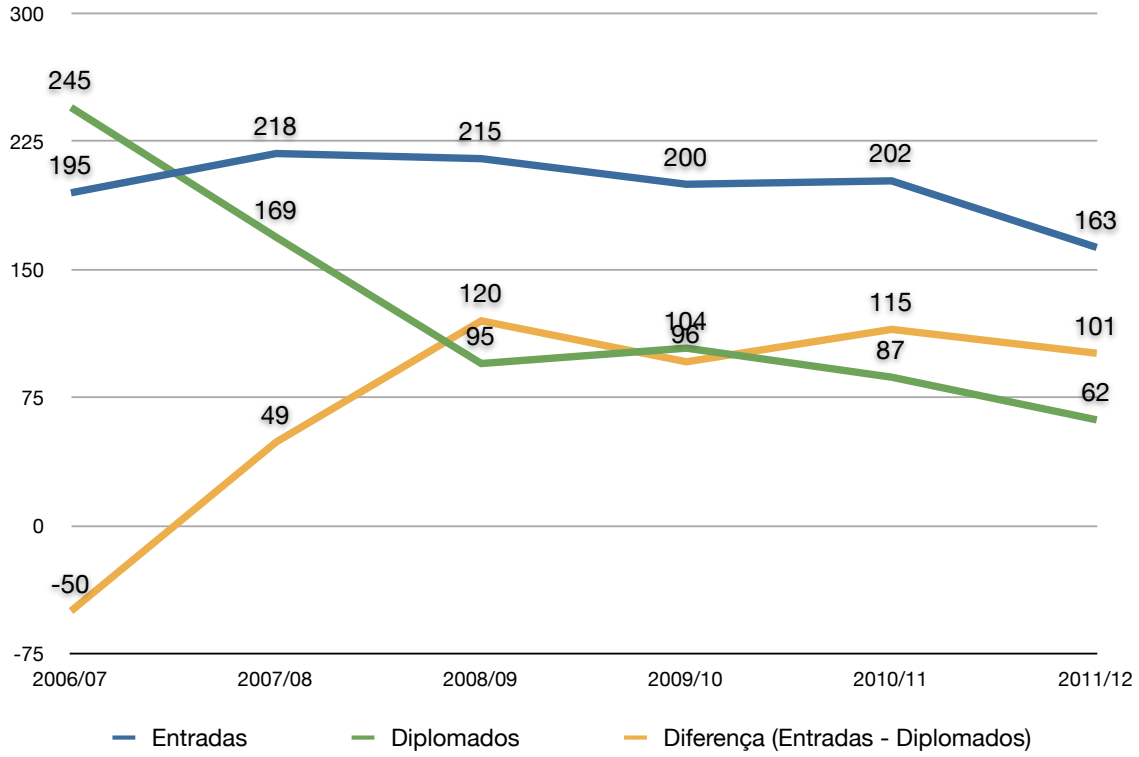


Figura 1.2: Número de Alunos com 1ª Matrícula e Diplomados na Licenciatura em Eng. Informática da FCT-UNL. (CLIP)

aumente e se mantenham as condições de ensino.

O grupo de alunos que entram no curso num determinado ano é designado de turma ou fornada. É possível analisar a progressão desta ao longo dos anos, estudando o número de alunos que desiste, diploma e passa para o ano curricular seguinte. Na figura 1.3 observa-se a evolução da turma 2006/07 até ao ano 2011/12. É possível distinguir diferentes desempenhos, nomeadamente: alunos que terminam o curso no tempo esperado (em três anos), alunos que não chegam a terminar (desistem ou transferem para outros cursos) e alunos que terminam o curso no dobro do tempo esperado. Estes desempenhos podem ter casos mais particulares, como alunos que não aprovam uma unidade curricular no primeiro ano e que nos três anos seguintes concluem o curso, e é interessante entender quão diferentes são estes desempenhos. Se se perceber o que distingue os alunos, se apenas varia uma unidade curricular feita ou se as diferenças são muito mais notáveis, é possível aplicar medidas apenas para determinado grupo, com o intuito de melhorar o desempenho e reduzir o tempo até diplomar.

Com a redução de financiamento e o aumento da população estudantil cresce uma necessidade de melhorar a gestão dos recursos utilizados, sem comprometer a qualidade do ensino. Actualmente, essa necessidade passa por monitorar o desempenho académico, nomeadamente a velocidade de obtenção de diploma, a taxa de desistência dos estudos e a taxa de continuação dos estudos para o ciclo seguinte. Em geral, a situação académica tem de ser melhorada e as medidas tomadas para esse efeito têm de ter em conta que o

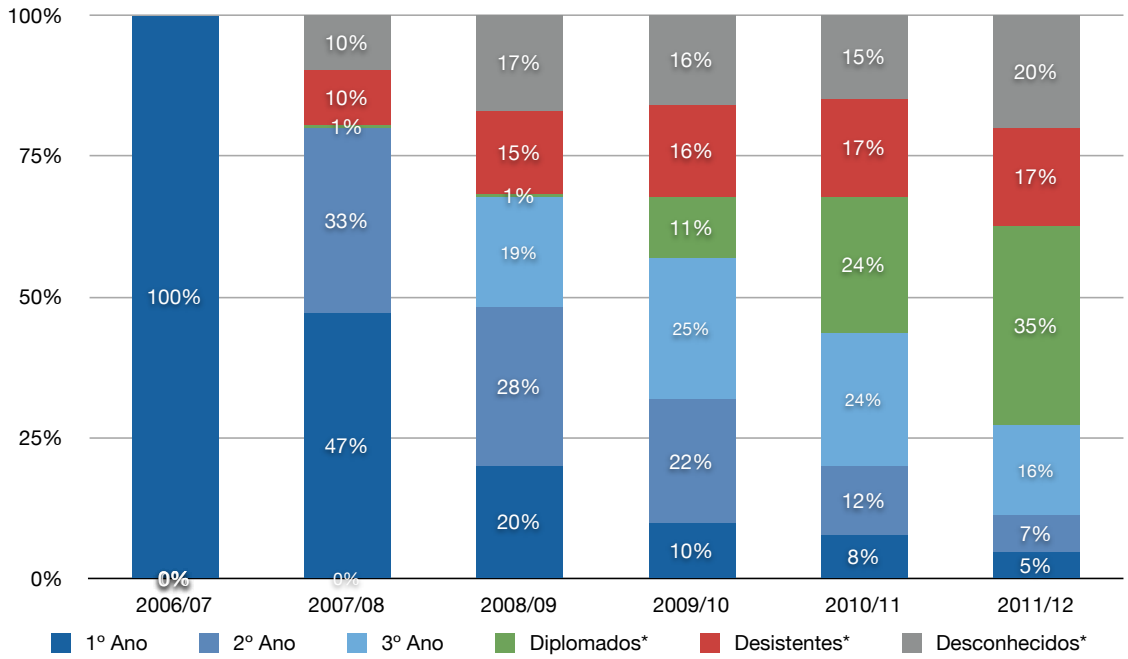


Figura 1.3: Evolução da fornada de 2006/07. (CLIP)
** Estes valores são acumulados com os dos anos anteriores.*

desempenho não é homogéneo no conjunto dos alunos do curso.

1.1.1 Contexto

O contexto deste trabalho centra-se na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa e na sua população estudantil. Os dados utilizados provêm do serviço de informação da UNL, o CLIP. Irá ser estudada a população de LEI, desde o ano lectivo de 2006/07 até ao presente. A escolha deste período de tempo é justificada pela reforma ao ensino superior derivada do processo de Bolonha [12b], que afectou o curso precisamente a partir do ano de 2006. Tendo esta reforma alterado o curso profundamente, deixa de fazer sentido considerar os períodos lectivos com o formato antigo da licenciatura.

Esta dissertação irá utilizar os dados do CLIP, com algumas ressalvas: o estudo do desempenho académico não irá ter em conta a frequência dos alunos, a docência das unidades curriculares estudadas, nem os métodos de avaliação utilizados. A não utilização da informação das frequências segue o mesmo principio seguido pelo projecto SAIA[2], que detectou que esta informação estava muitas vezes incompleta ou errónea e por isso não é fiável. A docência e os métodos de avaliação podem ser considerados como factores relevantes numa análise ao desempenho académico, mas dentro do contexto descrito não será viável a sua utlização.

[2]Projecto SAIA (Sistema de Análise de Informação Académica) visa a criação e gestão de um data warehouse que é utilizado na produção dos relatórios anuais de LEI.

1.2 Objectivos

A necessidade de melhorar a performance académica está directamente relacionada com a criação e aplicação de medidas aos alunos que frequentam o ensino superior. Seria uma grande mais valia desenvolver e aplicar estas medidas a grupos específicos de alunos (por oposição à aplicação de medidas globais), se fosse possível identificar os alunos com um comportamento diferenciado. Esta dissertação têm como objectivo desenvolver e propor uma metodologia para detectar e definir padrões de desempenho académico nos estudantes do ensino superior.

A validação deste objectivo será feita através da aplicação do método aos dados dos alunos de LEI, relativos ao período do ano lectivo de 2006/07 até ao ano lectivo de 2011/12.

1.3 Abordagem

São consideradas duas abordagens que analisam o desempenho dos alunos de perspectivas diferentes. Uma considera o desempenho num ano lectivo e a outra considera o percurso do aluno pelo curso.

A primeira abordagem considera o trabalho realizado pelos alunos num determinado ano lectivo e tenta encontrar padrões no desempenho apenas com os dados do ano. Esta abordagem permite não só analisar como é que a população do curso se comporta num ano lectivo, como também permite uma comparação ano a ano, olhando para os padrões detectados num determinado ano e avaliando a evolução de cada padrão nos anos seguintes (se os alunos mantêm comportamentos semelhantes ou se o padrão se extingue).

A segunda abordagem tem como propósito estudar o caminho percorrido pelo aluno durante todo o período em que esteve inscrito no curso, detectando padrões nas inscrições e aprovações do aluno às unidades curriculares presentes no plano de estudos. Nesta abordagem abstrai-se do desempenho dos alunos em cada ano lectivo e considera-se o desempenho globalmente, sendo que o número de inscrições para concluir uma unidade curricular passa a ser um factor de performance a considerar.

A validação das abordagens utilizando a população de LEI será útil como caso de uso para demonstrar a utilização da metodologia. Como ferramentas de análise serão utilizados os algoritmos de agrupamento: algoritmo hierárquico aglomerativo, *k-means*, SNN e SOM. A escolha destes algoritmos é explicada com mais pormenor na secção 2.1.1.

A secção seguinte descreve a estrutura e fonte dos dados que foram utilizados nesta análise.

1.4 Dados académicos

Esta secção introduz alguns conceitos fundamentais para o correcto entendimento do que é um curso superior, como se organiza e qual a informação mínima necessária para

a medição do desempenho escolar.

Cada instituição académica oferece em geral vários cursos, que habilitam a um grau académico. Os graus académicos do ensino superior em Portugal dividem-se em três ciclos: 1º ciclo - Licenciatura, 2º ciclo - Mestrado, 3º ciclo - Doutoramento, com durações médias de três, dois e quatro anos, respectivamente [12a]. A um curso está associado um ou vários pares, compostos por um plano curricular e um plano de estudos.

Um plano curricular define o conjunto de restrições que um aluno tem de cumprir para atingir o grau académico pretendido. As restrições vêm na forma de um limite mínimo de créditos ECTS obtidos em determinada àrea científica ou numa lista de unidades curriculares com aprovação obrigatória.

Um plano de estudos define a organização das unidades curriculares pelos semestres e pelos anos curriculares. O plano de estudos também define quais são as unidades curriculares de aprovação obrigatória e quais são as opcionais. Em conjunção com o plano curricular, estes dois planos definem o mínimo necessário para um aluno obter o grau académico do curso.

Uma unidade curricular (UC) é uma unidade de ensino a que os estudantes se podem inscrever, que pretende formar alunos sobre um determinado tema referente a uma área científica. Os alunos inscritos são avaliados e, quando a unidade termina, obtêm uma classificação final. A classificação mínima para um aluno aprovar a uma unidade curricular é de normalmente dez valores. Uma unidade curricular pode ter mais do que uma edição por ano, sendo leccionada em ambos os semestres.

A noção de frequência pode existir numa UC. A frequência a uma unidade curricular determina se um aluno pode ou não receber a avaliação final, ou por outras palavras, se realizou um conjunto mínimo de trabalho que o habilita a ser avaliado. A obtenção de frequência pode ser realizada, por exemplo, registando o número de aulas a que o aluno assistiu, realizando avaliações periódicas ao longo do semestre para avaliar o conhecimento do aluno ou ainda realizando um trabalho relacionado com o tema da disciplina.

A inscrição a uma unidade curricular pode ser limitada pela existência de precedências. Isto implica que para um aluno se inscrever a determinada disciplina, necessita de aprovar às unidades curriculares precedentes. Por exemplo, para um aluno se inscrever em Química II precisa de obter aprovação a Química I.

Se um aluno aprovar às unidades curriculares definidas no plano de estudos, considera-se que obteve aproveitamento académico. Também pode ser feita uma definição relativamente ao semestre do aluno, em que um aluno obtêm aproveitamento quando aprova a todas as unidades curriculares do plano de estudos para esse semestre.

No contexto da FCT-UNL, os dados referentes aos cursos, disciplinas, alunos, etc. podem ser consultados através do sistema de informação académica, conhecido como CLIP. Os dados foram extraídos para um ficheiro que contem a informação de um curso estruturada de forma a facilitar o processamento e será esta a fonte dos dados para este trabalho. O ficheiro vêm em formato XML e contêm dados sobre as unidades curriculares leccionadas e os alunos inscritos num determinado ano. Os dados das unidades

curriculares seguem a estrutura seguinte:

- **Detalhes Gerais** O nome da disciplina, o número de ECTS, se é obrigatória no curso, o departamento responsável e a área científica a que pertence;
- **Detalhes Específicos:** O número de estudantes inscritos na edição da disciplina, a linguagem principal, se as aulas da unidade curricular são de presença obrigatória ou não, quem é o professor regente e o professor responsável, e qual é a decomposição dos ECTS por horas dedicadas a estudo, a avaliações, a projectos e outros.

Se houver mais do que uma edição de unidade curricular por ano lectivo (uma em cada semestre), estão descriminadas no ficheiro. Para cada aluno os dados têm a estrutura seguinte:

- **Detalhes Genérios:** Nome, sexo e idade, número de inscrições no curso, ano curricular em que o aluno estava nesse ano, ano de entrada no curso, estatutos do aluno, estado do aluno (se está activo, desistente ou graduado);
- **Detalhes de Candidatura:** Fase de entrada, nota de entrada, nome, distrito e concelho da escola secundária de onde se candidatou e ainda o número da opção;
- **Detalhes de Desempenho Escolar:** Subdividem-se em três categorias, detalhes das inscrições a disciplinas, detalhes das avaliações feitas ao longo do ano nessas unidades curriculares e detalhes das aprovações conseguidas nesse ano.
  - **Inscrições:** Para cada unidade curricular a que o aluno se inscreveu esse ano existe o semestre, a vez da inscrição e o tipo (inscrição curricular ou extracurricular);
  - **Avaliações:** Para cada avaliação existe o semestre em que foi lançada, a classificação, a fase da avaliação (pode ser fase normal, recurso, especial e extra) e o tipo. Este último atributo define de que se trata a avaliação: se foi de aprovação, de frequência, de melhoria, ou de melhoria ad-hoc. A fase da avaliação não é considerada se for uma avaliação do tipo frequência.
  - **Obtenções:** Para cada unidade curricular aprovada nesse ano lectivo existe o tipo (se o aluno aprovou à unidade nesta instituição então é interna, se a unidade foi realizada noutra instituição é externa), a nota e a data da aprovação.

A secção seguinte apresenta a Licenciatura em Eng. Informática, o caso de estudo desta dissertação.

1.4.1 Licenciatura em Eng. Informática

Esta secção irá detalhar a estrutura da LEI, apresentando em pormenor o plano curricular que compõe o curso.

O ensino de informática a nível universitário começou 1975 na FCT-UNL, com a criação de um departamento responsável pelo primeiro curso inteiramente dedicado às Ciências e Engenharia dos Computadores [12c]. No ano lectivo de 2006/07, foi implementada a nova licenciatura, seguindo o processo de Bolonha [12b], correspondente ao 1º ciclo do ensino superior. É um curso de três anos e 180 créditos ECTS. As áreas científicas que este curso abrange são Informática, Matemática, Física, Engenharia Electrotécnica e Ciências Humanas e Sociais.

O curso oferece dois planos curriculares (ou perfis) que podem ser escolhidos pelos alunos. Os planos foram desenhados para possibilitar a escolha do percurso que melhor se adapta às necessidades do aluno: um plano curricular para alunos que queiram terminar a licenciatura e integrar o mercado de trabalho (designado de Informática Aplicada) e outro para os alunos que queiram prosseguir com os seus estudos para o ciclo seguinte (designado de Ciências da Engenharia). Os dois perfis partilham uma parte significante dos planos de estudo, nomeadamente referente aos dois primeiros anos. A escolha por parte dos alunos não é final e pode ser alterada a qualquer altura.

As restrições em comum dos planos curriculares oferecidos em LEI são as seguintes:

- Aprovar em todas as cadeiras obrigatórias;
- Obter no mínimo 180 créditos;
- Obter no mínimo 31 créditos na área de Matemática;
- Obter no mínimo 6 créditos na área de Física;
- Obter no mínimo 6 créditos na área de Engenharia Electrotécnica.

Quando a escolha de percurso é realizada, este conjunto é estendido com as seguintes restrições:

- Informática Aplicada:
  - Obter no mínimo 122 créditos na área de Informática;
  - Obter no mínimo 18 créditos nas unidades curriculares avançadas de Informática;
  - Obter no mínimo 9 créditos em Ciências Humanas e Sociais;
  - Aprovar na unidade curricular de Projecto Integrador.
- Ciências da Engenharia:
  - Obter no mínimo 110 créditos na área de Informática;
  - Obter no mínimo 15 créditos em Ciências Humanas e Sociais;
  - Aprovar nas unidades curriculares Aspectos Sócio-Profissionais de Informática e Estágio Professionalizante.

Existe ainda um plano curricular transitório, semelhante aos planos aqui apresentados, criado para facilitar a passagem dos alunos da antiga licenciatura (antes do processo de Bolonha) para a nova. Este plano curricular não é apresentado pois apenas existe interesse em estudar os alunos que entraram no curso depois da implementação do processo de Bolonha, que apenas têm para escolha os planos acima descritos.

1.4.1.1 Estatísticas Descritivas (desde 2006/07 até 2011/12)

Para melhor compreensão dos dados, esta secção apresenta uma descrição estatística da população, fazendo uso de alguns indicadores que são normalmente utilizados em relatórios anuais de desempenho. Designa-se um aluno que entrou antes do processo de Bolonha como pré-Bolonha e um aluno que entrou depois como pós-Bolonha.

A Licenciatura em Eng. Informática tem em média 884 alunos inscritos no curso e, todos os anos, uma média de 198 alunos candidatam-se e são aceites na licenciatura. A distribuição do número de alunos inscritos por ano lectivo pode ser vista na figura 1.4, com distinção entre novas entradas e alunos já inscritos.

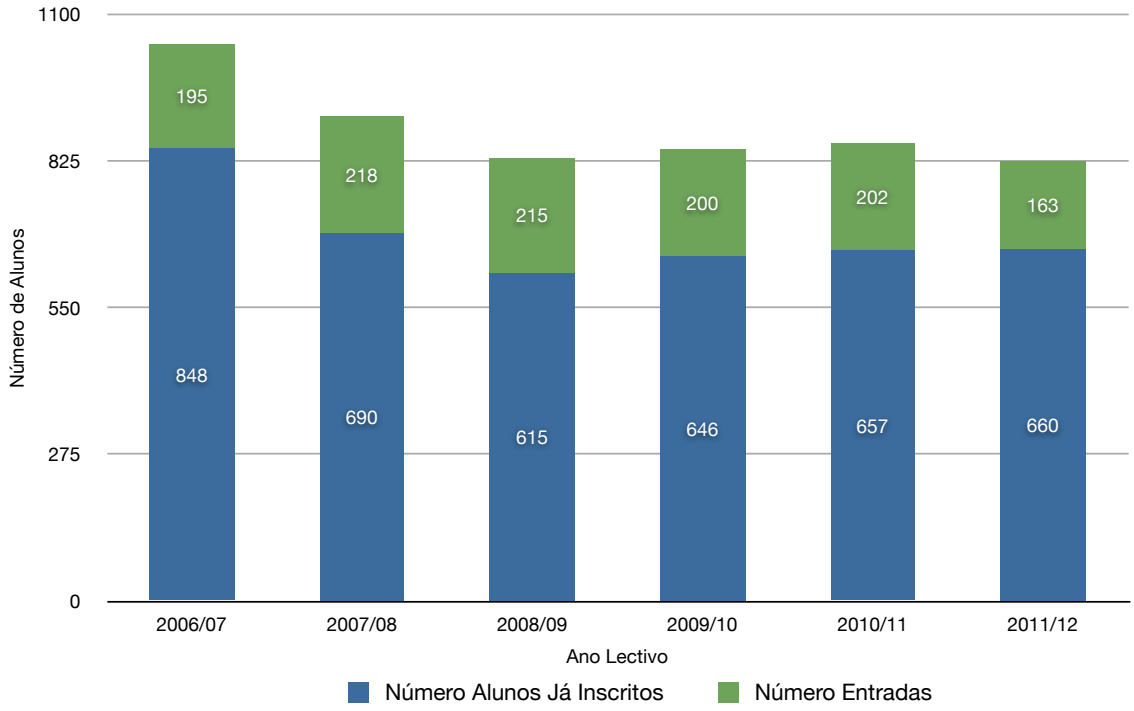


Figura 1.4: Número de Alunos Inscritos no Curso

Desde o ano lectivo 2006/07 até 2011/12, 785 alunos concluíram a licenciatura. Dos alunos pós-Bolonha, 164 diplomaram-se com o perfil de Ciências da Engenharia e 27 diplomaram-se com o perfil de Informática Aplicada. Na figura 1.5 pode se observar, para cada ano lectivo, o número de alunos pós-Bolonha diplomados. A nota média de conclusão da licenciatura é de 13 valores, enquanto o tempo médio de conclusão é de 6 anos. A figura 1.6, mostra o tempo de graduação médio por ano lectivo dos alunos pós-Bolonha, distinguindo o perfil. Para o perfil de Informática aplicada a nota média de

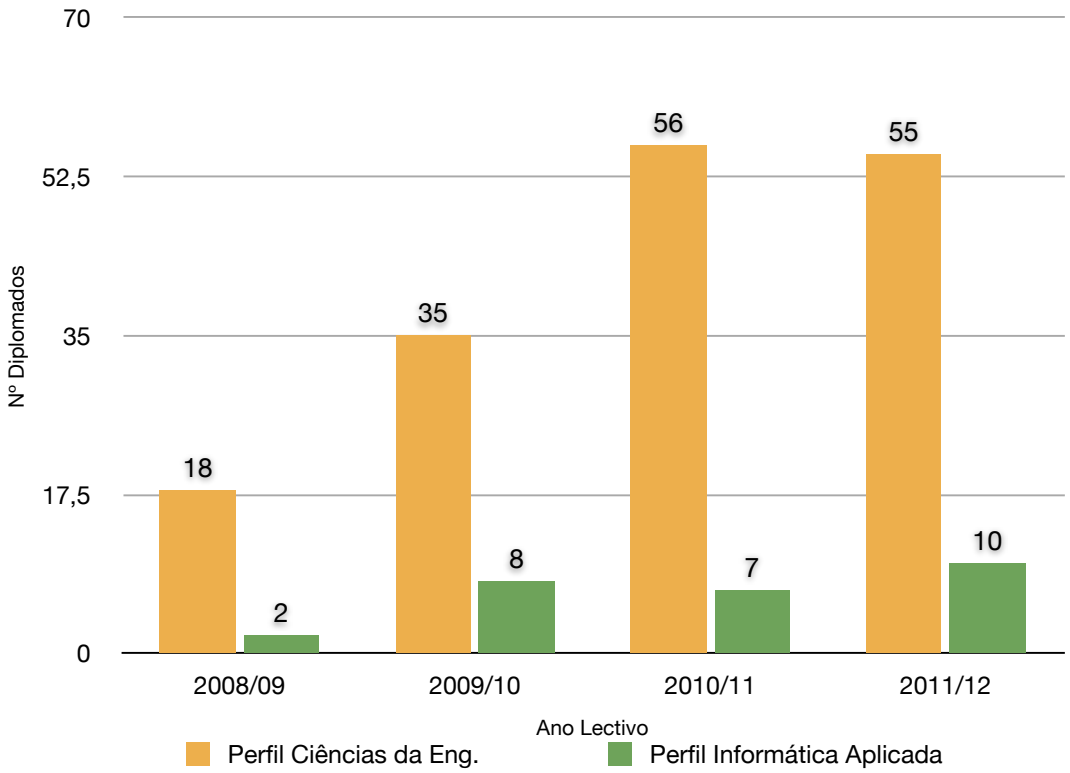


Figura 1.5: Número de alunos pós-Bolonha diplomados por perfil.

graduação é de 13 valores e para Ciências da Engenharia é de 14, com tempos médios de conclusão de 4 e 3 anos e meio, respectivamente.

1.5 Estrutura do Documento

Depois desta introdução, O documento presente está estruturado com a seguinte organização: o capítulo 2 apresenta o estado da arte e o trabalho relacionado com a análise, nomeadamente o que são técnicas de agrupamento e quais são as ferramentas usadas neste trabalho; de o capítulo 3 descreve com pormenor qual é então a metodologia desenvolvida, detalhando as abordagens escolhidas para a análise dos dados; por fim, o capítulo 4 apresenta as conclusões deste trabalho: a análise dos resultados obtidos, discussão do processo e indicações para trabalho futuro.

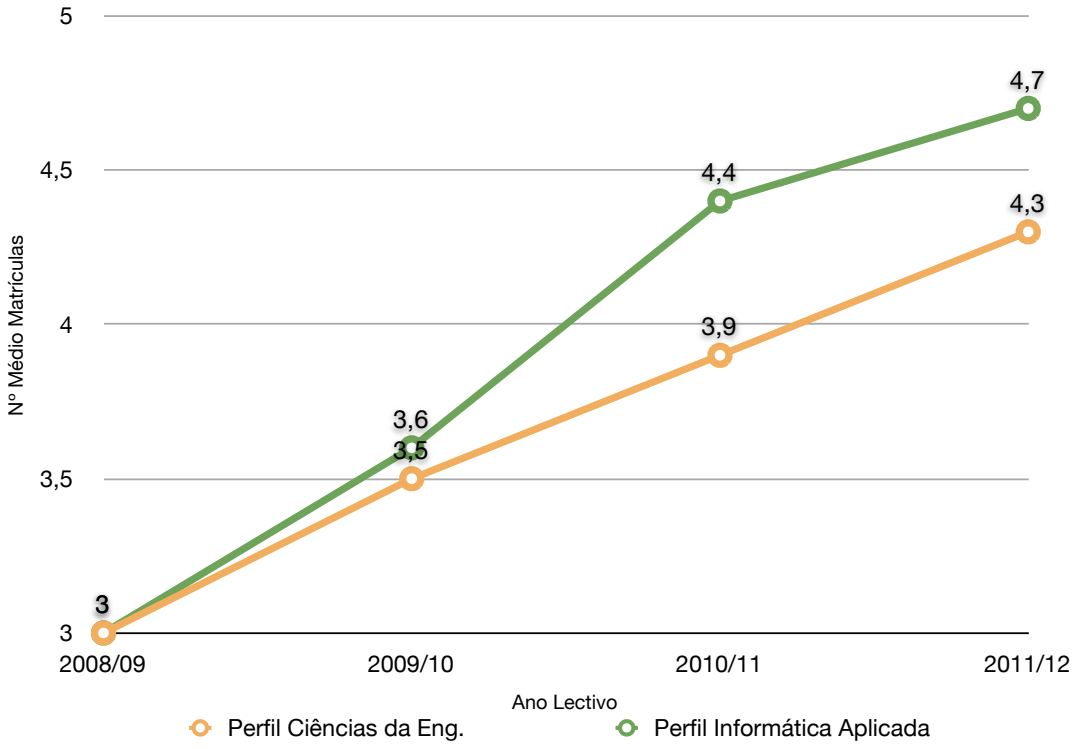


Figura 1.6: Tempo de conclusão médio por perfil.

2

Estado da Arte

O capítulo seguinte apresenta os trabalhos e técnicas relacionadas com o tema da dissertação, nomeadamente trabalhos relacionados com a utilização de técnicas de data mining para estudar a população estudantil do ensino superior. O domínio dos trabalhos apresentados foca-se sobretudo no estudo de alunos de instituições do ensino superior, utilizando como ferramentas de análise técnicas estatísticas ou de data mining.

Na secção 2.1 introduzem-se os conceitos chave de data mining, apresentam-se as técnicas mais utilizadas e as que melhor se aplicam ao problema em questão, descreve-se uma das melhores ferramentas criadas para análise de dados usando técnicas de data mining, a ferramenta WEKA (que foi usada extensivamente neste trabalho) e descreve-se também outras ferramentas utilizadas. A secção 2.2 apresenta os trabalhos mais relevantes no domínio do tema.

2.1 Data Mining

A quantidade de dados produzida nos dia de hoje é gigantesca. Qualquer acção tomada no dia a dia de uma pessoa, desde a compra de uma peça de roupa com um cartão de crédito, ao clique numa ligação para um website, à passagem de carro numa portagem, é registada e guardada. Tendo em conta isto, a expressão "vivemos na era da informação"deixa de ter significado. Vivemos sim na era dos dados, e visto existir uma relação inversa entre a geração de dados e a proporção desses mesmos que é compreendida, muita informação potencial é perdida no volume de elementos existente [HKP11].

É precisamente da necessidade de extrair conhecimento de grandes quantidades de dados que surge o processo conhecido por data mining. Data mining é na verdade um passo no processo de descoberta de conhecimento [HKP11] (em inglês *knowledge discovery*

from data ou KDD) que consiste em pré-processamento, data mining e pós processamento. Dado que a expressão "data mining"é normalmente associada a KDD, esta é aqui definida num sentido mais geral. Assim, data mining é o processo de analisar grandes volumes de dados, procurando por correlações ou padrões, na tentativa de extrair conhecimento útil [WFH11].

Este tipo de análise só é possível graças a alguns avanços tecnológicos [Lua02], nomeadamente: a grande capacidade de processamento nos computadores dos dias de hoje, que é necessária para processar grandes quantidades de dados em tempo útil; o custo reduzido de armazenamento, que permite guardar os volumes de data gerados; e a evolução da tecnologia das bases de dados, que permite não só guardar mas também aceder aos dados de uma forma estruturada, rápida e eficaz.

Data mining é um processo que depende bastante dos dados e pode ser adaptado, modelado e aplicado a vários domínios. Os tipos de dados analisados e os padrões detectados variam bastante. As fontes de dados mais comuns usadas em data mining são bases de dados relacionais, bases de dados transaccionais e data warehouses [HKP11]. Bases de dados relacionais são os repositórios de dados mais comuns e mais ricos em informação, sendo assim muito usados no estudo de data mining. Em bases de dados transaccionais, data mining é utilizado para *market basket analysis*, bastante útil para detectar grupos de produtos que se forem vendidos juntos, potenciam as vendas. Um data warehouse é um repositório de informação que agrega dados de várias fontes, em que data mining é normalmente utilizado para uma análise multi-dimensional, cruzando informação de várias dimensões em diferentes tipos de granularidade em busca de padrões.

Os tipos de padrões que resultam do uso de data mining podem ser definidos entre dois extremos [WFH11]. Por padrão entenda-se algo que pode ser usado para melhorar o conhecimento que se tem sobre os dados, algo que descreve as relações entre os elementos que foram analisados. Um extremo é um padrão que funciona como uma caixa preta, em que o seu funcionamento não se conhece e é usado sem se perceber o seu interior. O outro extremo é uma caixa transparente, em que a estrutura do padrão está visível. A diferença entre os dois extremos é se o padrão resultante é ou não é representado através de uma estrutura que possa ser analisada e utilizada em decisões futuras, ajudando a explicar os dados.

O padrão que se está a procurar acaba por ser definido pelo método que se utiliza[HKP11]. Os métodos podem ser divididos em duas categorias: métodos descritivos caracterizam as propriedades dos dados, descrevendo o data set que se está a analisar; e métodos predictivos, que induzem uma previsão a partir dos dados disponíveis. Como exemplo, dos métodos de data mining mais utilizados, associação e classificação são predictivos; e agrupamento e detecção de anomalias são descritivos, sendo que o foco da próxima secção será em métodos de agrupamento. Métodos de associação procuram as dependências entre os atributos dos dados. Métodos de classificação procuram por um modelo que descreva as classes presentes nos dados. Métodos de agrupamento procuram por grupos de objectos semelhantes entre si, e diferentes de objectos de outros grupos.

E métodos de detecção de anomalias procuram pelos objectos que variam bastante dos restantes. Em aprendizagem automática, os métodos de classificação e de agrupamento também são referidos como aprendizagem supervisada e não supervisada, respectivamente, visto técnicas de classificação terem acesso às classes dos objectos que estão a analisar e técnicas de agrupamento não.

No caso desta dissertação, como o propósito é descobrir padrões de desempenho ocultos nos dados, um método predicativo (como a classificação) não é aplicável. Os métodos de agrupamento são os mais indicados, sendo descritos em pormenor na secção seguinte.

De notar que a maior parte da informação sobre data mining e os seus algoritmos apresentada neste capítulo provem de dois livros bastante completos sobre a matéria: *Data mining: concepts and techniques* [HKP11], que oferece uma visão teórica sobre os métodos e os algoritmos de data mining; e *Data Mining: Practical machine learning tools and techniques* [WFH11], cuja publicação acompanha o lançamento da versão estável da ferramenta WEKA e que oferece exemplos da implementação e utilização dos algoritmos mais comuns.

2.1.1 Técnicas de Agrupamento

Nesta secção irão ser descritas técnicas de agrupamento, organizadas em quatro categorias: particionamento, hierárquicas, baseadas em densidade e baseadas numa grelha. Para as técnicas de particionamento e hierárquicas são descritos com mais pormenor os algoritmos mais comuns.

Agrupamento é o processo de agrupar objectos segundo a sua semelhança, de forma a que objectos no mesmo grupo sejam muito semelhantes e objectos em grupos diferentes sejam dissemelhantes [Jai10]. Por objecto entenda-se uma instância dos dados, por exemplo, numa análise a clientes de uma loja um objecto seria um cliente, representado por um conjunto de atributos que o descreve, como quantas vezes foi à loja, o dinheiro que já gastou lá ou a quantidade de produtos que já comprou. A medida de semelhança e dissemelhança é calculada a partir desse mesmo conjunto de atributos e normalmente é interpretada como uma medida da distância entre os dois objectos num espaço multidimensional.

Como irá ser visto nos trabalhos apresentados na secção 2.2, técnicas de agrupamento são frequentemente utilizadas em conjunção com outros métodos de data mining, como uma ferramenta para pré-processar os dados. Também são utilizadas como um processo de classificação automática, onde as classes do data set são os grupos encontrados. Técnicas de agrupamento são normalmente aplicadas para detectar a estrutura subjacente, para descobrir a classificação natural e para comprimir/resumir os dados [HKP11].

É difícil organizar as várias técnicas de agrupamento existentes em categorias claras, pois algumas contêm características que podem pertencer a mais do que uma categoria. Mesmo assim, é um procedimento necessário para melhor se perceber as várias técnicas.

No geral, as técnicas podem ser divididas nas seguintes categorias: técnicas de particionamento, técnicas hierárquicas, técnicas baseadas em densidade e técnicas baseadas numa grelha.

Técnicas de Particionamento As técnicas de particionamento criam partições nos dados, em que cada partição representa um grupo. Normalmente, os objectos só podem pertencer a uma partição. Esta condição pode ser relaxada, na técnica conhecida como *fuzzy clustering* [Dun73] . Em geral, o critério utilizado para encontrar grupos baseia-se na distância entre os objectos, ou seja, os elementos dentro do mesmo grupo são próximos entre si e afastados de elementos de outros grupos.

Técnicas Hierárquicas Técnicas hierárquicas formam uma hierarquia de grupos quando são utilizadas. Dependendo da decomposição hierárquica utilizada, podem ser classificadas em aglomerativas ou divisórias. Um algoritmo hierárquico aglomerativo começa considerando cada objecto individual como um grupo e a cada iteração vai aglomerando os grupos existentes, utilizando para isso um critério de semelhança. Um algoritmo divisório começa com apenas um grupo, contendo todos os elementos, e a cada iteração os grupos são divididos até só existir um elemento por grupo ou até se verificar uma condição de término. A condição utilizada varia, podendo ser baseada na distância entre os grupos ou na sua densidade. Separar os grupos numa hierarquia pode ser bastante útil para resumir os dados e para representá-los.

Técnicas baseadas em densidade Técnicas baseadas em densidade utilizam outra medida de semelhança para encontrar padrões. Em vez de se basearem em distâncias para criar os grupos, é utilizada a noção de densidade. Estas técnicas definem uma zona de vizinhança à volta de um objecto, sendo a densidade dada pelo número de objectos dentro da vizinhança. De uma forma geral, os algoritmos funcionam aumentando os grupos enquanto a densidade for maior que um valor mínimo definido pelo utilizador.

Técnicas baseadas numa grelha Estas técnicas dividem o espaço dos dados em células que formam uma estrutura em grelha. A grande vantagem deste tipo de algoritmos é que todos os processos de agrupamento são efectuados usando a grelha, e não o conjunto total de objectos, o que reduz os tempos de processamento. Esta técnica é independente do número de objectos a analisar e depende apenas do número de células criadas.

Um dos problemas da aplicação de técnicas de particionamento é utilizarem como input o número de grupos, o que geralmente é uma das incógnitas que se procura.

Já em algoritmos hierárquicos, a condição de fim pode ser fornecida, independentemente do algoritmo ser divisório ou aglomerativo. Utilizando algoritmos divisórios em data sets muito volumosos, torna-se demasiado pesado calcular as divisões e por isso

são utilizadas heurísticas. A utilização de heurísticas introduz algum erro na formação dos grupos mas em favor da eficiência a divisão é final, não voltando a ser considerada em iterações seguintes. É por isso natural que exista um maior número de algoritmos aglomerativos do que divisórios.

Algoritmos divisórios e aglomerativos são consideradas técnicas algorítmicas, pois olham para os dados deterministicamente, efectuando os cálculos das distâncias também de forma determinística. Para além de técnicas hierárquicas algorítmicas, existem técnicas probabílisticas, que calculam a semelhança entre objectos utilizando modelos probabilísticos (a semelhança entre dois objectos é dada por uma probabilidade em vez de um valor fixo); e técnicas bayesianas, que calculam não um conjunto de grupos, mas uma distribuição, ou seja, várias hipóteses de agrupamentos associadas a uma probabilidade.

Devido à medida de semelhança utilizada por técnicas de particionamento e hierárquicas (distância entre pontos), estas conseguem apenas detectar grupos com formas esféricas. Utilizando um critério de semelhança diferente, como a noção de densidade, é possível detectar grupos com formas arbitrárias.

De seguida vão ser descritos os algoritmos que foram utilizados neste trabalho para a procura de grupos de alunos. As razões por trás da escolha deste grupo não são muito fortes mas em primeiro lugar a utilização de técnicas diferentes permite cobrir uma área maior de resultados, visto não se conhecer à partida a estrutura dos dados. A posição adoptada para esta análise é sobretudo experimental e os algoritmos apresentados de seguida fornecem uma boa base de resultados para analisar.

Algoritmo *k-means*

O algoritmo *k-means* é uma técnica de particionamento baseada em centroides [HW79]. Cada grupo é representado pelo seu centroide, que é o centro conceptual do grupo. Os centroides são calculados utilizando a média de todos os pontos presentes no grupos. Um dos inputs deste algoritmo (designado por k) é precisamente o número de grupos que se pretende encontrar (o que pode ser visto como uma desvantagem).

Assim, na primeira iteração, são considerados aleatoriamente um número de objectos igual a k, que passam a representar os grupos iniciais. Os objectos restantes são atribuídos aos grupos mais semelhantes, utilizando a distância do objecto ao centroide como medida da semelhança. A cada iteração, a média de todos os pontos do grupo é calculada e o centroide de cada grupo é actualizado, procedendo-se à atribuição de todos os objectos aos grupos mais semelhantes. O algoritmo termina quando os grupos formados na iteração anterior não se alteram na iteração corrente.

Para medir a qualidade dos grupos formados, pode ser avaliada a variância do erro de cada objecto ao centroide do grupo onde se encontra. Ou seja, somando o quadrado da distância de cada ponto ao centro do grupo, obtemos uma variância que pode ser utilizada como medida de qualidade. A minimização desta variância é

um processo computacionalmente pesado e já foi demonstrado ser NP-hard apenas para dois grupos [HKP11]. É por essa razão que o algoritmo utiliza aproximações para chegar a uma solução em tempo útil.

Uma questão que se levanta com a utilização deste algoritmo é como escolher o melhor valor para k? A resposta a esta pergunta vem na forma de heurísticas que sugerem um valor para ser utilizado. Estas podem ser mais ou menos complicadas, por exemplo, escolher um intervalo para k, correr o algoritmo para todos esses valores e no fim escolher o valor de k que apresentar o menor erro.

Este algoritmo é bastante susceptível a ruído e valores de fronteira [HKP11], pois como utiliza a média dos objectos para encontrar os grupos, um número reduzido de amostras com valores bastante acima ou abaixo da média pode influenciar bastante os grupos produzidos. Além disso, só trabalha com atributos numéricos, em que cada objecto é um ponto num espaço multi-dimensional e em que cada atributo é uma dimensão. Apesar disso, *k-means* é dos algoritmos mais utilizados em data mining.

Algoritmo Aglomerativo Hierárquico

Um algoritmo hierárquico aglomerativo usa uma estratégia de formação de grupos de baixo para cima[HKP11]. A primeira iteração do algoritmo começa por baixo, com cada objecto a formar o seu próprio grupos, e a cada iteração fundem-se os dois grupos que se encontram mais próximos (mais semelhantes). O algoritmo termina quando se chega ao topo, ou seja, quando tem um grupo com todos os elementos, que é considerado a raiz da hierarquia. Uma representação habitual para este algoritmo é uma árvore dos grupos formados, em que cada nível representa uma iteração.

Independentemente do tipo de abordagem utilizada pelo algoritmo para formar os grupos, é sempre necessário uma medida da distância. As medidas mais comuns são: distância mínima, máxima e média. Quando é utilizada a distância mínima o algoritmo tende a formar grupos com grandes diâmetros e torna-se muito sensível a valores de fronteira, sendo que apenas um objecto pode alterar radicalmente a estrutura criada. Se for utilizada a medida máxima, os grupos formados ficam muito mais compactos, mas o algoritmo continua sensível a valores de fronteira. A medida média é um compromisso entre os extremos e pode ser calculada de duas maneiras diferentes. A primeira utiliza a média de todos os elementos do grupo para eleger um representante, semelhante a um centroide. Esta medida resulta bastante bem quando os objectos estão posicionados num espaço euclidiano e é possível definir com exactidão um centroide. No caso de só existirem relações de semelhança entre pares de objectos (em vez de uma função distância), ou existirem atributos categóricos em vez de numéricos, o centroide pode tornar-se impossível de definir,

visto não ser um objecto presente nos dados. A segunda maneira de definir a distância média entre dois grupos é calcular a média das distâncias dos objectos de um grupo para o outro.

Se os grupos existentes nos dados forem compactos e bem separados, a estrutura hierarquica produzida irá ser semelhante para qualquer medida de distância utilizada. Se isso não se verificar, a estrutura poderá ser bastante diferente. Este algoritmo necessita no máximo de um número de iterações igual ao número de objectos existentes, pois cada iteração funde dois grupos em um.

Algoritmo DBScan

O algoritmo DBScan baseia-se na noção de vizinhança para detectar os grupos existentes [Est+96]. Este algoritmo recebe dois valores de input, o raio da vizinhança e um valor mínimo para a densidade de uma zona. A zona de vizinhança à volta de um ponto é definida pelo raio e a densidade é definida pelo número de objectos dentro da esfera definida por esse raio (a vizinhança). O valor mínimo para a densidade de uma zona serve como a condição de paragem do algoritmo.

Na primeira iteração do algoritmo, todos os objectos são avaliados para determinar se a sua vizinhança é densa ou não, utilizando o valor mínimo de densidade. Se uma vizinhança for densa, o objecto é marcado como um *core object* e serve de referência para aquela zona. O processo de agrupamento fica simplificado a partir daqui, pois é efectuado apenas nos *core objects* e não em todo o data set. Nas iterações seguintes, o algoritmo procura ligar as zonas densas, procurando por *core objects* que tenham objectos dentro da mesma vizinhança. Se for encontrado um objecto que se encontre dentro de duas vizinhanças, então é criada uma ligação entre os *core objects* respectivos e as duas zonas são fundidas numa só.

O algoritmo encontra grupos existentes iterando sobre todos os objectos e verificando a sua vizinhança. Se um objecto ainda não tiver sido visitado e a sua vizinhança for maior que o valor mínimo então o objecto é marcado como "visitado"e é criado um grupo com esse único objecto. O algoritmo procede iterando sobre a vizinhança desse objecto, procurando pontos que não foram ainda visitados e cuja vizinhança seja maior que o valor mínimo. Os objectos que encontrar nessa situação são adicionados ao grupo criado até todos os pontos da vizinhança inicial já terem sido explorados. Quando isto acontece o algoritmo retorna o grupo formado. A iteração seguinte irá começar do início, mas os pontos deste grupo já não são tidos em conta. Os objectos que não têm uma vizinhança superior ao valor mínimo são marcados como ruído e ignorados na formação dos grupos.

Se os valores para o raio da vizinhança e número mínimo de densidade forem apropriados, o algoritmo é bastante eficaz a encontrar grupos com formas arbitrárias.

Algoritmo SNN

O algoritmo SNN é um algoritmo baseado em densidade, semelhante ao DBScan [MSC05]. A diferença está na forma como a noção de semelhança entre objectos é definida. No SNN, são utilizados os vizinhos mais próximos do objecto, sendo que o valor da semelhança é calculado como a soma das semelhanças dos pontos mais próximos. Os pontos com alta densidade são designados de *core objects* e pontos com baixa densidade são marcados como ruído. Os restantes objectos que são muito semelhantes a *core objects* representam novos grupos.

Este algoritmo recebe como três valores de input: o número de objectos que compõem a vizinhança, o valor de limite para a densidade e o limite que define os *core objects*. Depois dos valores de input definidos, o algoritmo começa por criar as vizinhanças de todos os objectos. Em seguida, a semelhança de cada objecto é pre-calculada também. Usando a noção de semelhança definida acima, a densidade de cada objecto pode ser calculada como o número de vizinhos em que o número de vizinhos partilhados é maior ou igual ao limite de densidade. Depois disto, os objectos são classificados como como *core objects*, se a sua densidade for maior ou igual ao limite. A partir deste ponto, o algoritmo já tem tudo o que precisa para iniciar o processo de agrupamento, que começa a partir dos *core objects*. No final, os pontos que não estão presentes em nenhum grupo são marcados como ruído.

O SNN é aqui apresentado porque é mais flexível que o DBScan, apesar de serem semelhantes. É capaz de detectar padrões cuja densidade varie ligeiramente, devido à maneira como a própria semelhança é definida.

Algoritmo SOM

Um mapa auto-organizado (*self-organizing map*) é um tipo de rede neuronal composta por nós ou neurónios [Koh90]. Cada nó tem associado um vector de pesos e os vários nós competem (e organizam-se) para representar os dados de entrada. Este algoritmo é bastante utilizado para criar representações de baixa dimensionalidade (normalmente 2-D) de dados com elevado número de dimensões.

O algoritmo recebe como parâmetros a largura e o comprimento da rede (o número de nós será comprimento x largura), e começa por iniciar o vector de pesos de cada nó com valores aleatórios. De seguida, para cada dado de input é calculado o nó que melhor se adapta, usando a distância euclidiana para procurar o nó com a distância mínima. Quando esse nó é encontrado, os nós vizinhos são actualizados e "puxados"para o nó escolhido. É com esta actualização que a rede se organiza para melhor representar a toponomia dos dados. O algoritmo termina quando todos os objectos forem processados e cada nó (e o seu vector de pesos) representa um grupo.

Todos os algoritmos aqui apresentados fazem uso de funções de distância para determinar a a semelhança entre dois objectos. A mais comum é a distância euclidiana, mas

também se poderia utilizar a distância de Manhattan ou outra. Esta função distância também é utilizada por algumas métricas para avaliar a qualidade dos grupos encontrados, como está descrito na secção seguinte.

2.1.2 Verificação e Validação do Agrupamento

Os resultados obtidos pela aplicação das várias técnicas de agrupamento serão dependentes de vários factores, incluindo a técnica utilizada, e por isso variam bastante. Para ser possível distinguir um bom resultado de um mau resultado é necessário utilizar métricas que atribuam à qualidade um valor. Dependendo da métrica, é possível comparar os resultados de uma forma global ou mais específica, grupo a grupo.

As métricas que vão ser utilizadas fazem quase todas uso das funções de semelhança e dissemelhança entre objectos. Nenhuma métrica é perfeita, têm todas as suas vantagens e desvantagens e por isso são utilizadas em conjunto. Idealmente, quer-se que os algoritmos utilizados retornem grupos em que objectos do mesmo grupo estejam bastante próximos e objectos de grupos diferentes estejam separados. As métricas seguem esta lógica, sendo que um bom resultado indica precisamente uma situação próxima da ideal.

Coesão A coesão é definida como a média das distâncias dos objectos de um grupo ao centro do seu grupo. Quanto menor for este valor, mais compacto é o grupo. Também é utilizado o valor global fazendo a média da coesão de todos os grupos. Se $\delta(O_x, C)$ for a distância do objecto O_x ao centro do grupo C, a coesão pode ser definida matematicamente como:

$$S = \sum_{x=0}^{N} \delta(O_x, C)$$

Afastamento O afastamento é definido como a média das distâncias dos objectos de um grupo ao centro do grupo vizinho mais próximo. Quanto maior este valor, mais separados são os grupos. O valor global também é utilizado, fazendo a média do afastamento de cada grupo. Se $\delta(O_x, C_n)$ for a distância do objecto O_x ao centro do grupo vizinho mais próximo C_n, o afastamento pode ser definido matematicamente como:

$$D = \sum_{x=0}^{N} \delta(O_x, C_n)$$

Silhueta A silhueta é definida para um objecto como o rácio entre a coesão e o afastamento desse objecto [Rou87]. Mais precisamente,

$$Silhueta_i = \frac{(D_i - S_i)}{D_i}$$

onde S_i é a coesão do objecto i ao seu grupo e D_i o afastamento desse objecto ao

grupo vizinho mais próximo. O valor deste indicador está compreendido entre 1 e -1, em que 1 indica que o objecto se encaixa bem no grupo em que está e -1 indica que o objecto encaixa melhor no grupo vizinho. Esta métrica é bastante interessante pois tanto pode ser aplicada a cada grupo, para tentar perceber quais os elementos que não se enquadram; como pode ser aplicada ao nível do dataset (fazendo a média da silhueta de cada objecto para cada grupo), para detectar agrupamentos que não estejam bem; como ao nível do esquema de agrupamento, permitindo comparar, por exemplo, as diferentes iterações do *k-means* com diferentes valores de *k*.

Índice de Davies–Bouldin Este índice é definido como o rácio entre a coesão dos grupos e o seu afastamento [DB79]. Se S_i for a coesão do grupo i e se M_{ij} for o afastamento entre os grupos i e j, R_{ij} é uma medida da qualidade definida como:

$$R_{ij} = \frac{\mathrm{S}_i + \mathrm{S}_j}{M_{ij}}$$

Quanto menor este valor melhor é o agrupamento calculado e para um grupo i, define-se o melhor valor de agrupamento como:

$$A_i \equiv \max_{j:i\neq j} R_{ij}$$

Assim o índice de Davies-Bouldin (DB) é calculado como a média da qualidade de cada grupo:

$$DB \equiv \frac{1}{N}\sum_{i=1}^{N} A_i$$

O melhor esquema de agrupamento terá o menor valor deste índice.

Índice de Dunn Esta métrica é definida como o rácio entre o afastamento dos grupos e o diâmetro do grupo para o qual este valor é máximo [Dun73]. A noção de diâmetro neste indicador é calculada como a distância entre os dois objectos mais afastados do grupo. Matemáticamente, o índice de Dunn (DI) pode ser definido da seguinte maneira:

$$DI_m = \min_{1\leq i\leq m}\left\{\min_{1\leq i\leq m, j\neq i}\left\{\frac{\delta(C_i, C_j)}{\max_{1\leq k\leq m}\Delta_k}\right\}\right\}$$

em que $\delta(C_i, C_j)$ é o afastamento entre os grupos C_i e C_j, $\max_{1\leq k\leq m}\Delta_k$ é o diametro máximo existente e m é o número de grupos detectados. Visto este índice procurar pelo valor mínimo, quanto maior for o valor melhor é o agrupamento calculado. Esta métrica é considerada como um indicador do pior cenário possível pois como utiliza o valor de diâmetro máximo, e se por acaso um dos grupos detectados for mal formado e tiver um diâmetro desproporcional ao resto dos grupos, o índice fica com um valor muito reduzido.

2.1.3 WEKA

A *Waikato Environment for Knowledge Analysis* (WEKA) foi criada em 1993, com o intuito de fornecer uma ferramenta a investigadores que facilitasse o acesso a técnicas de aprendizagem automática de topo [Hal+09]. Foi construída de maneira a não só fornecer um conjunto de algoritmos de aprendizagem automática prontos a usar, mas também a oferecer uma *framework* que permitisse a implementação de novos algoritmos. Este projecto tem como objectivo fornecer, a investigadores e praticantes, uma colecção de algoritmos de aprendizagem automática e data mining e de ferramentas para pré-processar dados. A ferramenta oferece algoritmos de regressão, classificação, agrupamento, regras de associação e de selecção de atributos. É aqui descrita pois foi utilizada extensivamente na análise de dados desta dissertação. WEKA tem várias interfaces gráficas que facilitam muito a utilização da funcionalidade existente. São oferecidas três interfaces: exploração, *knowledge flow* e experimentação.

A interface de exploração é a interface principal da WEKA e é composta por vários painéis que correspondem a acções específicas, como se pode ver na figura 2.1. O primeiro painel, *"Preprocess"*, é onde é feito o carregamento dos dados e o pré-processamento destes, utilizando filtros para esse efeito. Os dados carregados podem ter vários formatos, ARFF (formato específico da WEKA), CSV, formato LibSVM[1] e C4.5; e várias fontes como ficheiros, bases de dados e URLs. O painel *"Classify"* serve para aplicar algoritmos de classificação aos dados pré-processados no painel anterior. Por omissão, o painel efectua validação cruzada aos dados, utilizando o algoritmo seleccionado, para estimar o desempenho da classificação. Também representa o resultado em forma textual e com acesso a representações gráficas (no caso do resultado ser uma árvore de decisão, por exemplo). Assim, o feedback rápido torna este processo mais intuitivo, permitindo ver a diferença da afinação das condições de entrada do algoritmo. O utilizador também pode aplicar algoritmos de agrupamento e de associação aos dados pré-processados no painel *"Clustering"*. Este painel contem estatísticas descritivas dos grupos formados e permite a visualização dos mesmos quando é possível. O número de algoritmos de associação implementados não é tão elevado como os de classificação, regressão, ou mesmo de agrupamento, mas mesmo assim os algoritmos de associação mais utilizados estão presentes. Estes algoritmos podem ser acedidos no painel *"Associate"*. O painel seguinte, *"Select Atributes"*, permite aplicar técnicas para determinar quais os atributos que mais contribuem para o modelo gerado. Por último, o painel *"Visualize"* contem uma matriz de distribuição, onde se pode explorar os resultados dos algoritmos utilizados.

Na interface *knowledge flow* é possível fazer o mesmo que na de exploração, mas estão acessíveis versões dos algoritmos com carregamento de dados incremental, e o seu *workflow* de processamento permite pré-processar uma instância dos dados de cada vez antes de os introduzir no algoritmo. Esta interface também fornece nós para a visualização e

[1] "LIBSVM is an integrated software for support vector classification, (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM). It supports multi-class classification". - http://www.csie.ntu.edu.tw/čjlin/libsvm/

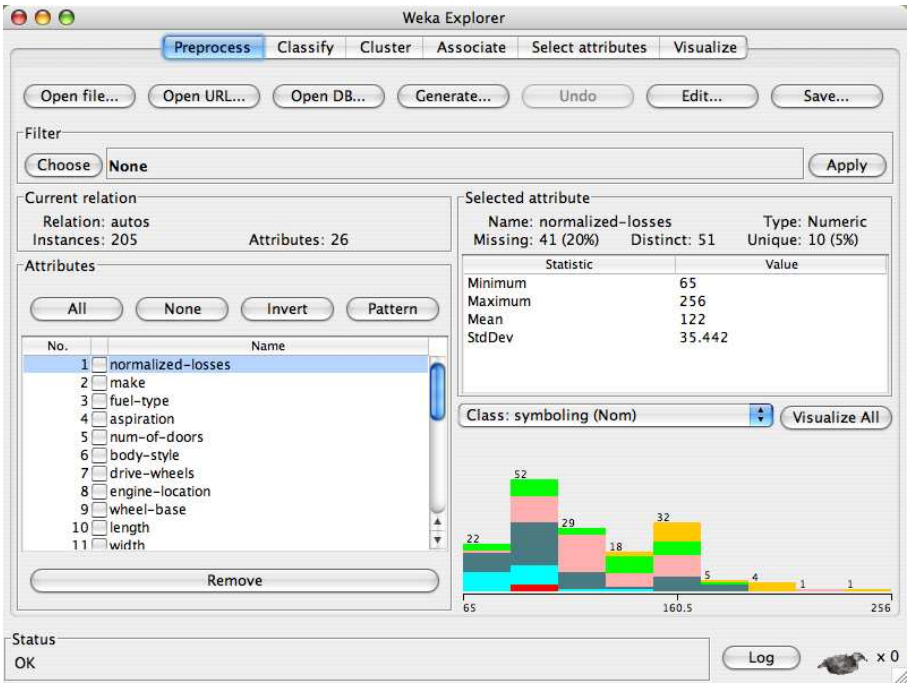


Figura 2.1: Interface de exploração do WEKA, com o painel *"Preprocess"* seleccionado.

avaliação dos resultados.

A última interface da WEKA é a de experimentação, criada com o intuito de facilitar a comparação experimental do desempenho dos algoritmos, utilizando para o efeito os vários critérios de avaliação disponíveis. É possível criar testes que corram vários algoritmos diferentes em vários data sets (utilizando validação cruzada, por exemplo) e depois guardar essas configurações para uma utilização posterior.

A primeira implementação da ferramenta foi feita em C, mas em 1999 foi descontinuada e WEKA 3.0 foi lançada, com implementação total em Java. Com a evolução da máquina virtual Java, os problemas de poder de processamento (em comparação com outras linguagens, como o C) já não se verificam e, visto que Java corre numa grande variedade de sistemas operativos e máquinas, consegue assim chegar a um maior número de utilizadores. Em 2006, a ferramenta foi patrocinada pela Pentaho Corporation[2], que a adoptou para a componente de data mining e análise preditiva da sua ferramenta de BI. Em 2008 foi lançada a última versão estável, em conjunção com a primeira edição do livro de data mining [WFH11] que tem acompanhado os lançamentos da WEKA.

2.1.4 Outras ferramentas

Em conjunção com o WEKA, foram utilizadas outras ferramentas e linguagens para analisar a informação disponível. Para além da aplicação dos algoritmos, que foi feita usando

[2]Pentaho Community - http://community.pentaho.com/

o WEKA e Java, a geração dos modelos está a cargo de transformações criadas na plataforma de ETL do Pentaho, o Data-Integration[3]. A visualização dos grupos gerados é feita em parte com gráficos disponibilizados pelo WEKA e com uma biblioteca de javascript dedicada à visualização de informação, D3.js[4].

2.2 Data Mining sobre Dados Académicos

O aumento da utilização de ferramentas de data mining no mundo académico deve muito à grande expansão que este tipo de ferramentas teve a nível empresarial. A utilização deste tipo de técnicas está sempre associado a um aumento de valor, o que se traduz em benefícios como redução de custos e aumento da produção [Lua02].

É natural a adopção e aplicação de técnicas de data mining por parte de grandes organizações corporativas se tivermos em conta as seguintes razões: estas organizações têm volumes de dados enormes a serem gerados naturalmente pelo seu funcionamento, e técnicas de data mining têm sucesso em ambientes com grandes volumes de dados para processar; a utilização destas técnicas traduz-se muitas vezes em ganhos directos bastante significativos para a organização, o que incentiva e facilita a aplicação deste tipo de técnicas; e num mundo corporativo há uma maior sensibilidade para a auto-avaliação do desempenho [Lua02]. Estas razões contrastam com as instituições académicas em níveis quase opostos. Apesar de ser no mundo académico que surgem artigos científicos com propostas de novos algoritmos e de novas aplicações desses algoritmos, é no mundo corporativo que existe o investimento para tal acontecer.

Os trabalhos apresentados em seguida estão organizados em duas secções: análises à retenção e persistência e análises ao desempenho dos alunos. A secção seguinte apresenta os trabalhos realizados no sentido de prever o comportamento dos alunos em relação à sua estadia na faculdade.

2.2.1 Análises à Retenção e Persistência

Os artigos seguintes abordam um domínio de aplicação um pouco paralelo ao que se pretende. Não são utilizadas técnicas de data mining para efectuar as análises dos dados, mas sim análises matemáticas recorrendo a técnicas como estatística e probabilidades. Mesmo assim, é importante abordar estes trabalhos, não só para analisar a abordagem e os resultados obtidos, mas também para melhor definir o domínio da solução que se está a tentar obter. As publicações apresentadas de seguida analisam a retenção e a persistência escolar. Por retenção entende-se a não conclusão do ano em que o aluno está, ou seja, o aluno ficar retido e não passar de ano. Por persistência entende-se a continuidade no curso, ou seja, a não transferência do aluno para outros cursos e a não desistência do aluno do curso.

[3]Pentaho Data-Integration - http://www.pentaho.com/explore/pentaho-data-integration/
[4]D3.js - http://d3js.org/

[FIO05] faz uma análise do sucesso e persistência escolar a dois grupos de alunos inscritos em engenharia numa universidade dos Estados Unidos, nos anos lectivos de 2000/2001 e 2001/2002. Utiliza regressões hierárquicas lineares e logísticas como ferramenta de análise e tenta relacionar atributos cognitivos e não cognitivos com a persistência dos alunos no curso de engenharia. Como atributos cognitivos são definidos a nota de candidatura e a nota média do aluno no curso, e como atributos não cognitivos são definidos o género dos alunos, a sua motivação para com o curso e a integração na instituição. A análise conclui que é possível relacionar a persistência dos alunos com os atributos definidos, e que um aluno para ter sucesso num curso de engenharia necessita de uma boa educação no secundário, de obter notas elevadas e de grande motivação.

[BAS97] analisa as atitudes e motivações dos alunos com o seu desempenho e com a retenção escolar. Como técnica de análise utiliza questionários aos alunos com perguntas acerca das atitudes perante a àrea de engenharia, das percepções do que irão aprender e das habilidades pessoais que consideram necessárias para ter sucesso em engenharia. Este estudo conclui que existe de facto uma relação entre as motivações que guiam o aluno e as atitudes que este apresenta perante o seu curso, com os problemas de retenção que se verificam.

[SN01] apresenta outra análise à retenção escolar dos alunos, estudando o impacto de características do aluno no seu desempenho no final do curso. Utiliza ferramentas estatísticas, mais precisamente regressões logísticas , e concentra-se maioritariamente em atributos não cognitivos na análise, como a idade, o género, o ambiente social, o ambiente escolar e o estado civil. Este trabalho concluí que, de facto, estas características influenciam, positivamente, o desempenho do aluno.

[TGB98], estuda quão profunda é a relação entre o género de um aluno e a retenção do mesmo em cursos de engenharia. A análise é puramente estatística e bastante centrada na questão das diferenças de géneros, analisando alunos de 17 instituições de ensino superior dos Estados Unidos da América utilizando atributos como o sexo, a nota de candidatura, a nota média do aluno e à quantos anos está na faculdade.

Estes trabalhos são apresentados com o propósito de delimitar com maior rigor o que se pretende da análise ao desempenho académico. Apesar da análise ao comportamento dos alunos ser um dos pontos centrais, as influências sociais e familiares na performance do aluno não estão dentro do âmbito. Também não se enquadra o ambiente escolar em que o aluno se encontra, para além do seu desempenho e percurso pelo curso. De seguida, apresentam-se os trabalhos recolhidos que efectuam análises ao desempenho escolar.

2.2.2 Análises ao Desempenho Escolar

Nesta secção irão ser apresentados os trabalhos que analisam o desempenho escolar dos alunos, distinguindo-se dos trabalhos da secção anterior pois utilizam técnicas de data mining no seu processo. Data mining aplicado à educação tem uma categoria própria,

data mining educacional (EDM - *Educational Data Mining*)[RV07; BY09].

[Rom+08] compara vários métodos de classificação, analisando a sua capacidade de prever se os alunos passam de ano ou não, utilizando os dados de utilização da plataforma de apoio à educação, neste caso o Moodle. Os dados de input são pré-processados e categorizados para melhor se comparar e perceber os resultados. Foi criada uma ferramenta, que se integra com o Moodle, que permite que este tipo de análise seja feito por um professor, de uma forma simplificada. A classificação dos vários algoritmos melhora se for realizada uma reamostragem dos dados e se alguns dos atributos forem categorizados antes do treino do algoritmo.

[ElH09] também utiliza técnicas de classificação e agrupamento para analisar dados de utilização do Moodle, tentando prever o comportamento dos alunos utilizadores no que toca ao aproveitamento. São utilizados algoritmos de associação para descrever a situação presente dos alunos e algoritmos de classificação para prever/descrever a situação futura. Técnicas de agrupamento são usadas no pré-processamento dos dados, para melhorar o resultado dos outros algoritmos. Os valores de fronteira detectados são avaliados de maneira especial, denotando os alunos que se encontram em casos extremos, permitindo ao professor responsável abordar estes casos sobre o comportamento fora do normal. Esta publicação sugere ainda que as técnicas utilizadas sejam incorporadas nos sistemas de *e-learning* para os utilizadores poderem tirar partido.

[Min+03] faz outra análise a um sistema de *e-learning*, o LON-CAPA[5]. Este artigo faz um trabalho semelhante aos anteriores, utilizando algoritmos de classificação para prever o desempenho do aluno analisando os dados de utilização da plataforma. Compara o desempenho de vários algoritmos de classificação, concluindo que a combinação de vários produz melhores resultados. Também conclui que se o número de atributos for reduzido, atribuir pesos a cada atributo tem melhores resultados do que escolher os atributos a utilizar na análise.

De toda a bibliografica encontrada, [Lua02] é o que melhor se enquadra no domínio desta dissertação, discutindo os aspectos teóricos e práticos das aplicações de data mining a dados do ensino superior. Apresenta um caso de uso, com os passos necessários para aplicar técnicas de data mining para estudar o aproveitamento dos alunos de determinado curso. O caso de uso utiliza uma rede neuronal e dois algoritmos de indução de regras para descrever a população estudante, e faz uma análise comparativa dos algoritmos utilizados, discutindo os resultados. Para além de apresentar uma correlação entre as necessidades de uma instituição académica e as necessidades do mundo corporativo, afirma também que a possibilidade de intervenção junto dos alunos que estão inclinados a desistir ou transferir, possibilitada pelo uso de data mining, não pode ser subestimada e tem grande valor para a instituição.

[BP11] tenta prever o sucesso escolar dos alunos utilizando árvores de decisão. Para facilitar o uso e interpretação do algoritmo de classificação todos os atributos são categorizados antes da análise. O modelo criado detecta os alunos com grande probabilidade

[5]The LearningOnline Network with CAPA - http://www.lon-capa.org/

de desistir no curso, o que permite a intervenção dos professores antes disso acontecer.

[VMS07] procura classificar, o mais cedo possível, os alunos que entram na instituição consoante o risco de desistirem de curso. São comparadas três técnicas diferentes para classificar os alunos em três classes diferentes (baixo risco de desistir, médio risco e alto risco). As técnicas usadas são redes neuronais, árvores de decisão e uma análise discriminante linear[6]. Os dados utilizado são divididos em 30% para validação e 70% para treino. A validação dos resultados, para todas as técnicas, fica entre os 40% e os 50%, ou seja, o algoritmo treinado consegue classificar correctamente entre 40 a 50% dos estudantes.

Por último, [Ogo07] testa algumas implementações de redes neuronais e de árvores de decisão para estudar o progresso e aproveitamento dos alunos de um curso do ensino superior. A utilização de ferramentas de data mining é recomendada para acompanhar e avaliar correctamente o desempenho escolar, recomendando também a criação de um data warehouse com os indicadores utilizados para medir o desempenho escolar. A implementação OLAP com ferramentas de reporting é utilizada pelo autor para criar relatórios de desempenho que automatizam o processo de minar os dados.

Resumindo, técnicas de data mining são certamente indicadas para a análise ao desempenho escolar que se pretende fazer. A capacidade de descobrir informação oculta será bastante importante para detectar perfis de desempenho académico nos alunos.

Em relação ao processo de aplicação das técnicas de data mining, um dos passos que consome mais tempo é a criação do modelo. Definir os atributos mais adequados para descrever um aluno é importante para assegurar a correcta detecção de padrões e a utilização de mais do que um algoritmo para minar os dados irá assegurar resultados mais apurados. Em análises ao desempenho académico com o propósito de detectar alunos em risco (seja de transferir, desistir, ou não aprovar), são utilizadas geralmente técnicas de classificação com técnicas de agrupamento para pré-processar os dados. Devido à natureza preditiva dos algoritmos de classificação, esta abordagem resulta bastante bem mas apenas para este propósito. Para esta dissertação a abordagem terá de ser diferente pois não se sabe à partida as classes a que os alunos podem pertencer (o que invalida a utilização de algoritmos de classficação). Algoritmos de agrupamento serão a principal ferramenta no estudo do comportamento dos alunos.

[6] Análise Discriminante Linear - http://en.wikipedia.org/wiki/Linear_discriminant_analysis

3

Detecção de Padrões de Desempenho Académico

Neste capítulo irá ser apresentado o trabalho que motivou esta dissertação, a detecção de padrões de desempenho académico em alunos do ensino superior. Foram desenvolvidas duas abordagens diferentes para o problema apresentado: a primeira que analisa o desempenho dos alunos num ano lectivo, e a segunda que analisa o percurso académico dos alunos enquanto permanecem no curso.

A secção seguinte irá descrever a metodologia utilizada e onde esta se enquadra nos vários modelos de KDDM (*Knowledge Discovery and Data Mining*) já existentes. Também apresenta os métodos que vão ser usados para avaliar os agrupamentos formados. As duas secções seguintes descrevem as abordagens: a secção 3.2 descreve a abordagem ao desempenho académico e a secção 3.3 descreve a abordagem ao percurso do aluno. Cada uma destas secções começa com uma discussão prévia dos modelos analisados e descreve em pormenor o processo utilizado com cada modelo.

3.1 Metodologia

A metodologia seguida nesta análise para avaliar os modelos propostos baseia-se num dos processos de KDDM mais utilizados no mundo académico. É um processo com seis passos apresentado pela primeira vez por [Cio+00] e considerado um dos mais importantes modelos de KDDM existentes [KM06]. Os passos deste processo são: perceber o domínio do problema, perceber os dados que estão disponíveis, preparar os dados disponíveis para serem analisados, aplicar as técnicas de data mining, avaliar o conhecimento adquirido e utilizar esse conhecimento.

Tendo em conta que a percepção do domínio do problema e dos dados que estão disponíveis já foi feita nos capítulos anteriores, o protocolo específico que irá ser descrito nas secções seguintes começa por escolher o algoritmo a utilizar. Depois da escolha do algoritmo, é realizada uma análise preliminar com o WEKA para perceber a influência dos parâmetros do algoritmo no agrupamento. Esta análise permite também ter uma noção dos tipos de agrupamento que o algoritmo forma com o modelo dado. A importância desta análise está no facto de não se saber à partida se existem padrões nos dados nem de que forma esses padrões se manifestam.

O passo seguinte é criar as ferramentas necessárias para melhor interpretar os resultados do modelo e iterar esse modelo no sentido de optimizar os resultados obtidos. Esta fase é bastante iterativa, onde é aplicado o algoritmo aos dados com um conjunto variado de parâmetros até se ter dados suficientes para analisar. O estudo destes dados pode eventualmente motivar a criação de novas visualizações para melhor compreensão dos resultados, o que oferece uma nova perspectiva sobre o agrupamento.

Quando forem seleccionados o modelo e o conjunto de parâmetros para cada algoritmo utilizado que ofereçam os melhores resultados, é feita a validação dos padrões detectados. Se nesta fase surgirem ideias para novos modelos ou visualizações, é com facilidade que se reinicia o processo e se introduz essa informação na próxima análise.

Avaliação dos Modelos

A validação e o estudo dos agrupamentos obtidos é feito de uma forma analítica, recorrendo às métricas descritas na secção 2.1.2; e de uma forma visual, através dos gráficos disponibilizados pelo WEKA e também de gráficos criados especificamente para uma determinada abordagem.

As métricas permitem não só fazer uma análise comparativa entre corridas do mesmo algoritmo aplicado a um modelo, mas também comparar resultados entre o algoritmo aplicado a modelos diferentes. Pela leitura das métricas só se consegue inferir a qualidade do agrupamento, a existência ou não de um padrão nesse agrupamento é uma questão subjectiva (que só faz sentido para quem procura o padrão) e não consegue ser respondida analiticamente.

É através da análise dos gráficos que se consegue perceber se o agrupamento formado é relevante ou não. Para além de um gráfico a três dimensões (do WEKA), em que cada dimensão e a cor dos pontos codifica um atributo dos dados; também foram utilizados gráficos de duas dimensões e gráficos de paralelas, para oferecer outra perspectiva sobre os resultados. A figura 3.1 mostra dois exemplos dos gráficos utilizados que foram desenvolvidos especificamente para este trabalho.

Para o algoritmo *k-means* é utilizado um método específico para avaliar o melhor valor de k, o método do cotovelo [KS96] . Este método consiste em desenhar uma função em que para cada valor de k se marca a soma do erro quadrático dos grupos para esse agrupamento. Chama-se método do cotovelo precisamente porque é escolhido o valor de

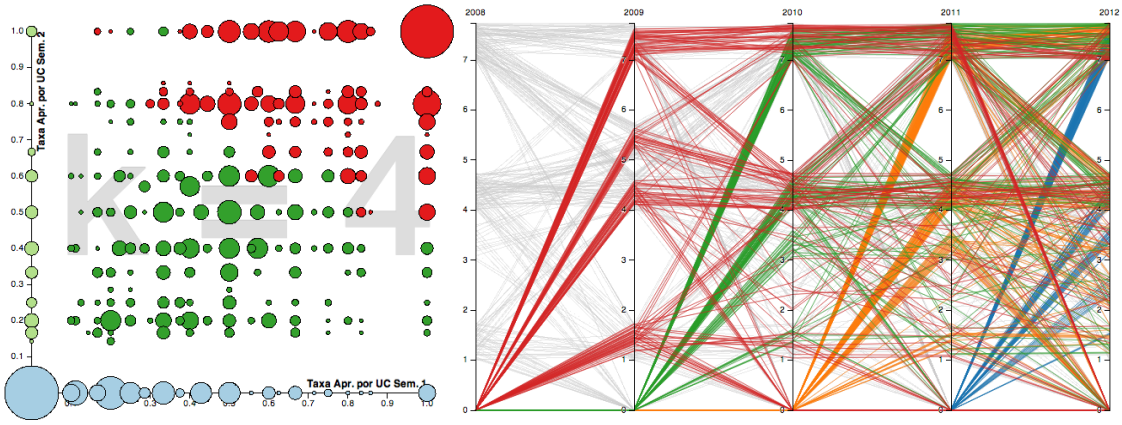


Figura 3.1: Exemplo de um gráfico 2-D (esquerda) e de um gráfico de paralelas (direita).

k que corresponde ao cotovelo da função, o que significa que a adição de mais um grupo não reduz de forma significativa o valor do erro.

Os gráficos de paralelas apresentados neste documento podem ser consultados no endereço http://pessoa.fct.unl.pt/l.ramos.

3.2 Análise ao Desempenho Académico

A abordagem descrita nesta secção pretende analisar o desempenho académico dos alunos durante um ano lectivo do curso. Esta abordagem considera o número de unidades curriculares a que o aluno se inscreveu e aprovou e ainda os resultados médios que obteve nessas UCs. A secção seguinte introduz uma discussão dos vários modelos possíveis de ser aplicados a esta abordagem, especificando os que irão ser analisados. Também descreve brevemente as conclusões das análises preliminares que tiveram lugar para testar a viabilidade dos modelos protótipo sugeridos. A secção 3.2.2 apresenta o protocolo da aplicação do *k-means* ao modelo escolhido para ser analisado, descrevendo a lógica por de trás de algumas decisões chave no processo e apresentando também os resultados obtidos. No final da secção é feita a validação do agrupamento resultante.

3.2.1 Discussão Prévia

Esta abordagem pretende analisar o desempenho de um ano lectivo do aluno, estando disponíveis para análise todos os anos lectivos desde Bolonha (2006/07) até ao ano lectivo de 2011/12. Foram considerados para esta abordagem dois modelos, o modelo de desempenho anual e o modelo de desempenho semestral. No modelo de desempenho anual resume-se a actividade académica considerando o número total de unidades curriculares inscritas durante o ano lectivo, enquanto no modelo de desempenho semestral consideram-se as unidades curriculares inscritas em cada um dos dois semestres. Inicialmente, foi apenas considerado o modelo anual mas depois de uma análise inicial ao

modelo conclui-se que seria interessante tentar modelar o aluno decompondo o seu desempenho por semestres. Apesar do ciclo de estudos ser anual, existe a percepção de que o comportamento dos alunos nos dois semestres é diferente, até devido a factores climáticos, sociais ou ao simples facto de um semestre ser anterior ao outro.

O modelo de desempenho semestral terá cada atributo descrito duas vezes, uma para resumir esse resultado no primeiro semestre e outro para o segundo semestre. Os atributos considerados para descrever o aluno são:

Estatísticas das notas As estatísticas das notas incluem o valor máximo, mínimo e médio conseguido pelo aluno nesse período. Estes valores são os indicadores base para a avaliação do desempenho do aluno, sendo utilizados em quase todas as análises ao desempenho revistas no trabalho relacionado;

Número de inscrições e obtenções As inscrições e obtenções são representadas em número de unidades curriculares e número de ECTS. A utilização destes indicadores distinguindo as inscrições/obtenções em UCs e em ECTS provem da possibilidade de dois alunos estarem inscritos ao mesmo número de unidades curriculares, mas o esforço associado a essas inscrições ser bastante diferente. O inverso também se pode verificar, ou seja, alunos inscritos ao mesmo número de ECTS terem um número diferente de inscrições a UCs;

Taxa aprovação A taxa de aprovação é outro atributo bastante utilizado em análises ao desempenho. É calculado dividindo o número de obtenções pelo número de inscrições. Este indicador permite aproximar alunos com um número de inscrições e de obtenções diferentes mas com a mesma taxa. A taxa de aprovação também é representada em ECTS e em UCs.

Depois de feita uma análise preliminar para avaliar a estabilidade destes protótipos, o modelo anual não foi escolhido porque não apresentava uma agrupamento satisfatório. Os grupos formados eram de difícil leitura e não seria sensato prosseguir a análise com este modelo visto o modelo semestral mostrar um agrupamento mais equilibrado.

Para além da escolha do modelo semestral, foi pensado expandir o conjunto de atributos com mais indicadores, numa tentativa de melhorar o agrupamento. Os indicadores utilizados foram: o ano curricular, se o aluno é caloiro ou não e se um aluno obteve aproveitamento nesse semestre (se obteu o número mínimo de créditos estabelecido no plano de estudos). Estes novos modelos foram descartados depois da análise preliminar ditar que o impacto que cada um tem no agrupamento formado não é relevante o suficiente para justificar o seu uso. Ou seja, os grupos formados por estes novos modelos eram bastante semelhantes aos grupos formados pelo modelo semestral original.

Também é necessário fazer algumas considerações sobre os dados antes de iniciar a análise. Visto estarem disponíveis vários anos lectivos para analisar é preciso escolher por onde começar a análise. Esta decisão tem de ter em conta os seguintes pontos: quanto mais informação estiver disponível melhor o agrupamento formado; os primeiros anos

lectivos vão ter uma população mal distribuída pelos três anos curriculares, o que pode influenciar de alguma maneira os resultados, e esses anos vão ter um número reduzido de alunos inscritos, pois só são considerados os alunos de Bolonha nesta análise.

Tendo em conta estes pontos, fazia sentido começar a análise pelo último ano lectivo disponível, mas o modelo escolhido permite uma abstracção do ano lectivo e permite combinar a população de cada ano disponível num ano lectivo global. Este ano lectivo global vai ser analisado pelos vários algoritmos e a análise a cada ano lectivo é feita posteriormente, para validar os resultados obtidos com o ano lectivo global. A secção seguinte apresenta a aplicação de alguns algoritmos ao modelo de desempenho semestral escolhido.

3.2.2 Protocolo Experimental

O primeiro passo neste processo de descoberta de padrões é escolher o algoritmo que vamos utilizar para analisar o modelo. A escolha do *k-means* como o primeiro a ser utilizado não tem nenhuma razão extraordinária, sendo um algoritmo simples de aplicar (com pouca parametrização) e que permite uma iteração rápida. A secção seguinte apresenta a aplicação do algoritmo, mostrando também os resultados obtidos e incluindo uma descrição semântica do melhor agrupamento encontrado. Também foram aplicados outros algoritmos ao modelo semestral, cujos resultados são apresentados na secção 3.2.2.2. A última secção mostra a validação do agrupamento obtido, comparando os grupos obtidos com o ano lectivo global aos grupos obtidos da análise a cada ano lectivo.

3.2.2.1 Algoritmo k-means

O *k-means* recebe como parâmetro o número de grupos que se pretende encontrar. Isto apresenta um problema à partida, pois o número de padrões que estão presentes nos dados é precisamente o que se procura. Para resolver esta questão foi utilizado o WEKA e foi feito um estudo do melhor valor para o parâmetro.

É possível determinar o melhor valor de k escolhendo um intervalo de valores possíveis e analisando os resultados de cada corrida do algoritmo, utilizando as métricas de qualidade definidas anteriormente. O intervalo escolhido foi entre 3 e 12, tendo em conta que não seria útil procurar por um número de grupos inferior a três (não iria revelar informação nova). O valor máximo foi estabelecido como 12 pois o algoritmo encontra sempre o número de grupos dado como parâmetro (mesmo que não existam nos dados) e valores maiores não teriam resultados interessantes.

Utilizando a soma do erro quadrático de cada grupo é possível efectuar a análise do cotovelo (figura 3.2) que permite encontrar visualmente o melhor valor de k. A figura 3.2 mostra também a diferença do erro de um k para o anterior. O que se procura é o ponto onde a redução do erro deixa de ser significativa, o que indica que o algoritmo está a entrar em overfitting e os grupos já não introduzem nova informação ao agrupamento. Isto verifica-se com $k = 8$, em que o erro decresce em cerca de 27%, face à descida de 4%

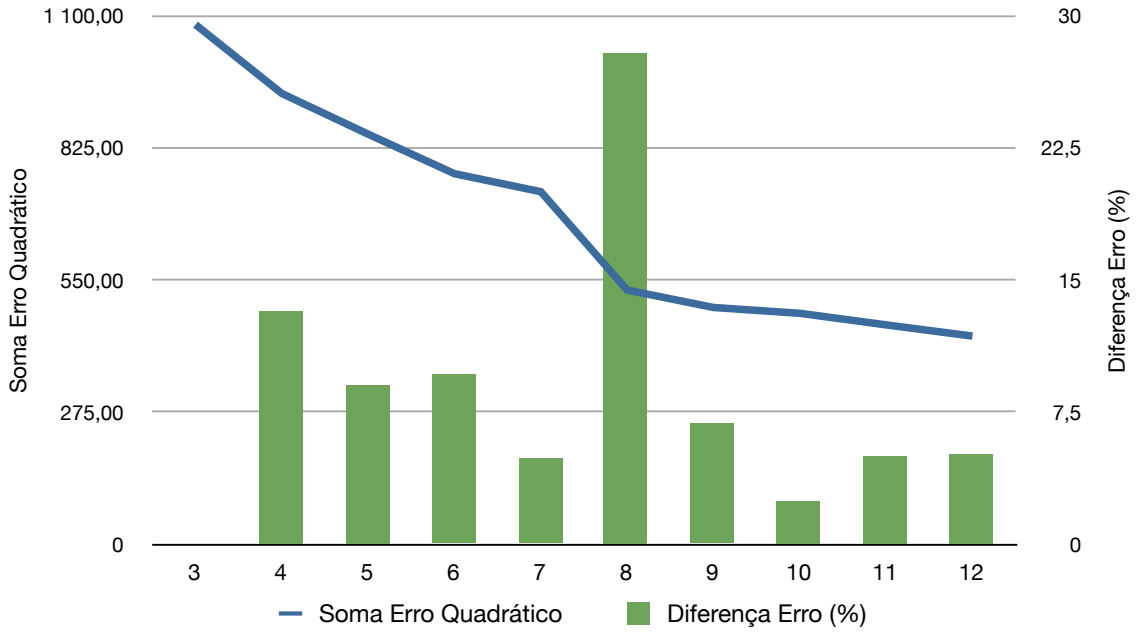


Figura 3.2: Gráfico da soma dos erros quadráticos de cada corrida do algoritmo para um k entre 3 e 12.

de $k = 6$ para $k = 7$ e de 7% de $k = 8$ para $k = 9$. Este pico marca o valor mais indicado para o algoritmo.

Para confirmar visualmente se este valor era o mais correcto, foram gerados os dados para os valores de k entre 3 e 12 e foi desenvolvida uma visualização a duas dimensões que permite analisar os grupos obtidos. Esta visualização representa os alunos por pontos num gráfico em que os eixos são as taxas de aprovação nos dois semestres. O tamanho de cada ponto representa o número de alunos com esse par de valores e a cor representa o grupo que tem mais representantes nesse ponto.

A figura 3.3 mostra uma representação dos dados para $k = 8$, que coloca os alunos consoante a sua taxa de aprovação em UCs no primeiro semestre versus a taxa de aprovação em UCs no segundo semestre. As linhas azuis delimitam de uma forma genérica os grupos. É possível distinguir com facilidade três grupos em cima dos eixos: o grupo na origem do gráfico é de alunos com mau desempenho em ambos os semestres e os outros dois grupos são de alunos com baixo desempenho num dos semestres (a taxa de aprovação é perto de 0) mas que no outro semestre têm uma taxa de aprovação mais elevada. Uma versão interactiva deste gráfico (e outros apresentados neste capítulo) está disponível em http://pessoa.fct.unl.pt/l.ramos.

Do estudo dos dados gerados pode-se fazer duas observações: parece existir uma grande dependência dos grupos formados às taxas de aprovação, como se estes atributos ditassem o agrupamento; e é possível criar uma semântica para descrever os grupos com base na taxa de aprovação. Para excluir a hipótese de que o agrupamento estava dependente da taxa de aprovação, foram criados três modelos apenas com dois atributos. Um só com a taxa de aprovação em UCs do primeiro e segundo semestre, outro com a taxa

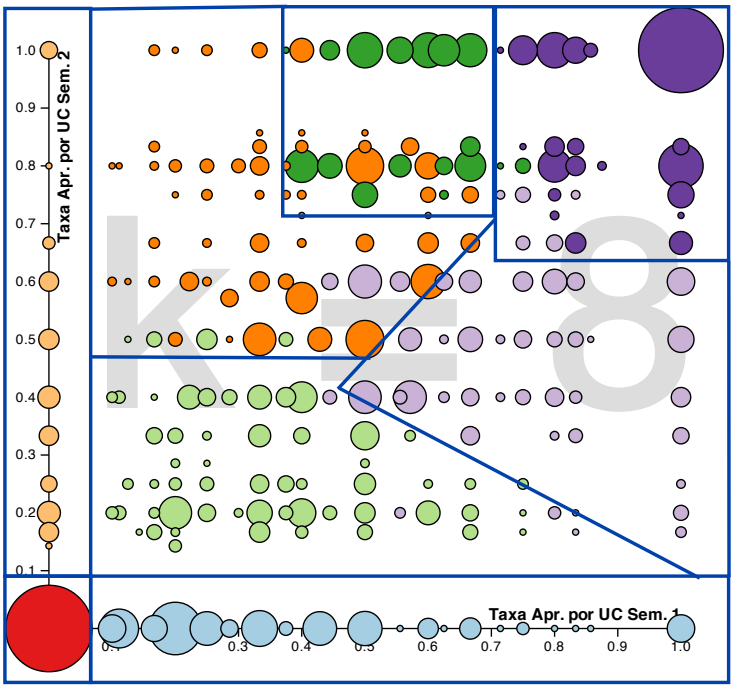


Figura 3.3: Agrupamento *k-means* para $k = 8$.

de aprovação em ECTS e ainda outro com o número de ECTS obtidos. Foi desenvolvida uma visualização que compara estes novos modelos com o modelo anterior e, como é possível ver na figura 3.4, os grupos formados pelos modelos de dois atributos são bastante diferentes do agrupamento original, o que verifica que não existe uma dependência única nas taxas. Cada cor representa um grupo diferente e o tamanho dos círculos in-

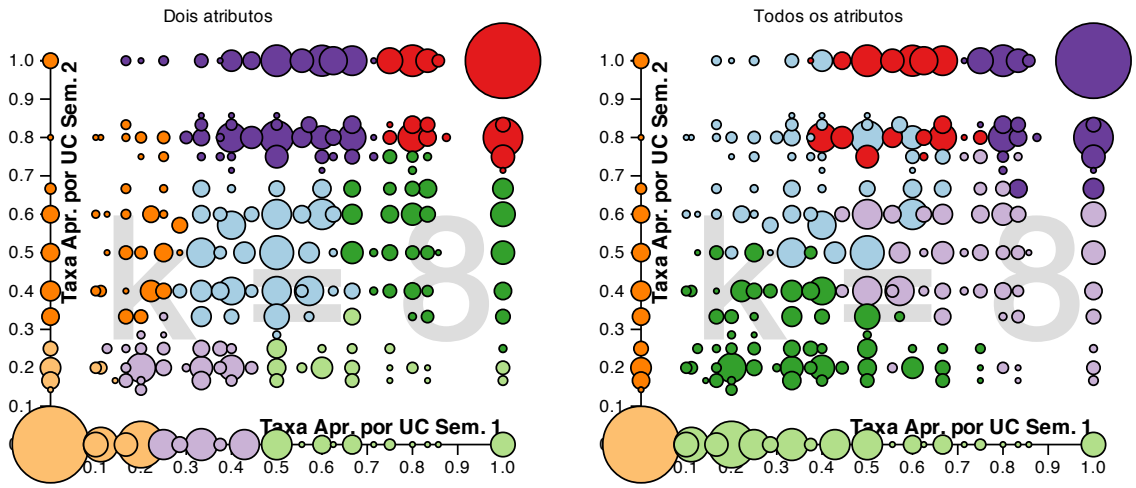


Figura 3.4: Comparação do agrupamento *k-means* para $k = 8$ com apenas a taxa de aprovação (esquerda) e com todos os atributos (direita).

dica o número de alunos nesse ponto. As cores de um gráfico para o outro não indicam grupos iguais.

Para ser possível criar uma semântica que descreva os grupos é necessário fixar o

k	Coesão	Afastamento	Silhueta	Índice Davies-Bouldin	Índice Dunn
3	19,497	19,427	-0,00300	2761,556	0,0008
4	18,598	18,512	-0,00414	2080,205	0,0009
5	17,872	17,726	-0,00732	1698,525	0,0009
6	17,224	17,030	-0,01089	1407,319	0,0011
7	17,043	16,841	-0,01221	1195,923	0,0010
8	15,627	15,429	-0,01239	1170,106	0,0012
9	15,347	15,139	-0,01325	1114,593	0,0011
10	15,161	14,929	-0,01461	1046,574	0,0009
11	14,856	14,639	-0,01384	940,430	0,0011
12	14,641	14,435	-0,01297	854,325	0,0013

Tabela 3.1: Métricas para os agrupamentos do algoritmo *k-means*.

valor de k para qual o algoritmo apresenta o melhor agrupamento. Para além da análise do cotovelo e da análise aos gráficos disponíveis foram também utilizadas as métricas na tabela 3.1, que não apontam para um valor concreto. Como se pode observar todas as métricas vão ficando progressivamente melhores à medida que o valor de k aumenta e é apenas conjugando as várias análises realizadas que se pode considerar $k = 8$ como o valor mais adequado para correr o algoritmo.

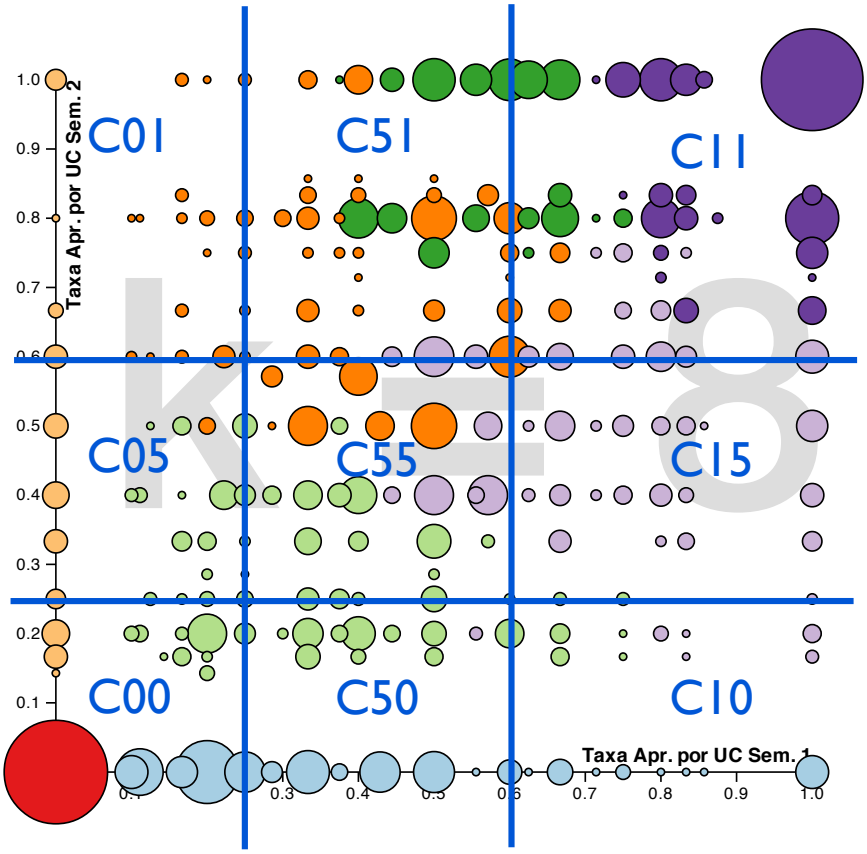


Figura 3.5: Semântica descritiva aplicada ao gráfico do agrupamento *k-means* com $k = 8$.

A semântica criada para descrever os grupos formados pelo algoritmo tem como base

a taxa de aprovação em UCs dos alunos. Esta semântica têm nove classificações com a forma "CXX", onde o primeiro X representa o desempenho no primeiro semestre, o segundo X representa o desempenho no segundo semestre e em que os valores que o X pode tomar são: "0" para alunos com baixo desempenho nesse semestre, "5" para alunos de médio desempenho e "1" para alunos de alto desempenho. Os intervalos que distinguem os vários desempenhos podem ser alterados para se adaptar ao agrupamento formado, mas neste caso baixo desempenho tem uma taxa de aprovação desde 0 a 0.25, médio desempenho vai de 0.25 a 0.6 e alto desempenho é para taxas acima dos 0.6. Por exemplo, um grupo descrito como "C05" indica que os alunos desse grupo tiveram um baixo desempenho no primeiro semestre e médio desempenho no segundo semestre.

Se aplicarmos esta semântica ao gráfico 2-D das taxas de aprovação, os grupos podem ser descritos pela sua localização na grelha, e o resultado está na figura 3.5. Esta definição não entra em demasiada especificidade para poder ser aplicada em outros contextos sem perder utilidade. Verifica-se que a semântica não encaixa em todos os grupos na perfeição, sendo difícil de discernir visualmente a que secção alguns grupos pertencem. Isto não se verifica para os grupos na origem e no extremo (1,1), mas para os outros grupos é necessário recorrer ao centro conceptual do grupo, ou seja, ao aluno médio, para que a semântica se aplique com mais à vontade.

3.2.2.2 Análise outros algoritmos

Para além do *k-means*, foram também aplicados outros algoritmos que não apresentaram resultados tão bons. Na tentativa de variar a maneira como os alunos eram agrupados, foram utilizados algoritmos de densidade para calcular novos agrupamentos.

Os primeiros resultados foram obtidos com o DBScan, que não mostrou nova informação nos seus grupos. Alguma optimização dos parâmetros pedidos pelo algoritmo, o raio de vizinhança ($epsilon$) e o número mínimo de pontos numa vizinhança ($minPoints$), não melhoraram os resultados pois o algoritmo perdia-se com as diferenças de densidade existentes e retornava na maioria das corridas apenas quatro grupos e uma grande parte de alunos eram classificados como ruído.

Ainda mantendo a ideia de que uma algoritmo de densidade poderia apresentar novos resultados, o DBScan foi trocado pelo SNN, que é mais resistente a variações elevadas de densidade. Este algoritmo recebe como argumento o tamanho da lista de vizinhos (k), o valor limite para a densidade (eps) e o valor limite para um ponto ser considerado *core point* ou não ($minPts$). Depois de alguma iteração e tentativa de optimização dos parâmetros, utilizou-se a heurística recomendada em [MSC05] e o valor de eps e $minPts$ são 30% e 70% do valor de k, respectivamente. Mesmo assim os resultados obtidos não foram satisfatórios e para o melhor agrupamento detectado, com um $k = 12$, eram encontrados 73 grupos diferentes e 37% dos alunos eram classificados como ruído. Se o valor de k fosse maior, o número de grupos reduzia mas continuava elevado demais para se perceber qualquer tipo de padrões. Para valores menores de k, o número de grupos aumentava

a ponto de não detectar qualquer tipo de comportamento relevante. A percentagem de alunos detectada como ruído mantinha-se a mesma independentemente do valor de k.

3.2.2.3 Validação do Modelo

Como os outros algoritmos utilizados não apresentaram resultados interessantes, prosseguiu-se com a validação dos resultados obtidos pelo *k-means*. Como já foi dito, foi utilizado um ano lectivo global para efectuar o primeiro agrupamento e em seguida vamos validar o melhor agrupamento encontrado, comparando a análise global a cada ano lectivo disponível.

Os resultados para cada ano variam ligeiramente, mas um $k = 7$ parece ser o mais indicado. Os anos lectivos 2006/2007 e 2007/2008 não apresentam os melhores resultados, como se verifica na figura 3.6, mas isto pode ser explicado por um overfitting do algoritmo aos dados. Como a população desses dois anos é reduzida e não está bem distribuída pelos anos curriculares, o algoritmo tenta encontrar mais grupos do que realmente existem, criando grupos que fazem menos sentido.

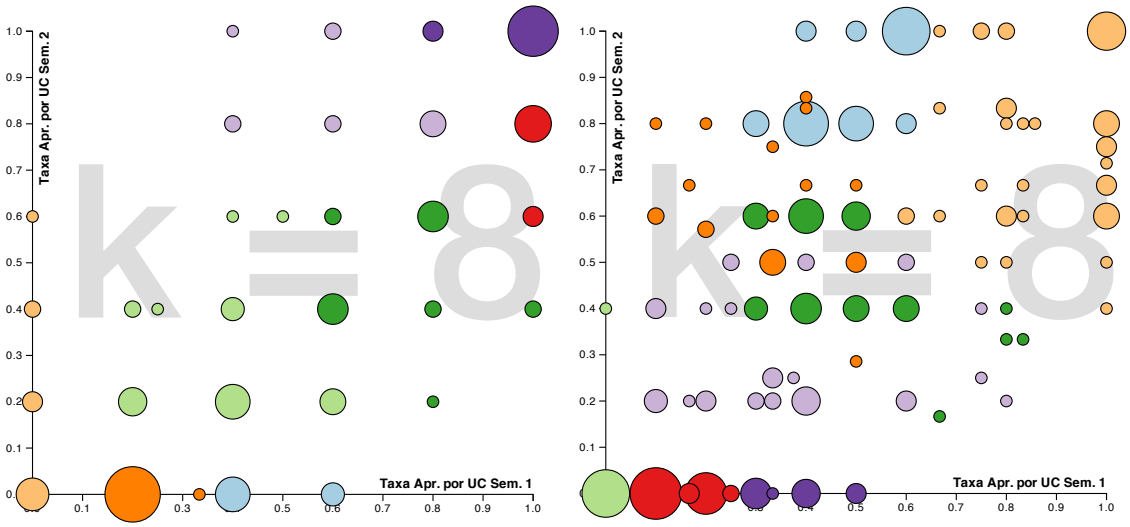


Figura 3.6: Agrupamento *k-means* para os anos lectivos de 2006/07 e 2007/08, respectivamente, com $k = 8$.

A figura 3.7 mostra a comparação do agrupamento de cada ano lectivo com o agrupamento global calculado antes. A variação dos resultados de ano para ano é reduzida a ponto de ser possível identificar em cada ano os grupos encontrados no ano lectivo global. Também é possível detectar a continuação de alguns grupos pelos anos lectivos, ou seja, de um ano lectivo para o seguinte é possível ver que um grupo com as mesmas características foi detectado.

Os três grupos de alunos com baixo desempenho registam-se em todos os anos, quase não variando em relação ao ano lectivo global. O grupo de alunos de baixo desempenho no primeiro semestre no ano lectivo global está repartido em dois grupos nos anos lectivos de 2009/10 e 2010/11, o que pode indicar algum overfitting nesses anos para $k = 8$.

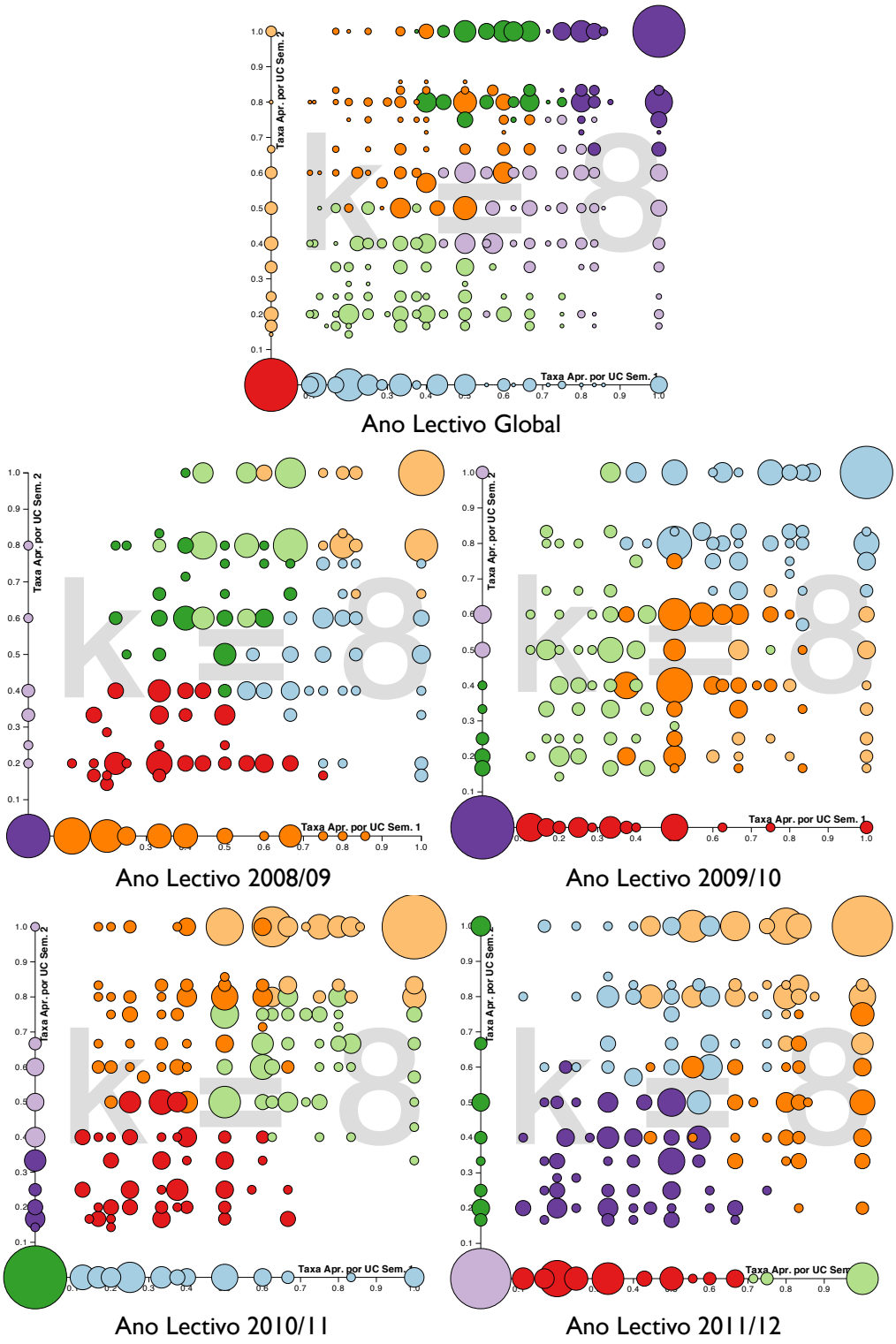


Figura 3.7: Comparação entre os agrupamentos *kmeans* ano lectivo global e anos lectivos de 2008/09 a 2011/12.

O mesmo se sucede no ano lectivo de 2011/12, mas desta vez com o grupo de alunos de baixo desempenho no segundo semestre. O grupo de alunos de alto desempenho no ano lectivo global também está presente em todos os anos lectivos, com uma ligeira

"expansão"no ano lectivo de 2009/10. Apenas os grupos de médio desempenho, que se encontram na região média do gráfico, é que apresentam variações maiores em relação aos grupos do ano lectivo global.

Considerando os padrões detectados em cada ano lectivo e a semântica criada para os descrever, é possível ainda tirar algumas conclusões sobre a evolução desses padrões de ano para ano. Aplicando a semântica aos grupos de cada ano lectivo, consegue-se chegar ao gráfico de paralelas presente na figura 3.8 que mostra a distribuição de duas fornadas de alunos pelos grupos de cada ano lectivo. O ano lectivo de 2006/07 não está presente pois, como já foi dito, os grupos detectados nesse ano não permitem a aplicação da semântica desenvolvida. A leitura deste gráfico diz-nos que os alunos que estão em grupos de alto desempenho assim se mantêm de ano para ano (até eventualmente acabarem o curso ou transferirem/desistirem). Também é possível ver a tendência de alunos na terceira matrícula e que estavam em grupos de médio desempenho alterarem para grupos com melhor desempenho. De notar que a leitura deste gráfico é feita com melhores condições no site http://pessoa.fct.unl.pt/l.ramos.

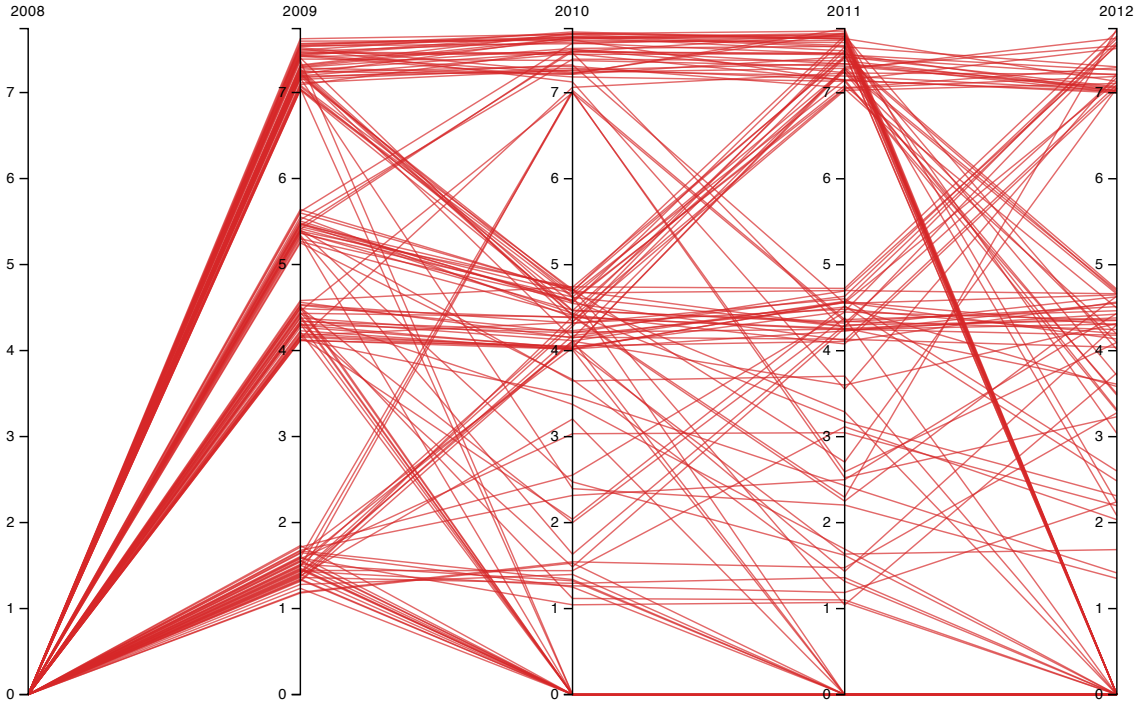
3.3 Análise ao Percurso Académico

Esta secção descreve a análise ao percurso académico de um aluno ao longo da sua estadia no curso. Este percurso é definido através de indicadores como o número de inscrições e o último resultado obtido numa unidade curricular. A secção seguinte apresenta algumas considerações a fazer antes de se descrever o protocolo experimental seguido. Irão apresentados os dados e discutidos os modelos que vão ser utilizados na análise, incluindo alguma discussão sobre os modelos que não foram escolhidos. A aplicação dos algoritmos *k-means* e SOM ao modelo protótipo escolhido está descrita na secção 3.3.2, que também inclui a validação dos grupos obtidos.

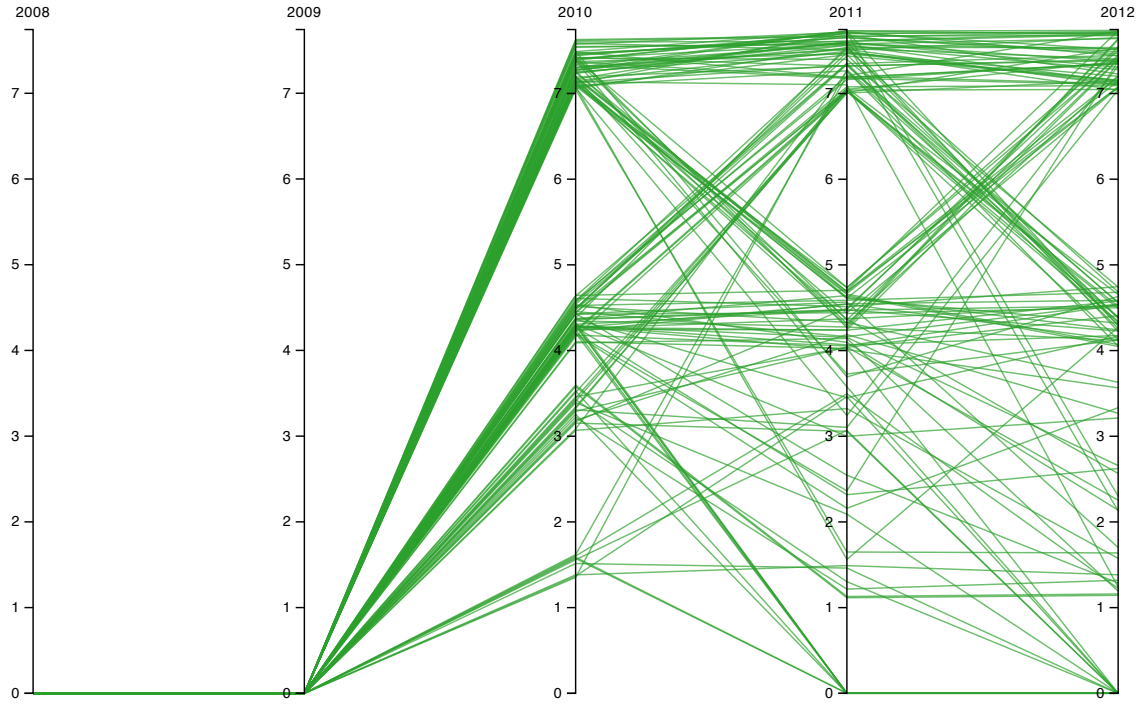
3.3.1 Discussão Prévia

A abordagem ao percurso académico de um aluno é definida fazendo uso das unidades curriculares presentes no plano de estudos, mais especificamente das inscrições e dos resultados obtidos nessas UCs. Visto o objectivo ser detectar padrões no percurso dos alunos, analisar alunos com menos de três matrículas que não tiveram ainda oportunidade de completar o curso pode comprometer os resultados. Assim, só foi extraída a informação dos alunos pós-Bolonha com três matrículas ou mais, o que dá um grupo de cerca de 500 estudantes.

O modelo proposto para este trabalho modela o aluno com base numa lista de unidades curriculares. Esta lista pode variar, podendo conter todas as disciplinas do plano de estudos ou apenas unidades curriculares da área de informática ou até apenas unidades da área de matemática. Como o objectivo desta abordagem é analisar os comportamentos dos alunos tendo em conta o curso na sua totalidade e, ao mesmo tempo, não criar



(a) Ano lectivo 2008/09



(b) Ano lectivo 2009/10

Figura 3.8: Distribuição dos grupos detectados de duas fornadas pelos anos lectivos seguintes.
Legenda: 1 - C00; 2 - C50; 3 - C05; 4 - C55; 5 - C15; 6 - C51; 7 - C11; 0 - Alunos sem grupo.

um modelo demasiado complexo, foram escolhidas as unidades curriculares obrigatórias

presentes no plano de LEI. Esta lista perfaz o total de 27 unidades curriculares, correspondendo a 90% dos ECTS para obtenção de diploma, cobrindo os seis semestres e os três anos curriculares da licenciatura.

Dos dados disponíveis, foi extraída a seguinte informação para cada unidade curricular presente no modelo: o número de inscrições do aluno nessa UC, o ano lectivo da primeira inscrição e o ano lectivo da última inscrição na UC e o último resultado do aluno à disciplina (seja uma aprovação ou melhoria). Pode ser calculada a diferença das datas da última e primeira inscrição para criar um indicador que abstrai o ano lectivo do modelo. Assim é possível analisar comportamentos semelhantes com alunos que passaram pelo curso em tempos diferentes. O número de unidades curriculares obrigatórias no plano curricular é 27 e se cada uma for modelada com três atributos, o modelo resultante tem 81 atributos.

Um modelo com um elevado número de atributos deixa de ser útil. Em alternativa, foram consideradas outras abordagens de maneira a diminuir o número de atributos. A primeira é efectuar um agrupamento em duas fases e a segunda é resumir os dados de uma unidade curricular apenas em um atributo, criando um índice.

O agrupamento por fases passa por criar três modelos para o aluno, em que a lista de unidades curriculares de cada modelo é composta por apenas um dos atributos descritos acima. Cada modelo gerado é agrupado e em seguida o aluno é modelado consoante os grupos em que se encontra. Este novo modelo é outra vez agrupado em busca de padrões de desempenho.

Depois de realizada alguma análise preliminar a este modelo, foi eliminado pois a interpretação da primeira fase de agrupamento não era fácil e comprometia a leitura dos grupos da segunda fase. Ou seja, os grupos encontrados pelo primeiro agrupamento realizado não revelavam nenhum padrão interessante e não era possível definir uma semântica que os descrevesse. Isto comprometia a leitura da segunda fase de agrupamento, visto estar bastante dependente da descrição dos grupos formados pela primeira fase. A complexidade em entender os grupos gerados e a dificuldade em criar uma visualização que permiti-se uma melhor leitura dos grupos levaram ao abandono deste modelo.

A segunda alternativa considerada, o modelo de índice, resume a informação de uma unidade curricular apenas num valor. Para atingir os resultados pretendidos, era necessário que o índice contivesse a informação do número de inscrições, da diferença entre as datas da primeira e última inscrição e do último resultado obtido. Para simplificar a criação do índice, a diferença entre datas não foi utilizada pois ao analisar as inscrições dos alunos, é habitual um aluno inscrever-se a todos os ECTS que tem disponíveis (mesmo que saiba à partida que não vai se dedicar a uma disciplina) e por isso a diferença entre primeira e última inscrição numa unidade curricular vai ser quase sempre igual ao número de inscrições. Assim, a primeira versão do índice regista na casa das dezenas o número de inscrições que o aluno têm na unidade curricular, e na casa das unidades o último resultado do aluno (alunos que não aprovaram ficam com nota 0).

Depois de alguma experimentação com o modelo, concluiu-se que este aproximava

alunos com notas de aprovação baixas de alunos que ainda não tinham aprovado à unidade curricular. Como a aprovação é um factor importante no desempenho de um aluno, alterou-se o modelo original para acentuar esta diferença. Em vez de os alunos terem nota 0 quando ainda não aprovaram, passam a ter um valor negativo igual ao número de inscrições. Por exemplo, dois alunos com três inscrições a uma unidade curricular, mas que o primeiro aprovou com 11 e o segundo ainda não aprovou, no modelo original têm um valor de 31 e 30, respectivamente, e no novo índice negativo irão passar a ter um valor de 31 e -3. Esta diferença permite detectar com maior exactidão comportamentos distintos.

A visualização criada para a análise ao modelo de desempenho académico não pode ser reutilizada para esta análise e por isso foi criado um novo gráfico para visualizar os resultados deste modelo. Este gráfico representa o índice de cada unidade curricular como um eixo vertical e cada aluno é representado por uma linha que intersecta os eixos das unidades curriculares consoante o índice da respectiva UC. Um exemplo pode ser visto na figura 3.9.

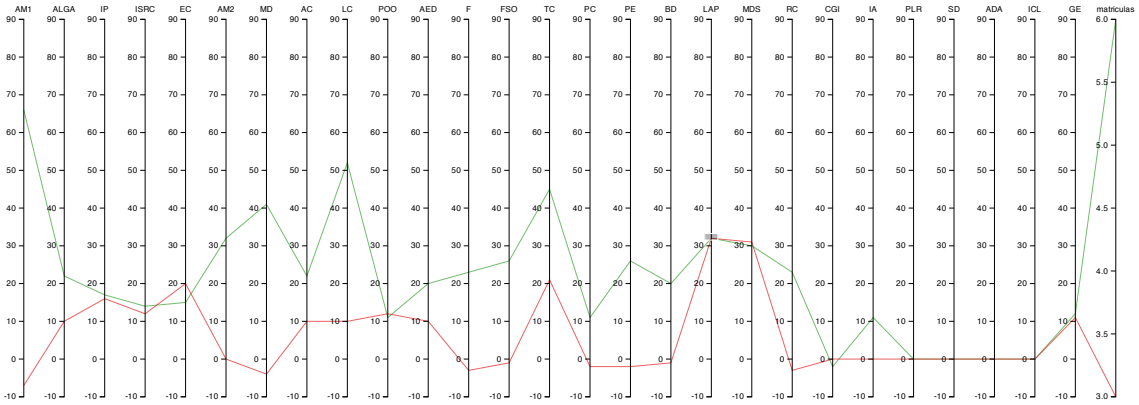


Figura 3.9: Exemplo de dois alunos representados na nova visualização para o percurso.

Nesta visualização, a ordem dos eixos é a mesma que a ordem de execução das unidades curriculares ditada pelo plano de estudos. Ou seja, os primeiros cinco eixos correspondem ao primeiro semestre, os segundos cinco eixos correspondem ao segundo e assim por diante. Picos superiores a zero num eixo indicam aprovações com um elevado número de inscrições e picos inferiores a zero indicam inscrições sem aprovação. Alunos que tenham zero num eixo indicam que ainda não se inscreveram nenhuma vez à unidade curricular e não existem valores entre 0 e 10. Na figura 3.9, o primeiro eixo corresponde a Análise Matemática I e o último ao número de matrículas do aluno. O aluno a verde aprovou com cerca de 15 valores na sexta inscrição a Análise Matemática I e tem seis matrículas no curso. O aluno a vermelho por sua vez, ainda não aprovou à unidade curricular (já vai com sete inscrições) e tem três matrículas no curso. É possível inferir o ano curricular do aluno olhando para as unidades curriculares que não têm inscrições. Ambos os alunos encontram-se no segundo ano curricular.

Valores inferiores a zero no eixo indicam que o aluno ainda não aprovou mas já se

inscreveu e esta visualização tem ainda um eixo extra que representa o número de matrículas de um aluno. Este atributo não é utilizado pelo algoritmo, mas é bastante útil para filtrar os resultados no gráfico. Na figura 3.9 ambos os alunos têm zero nos últimos eixos, o que indica que ainda não se inscreveram nessas unidades curriculares. Estes valores a zero são uma indicação do ano curricular em que o aluno se encontra, considerando que a maior parte dos alunos se inscreve ao máximo de unidades curriculares que tem disponível. Picos com valores

O modelo de índice negativo apresenta melhores resultados que o modelo de índice original (tanto visualmente como analiticamente) e por isso é o modelo escolhido para efectuar a análise ao percurso académico. Na secção seguinte irá ser descrita a aplicação dos algoritmos *k-means* e SOM a esta modelo, assim com a validação dos resultados obtidos.

3.3.2 Protocolo Experimental

O primeiro algoritmo escolhido para iniciar a análise ao percurso académico de um aluno foi o *k-means*. Esta escolha segue os mesmos princípios discutidos na análise ao desempenho: este algoritmo é simples de aplicar e permite uma iteração rápida entre analisar resultados, introduzir o feedback dessa análise no modelo e voltar a correr o algoritmo. A aplicação deste algoritmo, assim como os resultados obtidos, é apresentada na secção seguinte. A secção 3.3.2.2 apresenta os resultados obtidos da aplicação do SOM, o segundo algoritmo escolhido. Para finalizar, a secção 3.3.2.3 apresenta a validação dos resultados obtidos, comparando os agrupamentos de cada algoritmo.

3.3.2.1 Algoritmo *k-means*

Tal como foi dito na análise ao desempenho académico, o algoritmo *k-means* apresenta um problema: o parâmetro de entrada do algoritmo (o número de grupos) é precisamente o que se pretende encontrar. Para contornar esta questão foi realizado um estudo do melhor valor para k, baseado nas métricas definidas anteriormente e na análise dos gráficos criados para o modelo de índice negativo.

As métricas presentes na tabela 3.2 não oferecem resultados conclusivos na procura do melhor valor para k. Se forem consideradas a silhueta e o índice de Dunn, o valor de k mais indicado é 7. Se for considerado o índice de Davies-Bouldin, a coesão, o afastamento e a soma do erro quadrático então esse valor é 12. O método do cotovelo (figura 3.10) também não oferece um resultado conclusivo por si só, indicando que a maior descida de erro é para $k = 4$ e para $k = 6$. Uma análise visual a estes agrupamentos ajuda a decidir qual é que apresenta melhores resultados. Para um $k = 7$ é possível detectar grupos distintos e consegue-se definir uma semântica para os descrever, o que não acontece com os outros agrupamentos analisados.

A semântica criada para o agrupamento obtido com $k = 7$ distingue os grupos como grupos de baixo, médio ou alto rendimento. Grupos de baixo rendimento, como se pode

k	Coesão	Afastamento	Silhueta	DB Index	Dunn Index	Sum Squared Error
3	61.51629	60.866	-0.01213	313.40938	0.00456	325.03896
4	58.42604	58.29649	-0.0025	277.4982	0.00611	298.1406
5	57.75158	57.56813	-0.00378	215.39264	0.00617	287.11448
6	57.18899	56.97467	-0.00443	209.12377	0.00809	268.23317
7	56.74612	56.68766	-0.00096	179.41473	0.00874	263.24211
8	56.37125	56.14042	-0.00664	150.11418	0.00755	258.03165
9	55.95355	55.74166	-0.00605	133.73256	0.00701	254.25273
10	55.43415	55.10984	-0.00742	113.22632	0.00751	247.88345
11	54.94121	54.53716	-0.00909	100.44893	0.00723	242.16527
12	54.62805	54.27767	-0.00917	92.35291	0.00725	238.87834

Tabela 3.2: Métricas para os agrupamentos do algoritmo *k-means*.

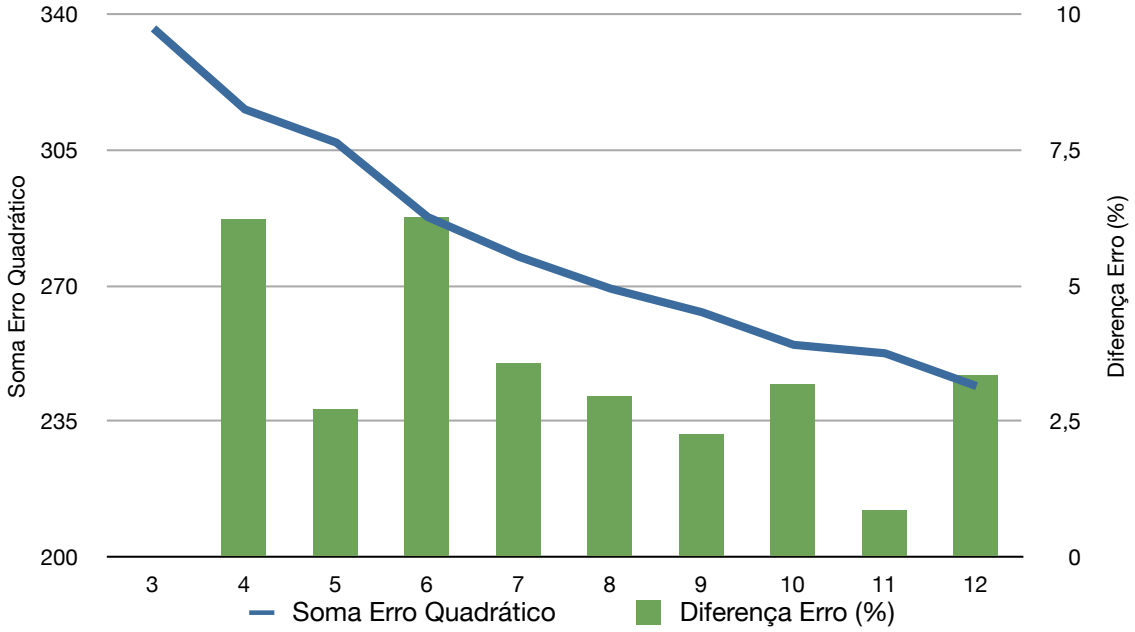


Figura 3.10: Gráfico da soma dos erros quadráticos de cada corrida do algoritmo para um k entre 3 e 12.

ver na figura 3.11, são caracterizados por um número elevado de unidades curriculares com índices negativos. São alunos que apesar de estarem no curso à mais de três anos, ainda precisam de pelo menos mais dois anos para o terminar. O grupo azul detecta um padrão interessante: são alunos que apesar de terem um rendimento maior do que o grupo vermelho, têm um grande número de um índices negativos nas unidades curriculares de segundo ano e a Análise Matemática II.

A figura 3.12, mostra os grupos designados de médio rendimento. São alunos que irão precisar de pelo menos mais um ano lectivo para terminar o curso e dividem-se em dois grupos. Ambos os grupos evidenciam um grande número de inscrições até obter aprovação nas unidades curriculares de primeiro e segundo ano e a grande maioria de alunos ainda não aprovou a unidades curriculares do terceiro ano. A diferença entre os dois grupos centra-se nas aprovações a unidades curriculares de primeiro e segundo ano:

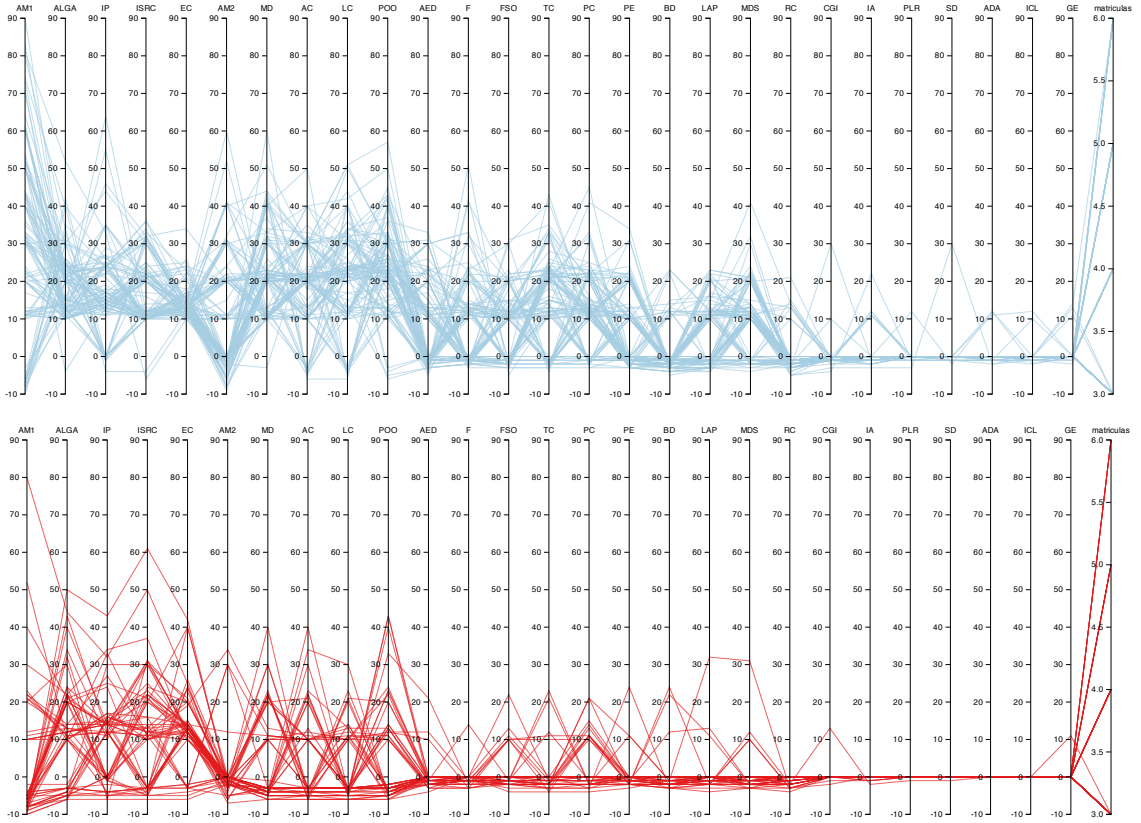


Figura 3.11: Grupos com percursos de baixo rendimento detectados pelo *k-means*.

os alunos do grupo azul aprovaram a quase todas as unidades curriculares do primeiro ano (excepto a Análise Matemática I e II), por oposição aos do grupo vermelho que têm um número elevado de inscrições sem aprovação. Outra diferença está nas aprovações a Análise Matemática I: os alunos do grupo vermelho têm um elevado número de inscrições sem aprovação face aos alunos do grupo azul cujo um grande grupo já aprovou.

Os grupos designados de alto rendimento, figura **??**, são caracterizados por alunos que já acabaram o curso ou estão prestes a acabar. Em média demoram cerca de duas inscrições a aprovar a uma unidade curricular e são alunos com um número muito reduzido de índices negativos.

A distinção entre os grupos dos vários tipos de rendimento é difícil de descrever neste documento, pois as diferenças são muitas e variadas, e são muito facilmente detectadas numa análise visual utilizando as ferramentas criadas para o efeito, que podem ser consultadas no site http://pessoa.fct.unl.pt/l.ramos. Em seguida são apresentados os resultados da aplicação do algoritmo SOM ao modelo de percurso negativo.

3.3.2.2 Algoritmo SOM

O número de grupos obtidos pelo algoritmo SOM, explicado com mais pormenor na secção 2.1.1, é ditado pela largura e altura da rede utilizada. O número de épocas de

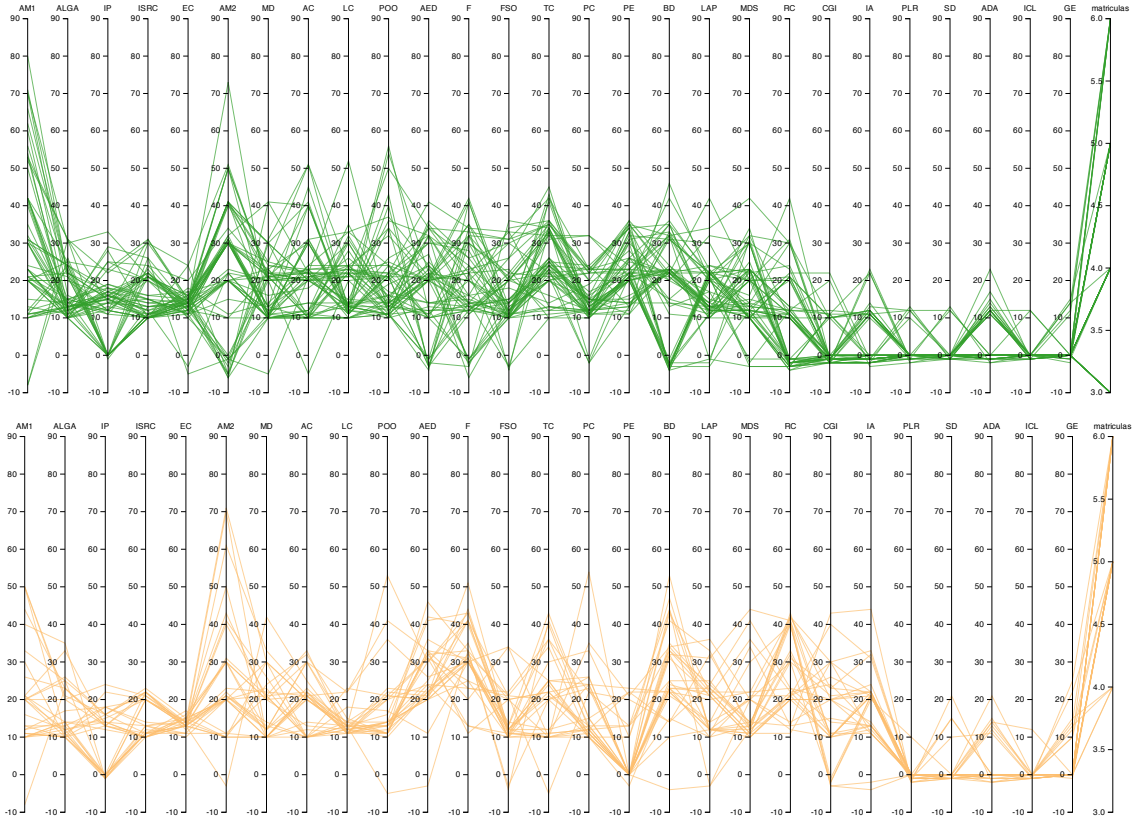


Figura 3.12: Grupos com percursos de médio rendimento detectados pelo *k-means*.

convergência e de ordenação não foi alterado e foram usados os valores por omissão do WEKA. Para descobrir o melhor agrupamento foi realizado um estudo aos valores de largura e altura da rede utilizada pelo algoritmo.

O intervalo escolhido para análise foi de uma largura e altura entre 1 e 4. Como o número de grupos detectados é igual ao número de nós na rede, o menor agrupamento relevante detectado será de dois e o máximo de 16. Uma quantidade maior de grupos provavelmente não irá revelar nova informação, e como o melhor agrupamento detectado com o algoritmo *k-means* é de 7 grupos, é pouco provável que o melhor agrupamento tenha mais do que o máximo possível neste intervalo. De notar que os agrupamentos cuja rede tem o mesmo número de nós mas têm diferentes valores de largura e altura, por exemplo, um agrupamento com largura 2 e altura 3 e outro agrupamento com largura 3 e altura dois, detectam grupos bastante semelhantes (a ponto de terem as mesmas métricas).

Como no algoritmo anterior, o melhor agrupamento foi escolhido conjugando a análise visual à análise das métricas obtidas. As métricas de qualidade, presentes na tabela 3.3, não revelam imediatamente um agrupamento óptimo. Cada métrica indica um agrupamento diferente, sendo que apenas a silhueta aponta para um agrupamento, de largura 2 e altura 4, cujo número de grupos não é um extremo. Uma análise visual a este agrupamento permite detectar grupos e criar uma descrição para cada um.

Seguindo a mesma lógica do algoritmo anterior, é possível criar uma semântica com

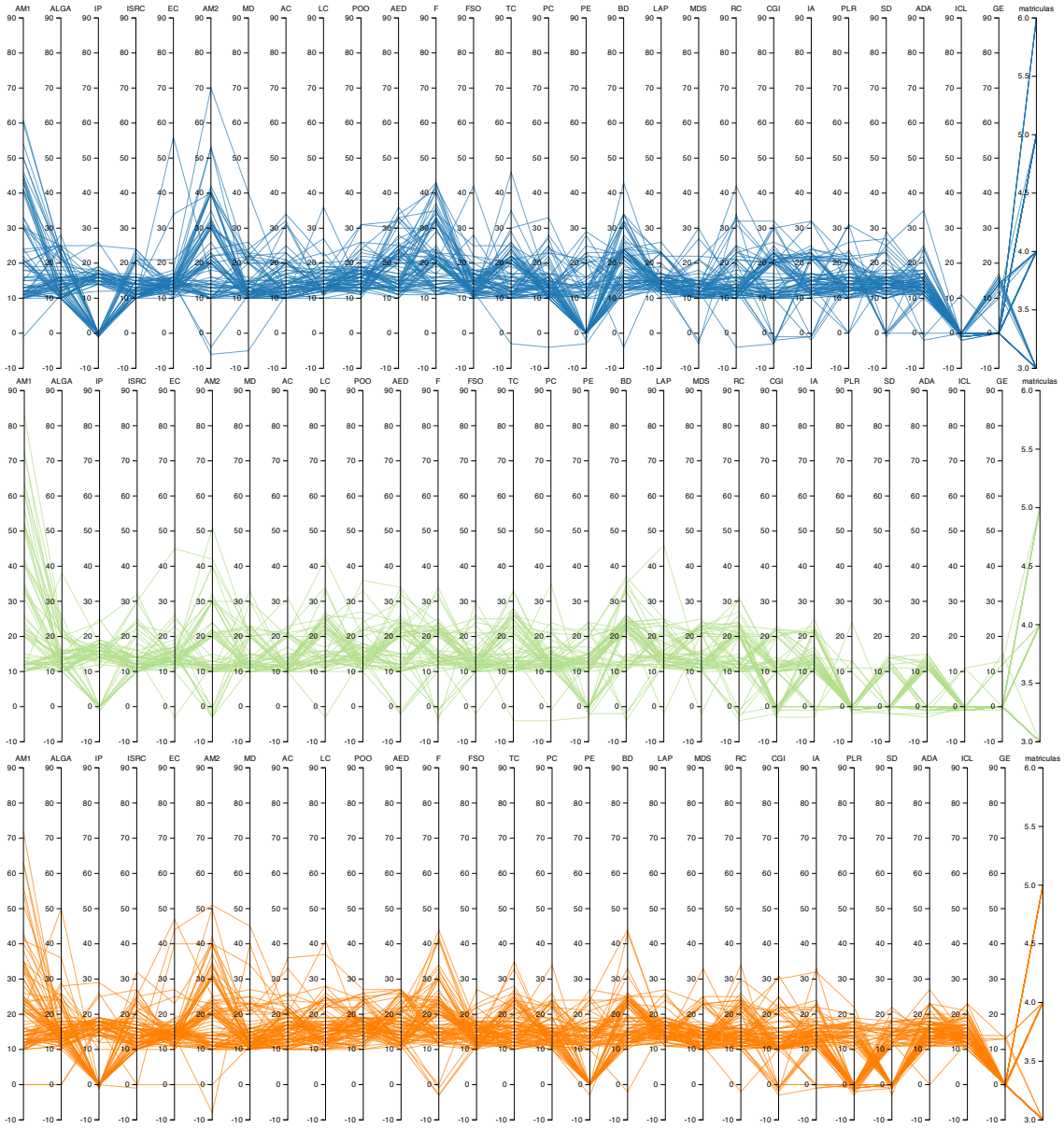


Figura 3.13: Grupos com percursos de alto rendimento detectados pelo *k-means*.

base no rendimento dos alunos que divide os grupos do agrupamento em baixo, médio e alto rendimento. Os grupos apresentados na figura 3.14, representam grupos de alunos de baixo rendimento. A figura 3.15 apresentam os grupos de médio rendimento, em que a maioria dos alunos tem índices negativos nas unidades curriculares do último semestre e precisarão de pelo menos mais um semestre para completar a licenciatura. Os grupos de alto rendimento, na figura 3.16, são de alunos que provavelmente já terminaram o curso com uma média de duas inscrições por aprovação.

A validação dos resultados apresentados, tanto deste algoritmo como do algoritmo *k-means*, irá ser discutida na secção seguinte.

Largura	Altura	Coesão	Afastamento	Silhueta	DB Index	Dunn Index
1	2	63.53242	63.42708	-0.00154	320.39745	0.00359
1	3	60.95311	60.47688	-0.00853	300.93598	0.00502
1	4	58.48273	58.28023	-0.00382	295.15256	0.00607
2	1	63.53242	63.42708	-0.00154	320.39745	0.00359
2	2	58.45585	58.23252	-0.00402	262.24258	0.00614
2	3	57.23524	57.06889	-0.00332	193.04945	0.00604
2	4	55.7827	55.56409	-0.00412	137.78975	0.00817
3	1	60.95311	60.47688	-0.00853	300.93598	0.00502
3	2	57.23524	57.06889	-0.00332	177.90746	0.00604
3	3	55.4003	55.14589	-0.00488	128.88586	0.00885
3	4	54.21878	53.57172	-0.01235	86.86354	0.01145
4	1	58.48273	58.28023	-0.00382	295.15256	0.00607
4	2	55.7827	55.56409	-0.00412	158.93885	0.00817
4	3	54.21878	53.57172	-0.01235	105.61602	0.01145
4	4	53.48499	52.77458	-0.01425	78.43515	0.0145

Tabela 3.3: Métricas para os agrupamentos do algoritmo SOM.

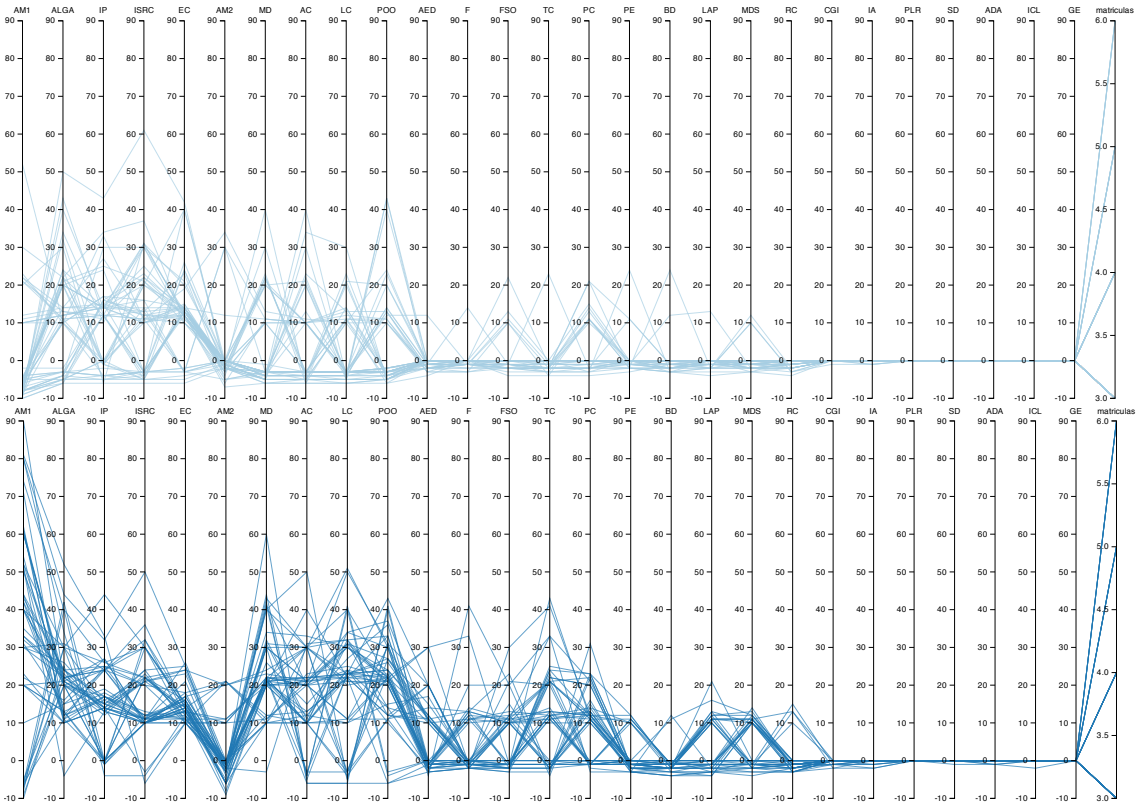


Figura 3.14: Grupos com percursos de baixo rendimento detectados pelo SOM.

3.3.2.3 Validação do Modelo

A validação dos resultados obtidos foi realizada comparando os melhores agrupamentos de cada algoritmo. A hipótese é que os padrões encontrados são estáveis se forem encontrados por ambos os algoritmos. A semelhança entre os grupos de algoritmos diferentes

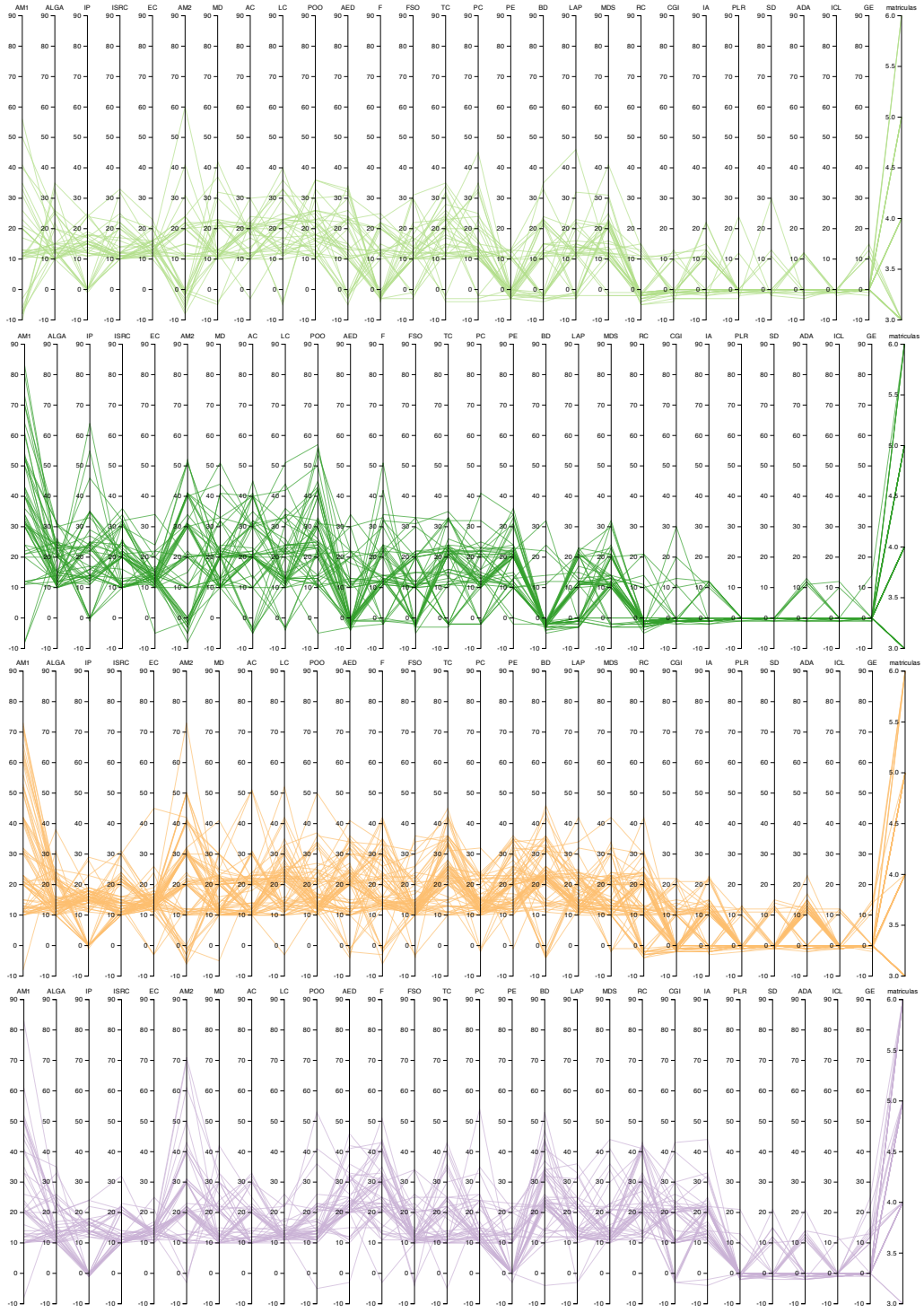


Figura 3.15: Grupos com percursos de médio rendimento detectados pelo SOM.

pode ser analisada verificando a distribuição dos alunos de um grupo de um algoritmo pelos grupos do outro algoritmo. Se se conseguir fazer uma mapeamento quase directo entre os grupos de ambos os algoritmos, então verifica-se a integridade dos padrões detectados.

Esta comparação pode ser encontrada na tabela 3.4, que mostra a distribuição dos

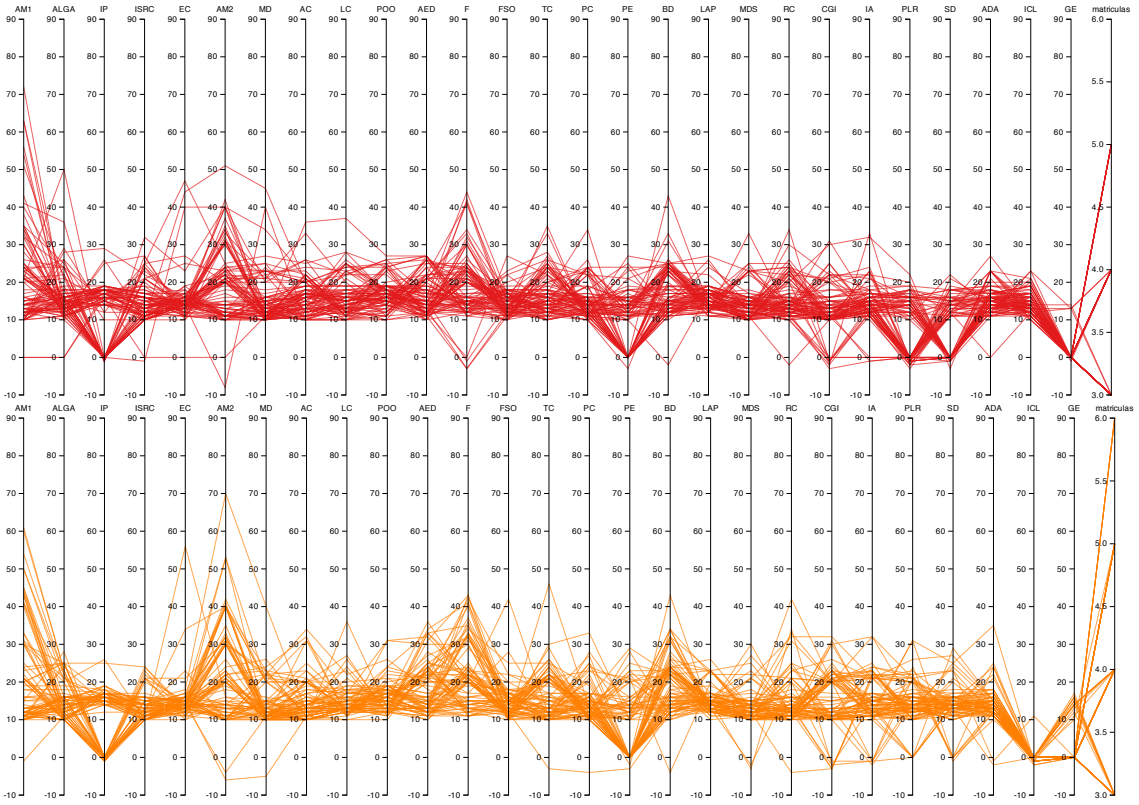


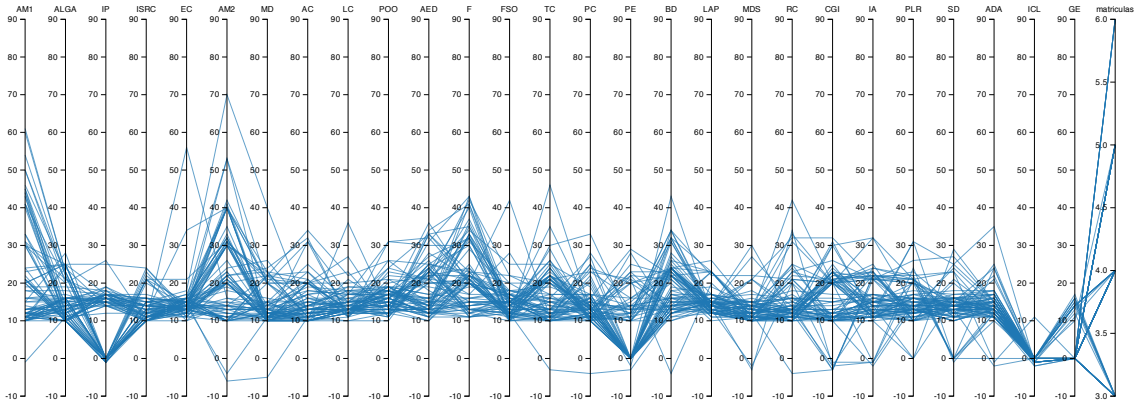
Figura 3.16: Grupos com percursos de alto rendimento detectados pelo SOM.

alunos de cada grupo do algoritmo *k-means* pelos grupos do algoritmo SOM. Como se pode verificar, a distribuição de alguns grupos é bastante sólida, quando a maioria dos alunos se encontra apenas num grupo (G1, G5 e G6 da coluna *k-means*), e para outros é mais variada (G2 e G3 da coluna *k-means*).

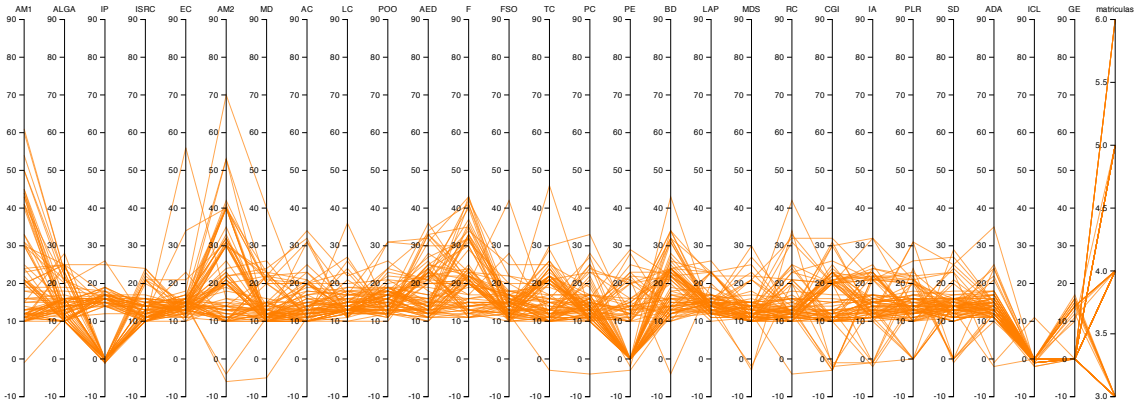
Os grupos designados como alto rendimento são detectados por ambos os algoritmos sem existir quase diferenças no número de alunos. A figura 3.17 mostra a comparação entre dois grupos de alto rendimento dos dois algoritmos, sendo muito difícil uma distinção visual. Os grupos de baixo rendimento também são detectados por ambos os algoritmos de forma bastante semelhante. A diferença entre os alunos detectados existe, mas não é relevante o suficiente para alterar a designação do grupo. A comparação entre

k-means / SOM	G0	G1	G2	G3	G4	G5	G6	G7
G0	.	48	24	36	.	.	.	.
G1	.	.	.	.	.	.	81	1
G2	.	.	9	3	.	28	3	19
G3	.	.	2	13	.	44	.	.
G4	55	3	2	.	.	.	.	.
G5	.	.	.	.	.	1	.	29
G6	.	.	.	.	94	1	.	.

Tabela 3.4: Distribuição dos alunos pelos grupos dos algoritmos *k-means* e SOM.



(a) *k-means*



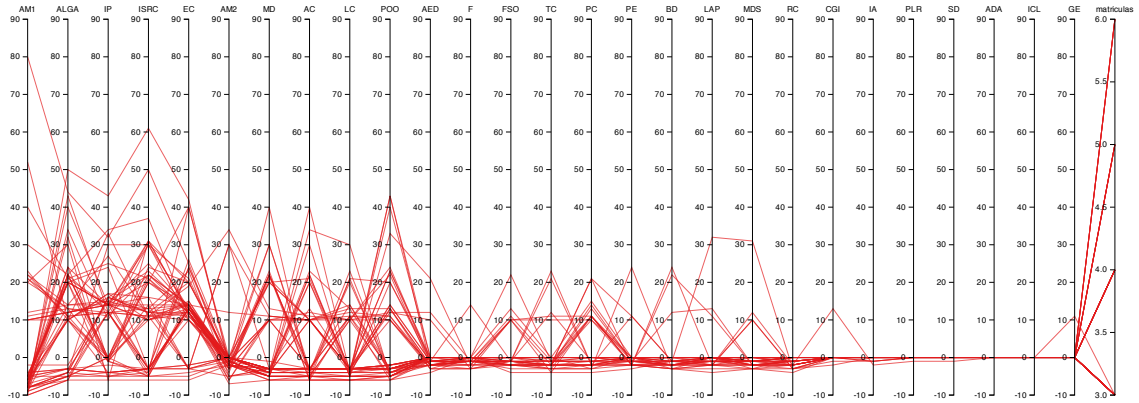
(b) SOM

Figura 3.17: Comparação grupos de alto rendimento.

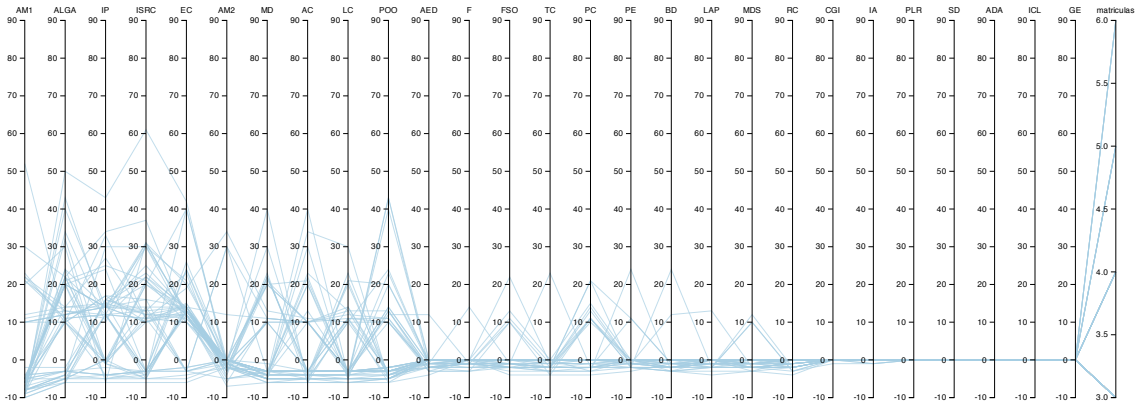
os grupos de baixo rendimento pode ser vista na figura 3.18. Apenas nos grupos de médio rendimento é que existe uma detecção menos linear de um algoritmo para o outro, como se pode ver na figura 3.19. O mapeamento entre os grupos não é directo e pode-se concluir que estes padrões não têm tanta estabilidade como os anteriores. Ainda assim, é possível observar os mesmo picos nas unidades curriculares de Análise Matemática I e II e ainda (se bem que menos pronunciado) um número de inscrições sem aprovação elevado a Base de Dados.

Para além da comparação entre agrupamentos diferentes, é possível validar o melhor agrupamento estudando a hierarquia formada à medida que se aumenta o número de grupos procurados pelo algoritmo. A figura 3.20 apresenta a hierarquia para o algoritmo *k-means*, em que cada eixo representa uma corrida do algoritmo para o valor do parâmetro indicado. Os valores nos eixos são os grupos detectados nessa corrida e cada aluno é representado por uma linha independente que intersecta um eixo no grupo a que pertence nessa corrida.

A partir desta figura é possível ver como se distribuem os alunos pelos grupos e a relação entre os grupos à medida que o valor de k aumenta. É possível inferir uma indicação para o melhor agrupamento quando as separações de grupos deixar de ser significativas,



(a) *k-means*



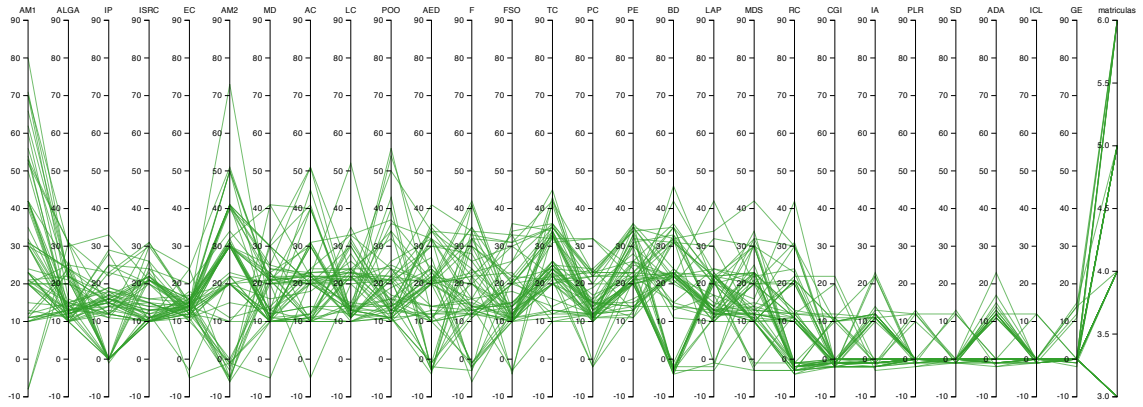
(b) SOM

Figura 3.18: Comparação grupos de baixo rendimento.

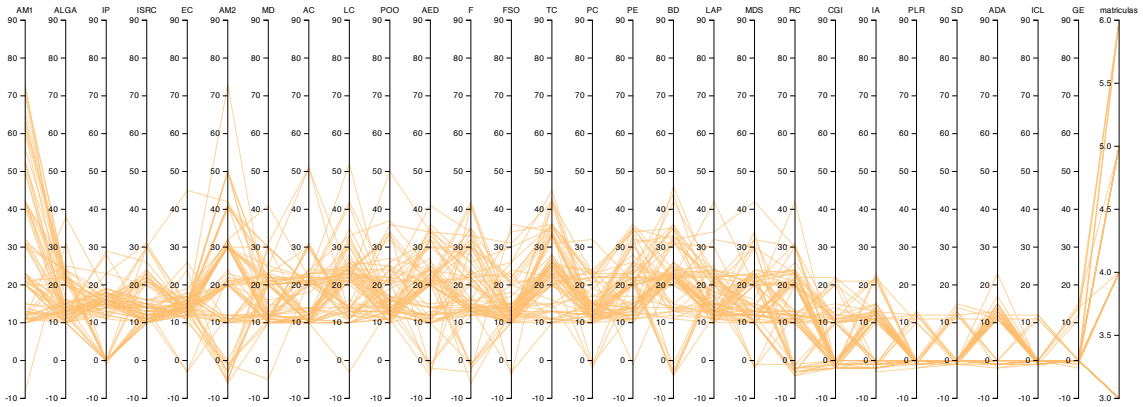
ou seja, quando um grande número de alunos forma um grupo inteiramente novo.

No caso do *k-means*, os grupos formados só se separam com algum significado quando se aumenta o valor de k de 6 para 7 e de 9 para 10. A escolha do $k = 7$ como melhor agrupamento parece ser a mais acertada à luz destes resultados. Note-se que os grupos 0, 3 e 4 (a contar do primeiro eixo) tem uma grande estabilidade ao longo das várias iterações, quase nunca sofrendo alterações. O grupo 1 separa-se em dois grupos logo na passagem para $k = 6$ e estes mantêm-se estáveis até $k = 7$. O grupo 2 mantêm-se estável na passagem para $k = 6$ e só depois se separa em dois, formando um grupo que recebe contribuições de muitos outros em $k = 7$.

No caso do SOM, a separação dos grupos a cada iteração não é tão linear. A figura 3.21, mostra a hierarquia para este algoritmo e pode-se observar que a separação têm sempre algum significado até chegar aos oito grupos. O grupo 3 (no último eixo) é o grupo que parece menos estável, pois é formado de vários contribuições de outros grupos.



(a) *k-means*



(b) SOM

Figura 3.19: Comparação grupos de médio rendimento.

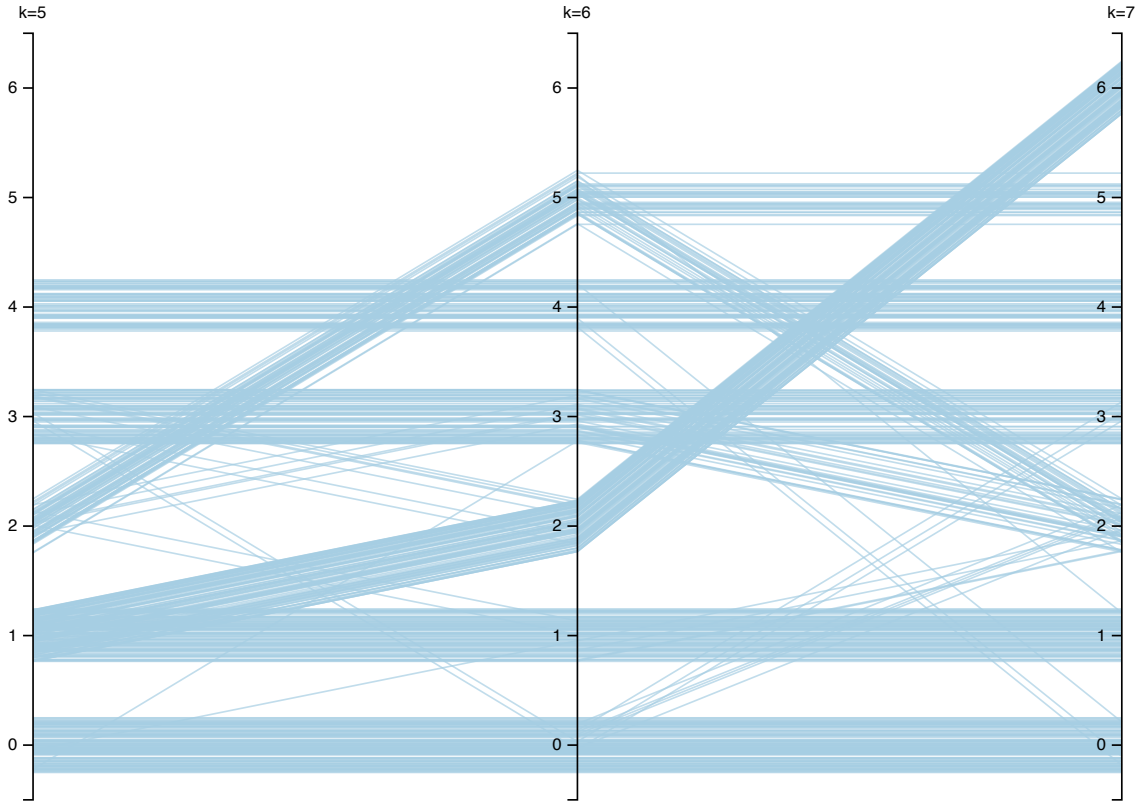


Figura 3.20: Hierarquia dos grupos do algoritmo *k-means* para k entre 5 e 7.

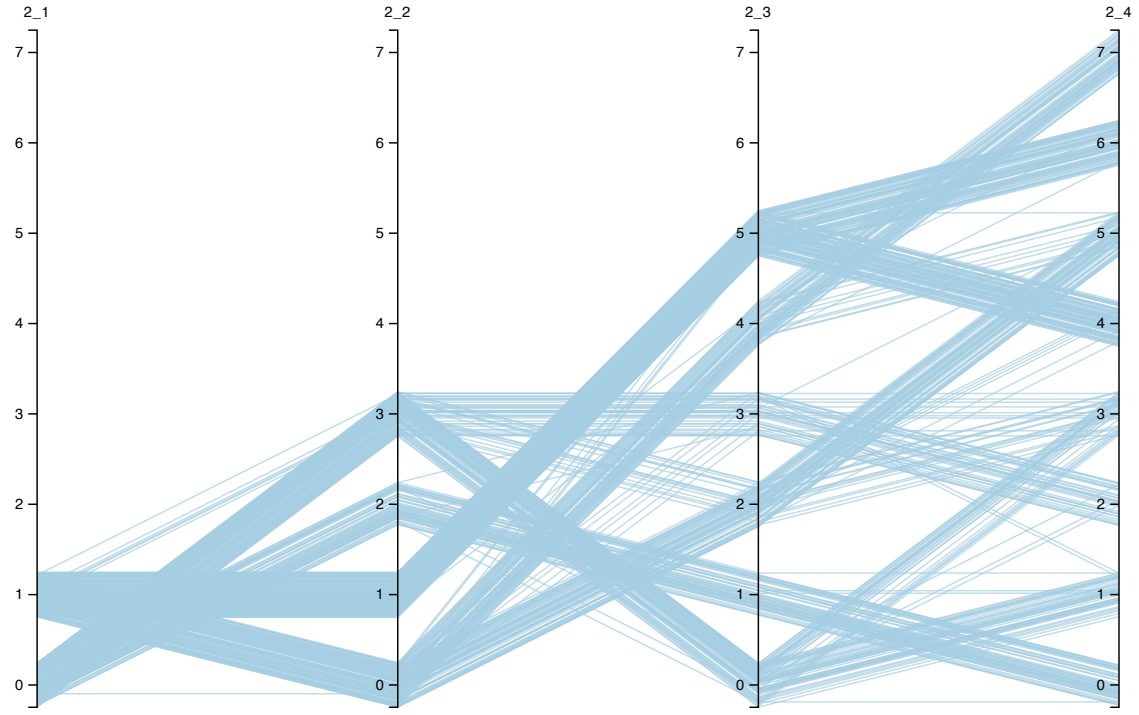


Figura 3.21: Hierarquia dos grupos do algoritmo SOM para k entre 5 e 10.

4

Conclusões

Com este trabalho pretende-se dar uma visão mais específica da aplicação de técnicas de data-mining na procura e detecção de padrões de desempenho académico. Esta área é relativamente inexplorada, estando disponível pouca informação sobre o assunto.

A secção seguinte deste último capítulo apresenta uma análise aos resultados obtidos, descrevendo algumas conclusões resultantes da aplicação de diferentes algoritmos aos modelos. Na secção 4.2 é feita uma discussão do processo tomado, sublinhando alguns pontos importantes e apresentando algumas alternativas em alguns dos passos do protocolo. A última secção deste capítulo, secção 4.3, discute a continuação deste trabalho, principalmente no esforço de melhorar o desempenho académico do ensino superior.

4.1 Análise de Resultados

No geral, os resultados obtidos pelas duas análises descritas no capítulo anterior foram positivos. Foram desenvolvidos modelos para ambas as abordagens que retornaram agrupamentos estáveis e foi possível criar semânticas que descrevem esses grupos.

Um desses modelos, o modelo de desempenho académico, permite uma análise à população activa de um ano lectivo específico, permitindo perceber que comportamos a nível de performace os alunos têm. Estes comportamentos foram posteriormente confirmados comparando-os com os grupos detectados na população de cada ano lectivo disponível. Os padrões só não se verificam nos primeiros dois anos lectivos (2006/07 e 2007/08) visto a população não ser muito elevada e não estar bem distribuída pelos anos curriculares.

Os resultados da análise ao percurso académico de um aluno também foram bastante positivos. Apesar dos bons resultados obtidos com o modelo de percurso, este não

foi explorado na sua totalidade. A lista de unidades curriculares obrigatórias do plano curricular é apenas uma das várias combinações possíveis e uma alteração nesta lista permite realizar uma análise com resultados diferentes dos obtidos. De qualquer modo, a análise realizada em conjunção com a visualização desenvolvida permitiram a detecção de padrões e a criação de uma semântica que descreve esses padrões.

A partir do gráfico de paralelas na figura 4.1, facilmente se distinguem os padrões descritos pela mesma semântica (que define grupos com baixo, médio e alto rendimento). De uma análise ao gráfico é possível dizer que dos dois grupos de baixo rendimento, um

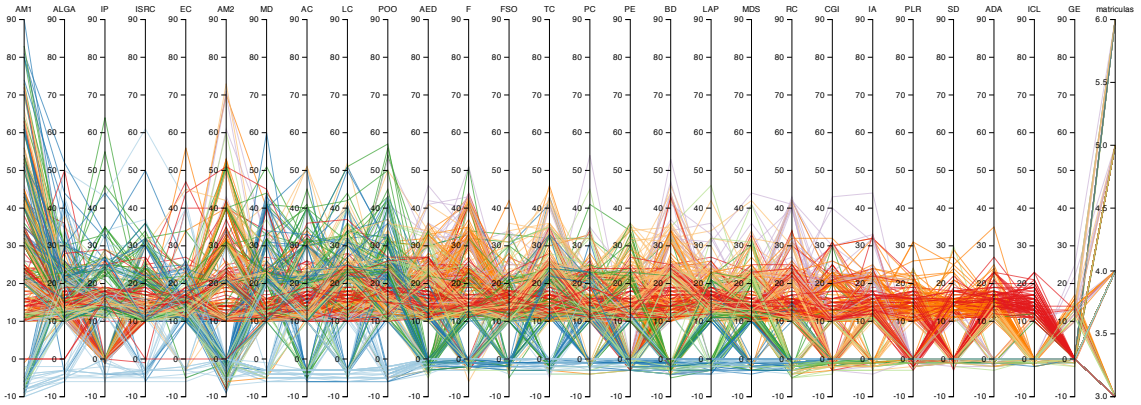


Figura 4.1: Gráfico de paralelas para o algoritmo SOM com oito grupos sobrepostos.

grupo tem alunos que ainda não obteve a aprovação nas unidades curriculares de segundo ano e o outro grupo tem alunos que ainda não aprovaram em unidades curriculares de terceiro ano. Também é possível dizer que, de entre os grupos de alto rendimento, distinguem-se um do outro pela aprovação (ou não) nas unidades curriculares de Introdução à Programação e Probabilidade e Estatística. Esta distinção é feita com base nos picos do gráfico, que indicam um grande número de inscrições. Uma conclusão que facilmente se retira da leitura do gráfico é que as unidades curriculares com mais inscrições por aluno são Análise Matemática I e II (em que mais de um terço dos alunos demora mais de 4 anos a terminar AMI).

4.2 Discussão do Processo

O protocolo seguido descrito na secção 3.1 foi um processo sobretudo de descoberta. Apesar de se saber qual a lógica a seguir para uma análise deste tipo, o que fazer em pormenor a cada passo do processo não era uma incógnita. Esta secção irá discutir alguns problemas encontrados ao longo da análise, quais as tarefas que consumiram mais tempo e o que pode ser feito para melhorar uma análise deste tipo.

Um dos maiores entraves considerados neste trabalho foi a pouca informação prática disponível sobre o assunto. Com um número muito reduzido de trabalhos na área para comparar resultados, é difícil guiar a análise e perceber se as decisões tomadas irão ter um impacto positivo ou não. Visto que a análise dos resultados é bastante subjectiva,

ou seja, depende muito do conhecimento de domínio e do que se procura, a existência de outros trabalhos é benéfica. Não só se ganha pelas possíveis perspectivas diferentes que estes trabalhos oferecem mas também porque ajuda na leitura dos resultados, que pode ser influenciada quando quem conduz a análise quer encontrar padrões onde não existem. Sem termos de comparação, fica difícil de distinguir um mau resultado de um bom resultado e foi criada uma métrica de qualidade para fazer esta distinção.

O que consumiu mais tempo neste trabalho foi conseguir o domínio das várias ferramentas utilizadas na criação do workflow de análise. Para permitir uma rápida iteração entre gerar um modelo e introduzir o feedback da análise dos resultados num novo modelo, foi preciso utilizar (ou criar) uma ferramenta específica a cada passo. O conhecimento prévio do que está disponível para efectuar a análise e quais as ferramentas que vão ser aplicadas reduzem significativamente o tempo dispensado em aprendizagem.

A partir do momento em que existe um workflow, os problemas encontrados passam a estar relacionados com a análise em si. A escolha do grupo de algoritmos é outro ponto que pode atrasar bastante a análise. Neste trabalho os algoritmos utilizados estavam limitados pela oferta do WEKA e também pelo tempo de implementação de novos algoritmos. Possivelmente, a utilização de outros algoritmos para além daqueles que foram utilizados poderá resultar em agrupamentos completamente diferentes.

A experimentação com outros algoritmos faz parte do trabalho futuro, que é discutido na próxima secção.

4.3 Trabalho Futuro

O trabalho realizado nesta dissertação deixa ainda muito por explorar. Há espaço para continuar a análise realizada, experimentando com diferentes algoritmos que não foram utilizados. Sendo data mining uma área em constante expansão, a aplicação de novos algoritmos pode trazer novos padrões e conclusões à análise.

A aplicação dos modelos e algoritmos a dados de outra instituição é outro caminho a explorar. Como já foi dito, não existem publicações académicas que registem análises semelhantes à que foi feita e seria bastante interessante ver como se comportam os modelos criados com dados de outras instituições.

O trabalho realizado também pode ser utilizado numa tentativa de melhorar o desempenho académico da instituição, manifestamente no desenvolvimento de medidas específicas para os grupos encontrados. Uma análise do impacto que este tipo de medidas pode ter para a instituição de ensino superior (em termos de gestão de recursos e não só) pode ser bastante interessante. Para além do desenvolvimento de medidas especificas, também poderá ser feita uma análise de data mining que utilize técnicas de classificação em vez de agrupamento. O objectivo seria tentar determinar situações de desistência ou de degradação do desempenho escolar enquanto estas ocorrem, tendo como base os resultados obtidos nesta dissertação. Uma análise deste tipo que possibilite a intervenção junto dos alunos nestas situações é uma mais valia para qualquer instituição académica.

Por último, pode ser pensada uma ferramenta que permita realizar análises no domínio desta dissertação, sem ser necessário conhecimento extensivo de data mining. A ferramenta teria de ser extensível para permitir a adição de novos modelos, algoritmos ou visualizações. O propósito seria possibilitar a um professor ou investigador, sem conhecimentos de técnicas de data mining, a possibilidade de estudar uma instituição académica utilizando essas técnicas. Uma ferramenta deste tipo disseminaria de forma facilitada o uso deste tipo de métodos no estudo e melhoria do desempenho académico a nível nacional.

Bibliografia

[BY09] R. Baker e K. Yacef. "The state of educational data mining in 2009: A review and future visions". Em: *Journal of Educational Data Mining* 1.1 (2009), pp. 3–17.

[BP11] B. Baradwaj e S. Pal. "Mining Educational Data to Analyze Students" Performance". Em: *International Journal* 2 (2011).

[BAS97] M. Besterfield-Sacre, C. Atman e L. Shuman. "Characteristics of Freshman Engineering Students: Models for Determing Student Attrition in Engineering". Em: *Journal of Engineering Education - Washington* - 86 (1997), pp. 139–150.

[Cio+00] K. J. Cios, A. Teresinska, S. Konieczna, J. Potocka e S. Sharma. "Diagnosing Myocardial Perfusion SPECT Bull's-eye Maps-A Knowledge Discovery Approach". Em: *IEEE Engineering in Medicine and Biology Magazine* 19.4 (2000), pp. 17–25.

[DB79] D. Davies e D. Bouldin. "A cluster separation measure". Em: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2 (1979), pp. 224–227.

[12a] *DGES - Ensino Superior*. 2012. URL: http://www.dges.mctes.pt/DGES/pt/Reconhecimento/NARICENIC/Ensino%20Superior/Sistema%20de%20Ensino%20Superior%20Portugu%C3%83%C2%AAs.

[12b] *DGES - Processo de Bolonha*. 2012. URL: http://www.dges.mctes.pt/DGES/pt/Estudantes/Processo+de+Bolonha/Processo+de+Bolonha/.

[12c] *DI História*. 2012. URL: http://www.di.fct.unl.pt/sobre/historia-do-departamento.

[Dun73] J. Dunn. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters". Em: (1973).

[Est+96] M. Ester, H.-p. Kriegel, J. S e X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise". Em: AAAI Press, 1996, pp. 226–231.

[FIO05] B. French, J. Immekus e W. Oakes. "An examination of indicators of engineering students' success and persistence". Em: *Journal of Engineering Education - Washington* - 94.4 (2005), p. 419.

[ElH09] A. El-Halees. "Mining Students Data to Analyze Learning Behavior: A Case Study". Em: *Department of Computer Science, Islamic University of Gaza PO Box* 108 (2009).

[Hal+09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann e I. Witten. "The WEKA data mining software: an update". Em: *ACM SIGKDD Explorations Newsletter* 11.1 (2009), pp. 10–18.

[HKP11] J. Han, M. Kamber e J. Pei. *Data mining: concepts and techniques*. Morgan Kaufmann Pub, 2011.

[HW79] J. A. Hartigan e M. A. Wong. "Algorithm AS 136: A k-means clustering algorithm". Em: *Applied statistics* (1979), pp. 100–108.

[Jai10] A. Jain. "Data clustering: 50 years beyond K-means". Em: *Pattern Recognition Letters* 31.8 (2010), pp. 651–666.

[KS96] D. J. Ketchen e C. L. Shook. "The application of cluster analysis in strategic management research: an analysis and critique". Em: *Strategic management journal* 17.6 (1996), pp. 441–458.

[Koh90] T. Kohonen. "The self-organizing map". Em: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480.

[KM06] L. A. Kurgan e P. Musilek. "A survey of Knowledge Discovery and Data Mining process models". Em: *Knowledge Engineering Review* 21.1 (2006), pp. 1–24.

[Lua02] J. Luan. "Data mining and its applications in higher education". Em: *New directions for institutional research* 2002.113 (2002), pp. 17–36.

[Min+03] B. Minaei-Bidgoli, D. Kashy, G. Kortmeyer e W. Punch. "Predicting student performance: an application of data mining methods with an educational web-based system". Em: *Frontiers in Education, 2003. FIE 2003. 33rd Annual*. Vol. 1. IEEE. 2003, T2A–13.

[MSC05] A. Moreira, M. Y. Santos e S. Carneiro. "Density-based clustering algorithms–DBSCAN and SNN". Em: *Available at: get. dsi. uminho. pt/local/download/SNN&DBSCAN. pdf* (2005).

[Ogo07] E. Ogor. "Student academic performance monitoring and evaluation using data mining techniques". Em: *Electronics, Robotics and Automotive Mechanics Conference, 2007. CERMA 2007*. IEEE. 2007, pp. 354–359.

[12d] *PORDATA*. 2012. URL: http://www.pordata.pt/.

[Rep05] D. da República. "Alteração à Lei de Bases do Financiamento do Ensino Superior". Em: *Diário da República* (ago. de 2005).

[Rep06] D. da República. "Fórmula de Cálculo do Orçamento para Financiamento do Ensino Superior". Em: *Diário da República* (jan. de 2006).

[RV07] C. Romero e S. Ventura. "Educational data mining: A survey from 1995 to 2005". Em: *Expert Systems with Applications* 33.1 (2007), pp. 135–146.

[Rom+08] C. Romero, S. Ventura, P. Espejo e C. Hervás. "Data mining algorithms to classify students". Em: *Proceedings of Educational Data Mining* (2008), pp. 20–21.

[Rou87] P. Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". Em: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[SN01] J. Smith e R. Naylor. "Determinants of degree performance in UK universities: a statistical analysis of the 1993 student cohort". Em: *Oxford Bulletin of Economics and Statistics* 63.1 (2001), pp. 29–60.

[TGB98] S. Takahira, D. Goodings e J. Byrnes. "Retention and performance of male and female engineering students: An examination of academic and environmental variables". Em: *women* 11 (1998), p. 14.

[VMS07] J. Vandamme, N. Meskens e J. Superby. "Predicting academic performance by data mining methods". Em: *Education Economics* 15.4 (2007), pp. 405–419.

[WFH11] I. Witten, E. Frank e M. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

Apêndice: Glossário

Data Mining É o processo computacional de descoberta de padrões em dados de grande volume, fazendo uso de técnicas vindas da inteligência artificial, aprendizagem automática, estatística e sistemas de base de dados. Também é um dos passos do processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* ou KDD).

Técnicas de Agrupamento São algoritmos que agrupam objectos de forma não supervisionada, de tal maneira que objectos do mesmo grupo são semelhantes entre si e dissemelhantes de objectos noutros grupos.

Grupo/Padrão/Cluster São conjuntos de objecto considerados semelhantes encontrados autonomamente por um algoritmo de agrupamento.

Desempenho Académico Os resultados conseguidos por um aluno num grupo de unidades curriculares, normalmente traduzido em média de notas ou média de créditos obtidos.

Plano Curricular Conjunto de restrições que um aluno tem de cumprir para concluir o curso a que o plano curricular se refere. Essas restrições podem vir na designação de que unidades curriculares são obrigatórias e quantos créditos é necessário ter em que áreas.

Plano de Estudos Conjunto de unidades curriculares divididas em anos lectivos e semestres, que estão disponíveis para um aluno assistir. A obrigatoriadade de aprovação a uma unidade curricular é definida no plano curricular.

ECTS Medida europeia de esforço de um aluno. Em Portugal 1 crédito ECTS equivale a 28 horas de trabalho dedicado à unidade curricular, seja estudo, aulas, exames, testes, trabalhos práticos, visitas de estudo, etc.

Unidade Curricular (UC) É uma unidade de ensino com objectivos de formação próprios a que os alunos se podem inscrever e são avaliados pelo conhecimento adquirido (traduzido numa classificação final).