Association for Information Systems
AIS Electronic Library (AISeL)

2018 Proceedings | Portugal (CAPSI)

2018

Factors Influencing School Success in Undergraduate Programs at a Portuguese Higher Education Institution

Cláudia Pimenta
Instituto Politécnico de Coimbra, iscac15464@alumni.iscac.pt

Renato Ribeiro
Instituto Politécnico de Coimbra, iscac15064@alumni.iscac.pt

Vera Sá
Instituto Politécnico de Coimbra, iscac15093@alumni.iscac.pt

Fernando Paulo Belfo
Instituto Politécnico de Coimbra, pbelfo@iscac.pt

Fatores que Influenciam o Sucesso Escolar das Licenciaturas numa Instituição de Ensino Superior Portuguesa

Factors Influencing School Success in Undergraduate Programs at a Portuguese Higher Education Institution

Cláudia Pimenta, Instituto Politécnico de Coimbra*, iscac15464@alumni.iscac.pt

Renato Ribeiro, Instituto Politécnico de Coimbra*, iscac15064@alumni.iscac.pt

Vera Sá, Instituto Politécnico de Coimbra*, iscac15093@alumni.iscac.pt

Fernando Paulo Belfo, Instituto Politécnico de Coimbra*, pbelfo@iscac.pt

* Instituto Superior de Contabilidade e Administração de Coimbra, Portugal

Resumo

Este artigo apresenta um projeto de descoberta de conhecimento em bases de dados, o qual pretendeu ajudar a entender a influência de alguns fatores no sucesso escolar dos alunos das licenciaturas de uma instituição do ensino superior portuguesa. As etapas contemplaram tarefas como entendimento do tema, compreensão dos dados, preparação dos dados, modelação e avaliação dos resultados. A investigação contemplou atividades preditivas e descritivas, utilizando técnicas de indução de árvores de decisão e de agrupamento. Um inquérito, feito a 86 alunos, contemplou fatores como assiduidade nas aulas, tempo de estudo, frequência de estudo em grupo, distância à localidade da residência e sucesso escolar. As árvores de decisão induzidas revelaram que o fator mais explicativo do sucesso escolar é o tempo de estudo e classificaram esse sucesso em função dos diversos fatores explicativos. Os resultados de agrupamento permitiram caraterizar o aluno típico e os dois ou três principais grupos de alunos.

Palavras-Chave: Sucesso escolar, descoberta de conhecimento em bases de dados, mineração dos dados, árvore de decisão, agrupamento.

Abstract

This article presents a project of knowledge discovery in databases, which intended to help understand the influence of some factors on the academic success of undergraduate students of a Portuguese higher education institution. The steps in this process included tasks such as business understanding, data understanding, preparation of data, modelling and evaluation of results. The research contemplated predictive and descriptive activities, using the techniques of induction of decision trees and grouping. A survey, made to 86 students, included factors such as attendance in class, time of study, group study frequency, distance to the place of residence and school success. The induced induction trees revealed that the most explanatory factor of school success is the time of study and classified this success in function of the various explanatory factors. The clustering results allowed characterizing the typical student and the two or three main groups of students.

Keywords: *School success, knowledge discovery in databases, data mining, decision tree, clustering.*

1 Introdução

O atual quadro de uma economia cada vez mais alicerçada no conhecimento e na inovação implica uma importância crescente das qualificações de grau superior. E, se é clara a importância dessas qualificações, também é clara a importância de que essa formação se faça o mais eficientemente possível. Um dos objetivos fundamentais das políticas públicas de educação no ensino superior, quer nas instituições universitárias, quer nas politécnicas, deverá ser a promoção do sucesso escolar dos seus alunos. O insucesso e o abandono escolar são considerados problemas preocupantes para muitos estudantes, e, consequentemente, também o devem ser para as suas instituições de ensino, para o sistema de ensino superior e para toda a sociedade portuguesa. Apesar de Portugal ter registado uma forte queda na taxa de abandono escolar entre 2008 e 2016 (de 34,9% para 14%), ainda representou em 2016 a quarta taxa de abandono escolar mais elevada da União Europeia, com grande parte significativa dos seus jovens entre os 18 e 24 anos a deixarem prematuramente a educação e a formação (Eurostat, 2017).

Este artigo apresenta um trabalho que investigou a influência de diversos fatores no sucesso escolar de alunos de licenciaturas de uma instituição do ensino superior portuguesa. Foi desenvolvido um inquérito, respondido por uma amostra de alunos dessa instituição, o qual permitiu construir uma bases de dados que serviu de base a um projeto de descoberta de conhecimento nesses mesmos dados com base na metodologia CRISP-DM, um acrónimo para a expressão inglesa de "Cross-Industry Standard Process for Data Mining". A secção seguinte apresenta os princípios desta metodologia e a secção 3 apresenta a forma como a metodologia foi seguida em concreto no projeto desenvolvido. A última secção apresenta as conclusões e trabalhos futuros.

2 Descoberta de conhecimento em bases de dados

A mineração de dados (*data mining*), é, de acordo com o grupo Gartner, o processo de descobrir novas correlações, padrões e tendências significativas analisando grandes quantidades de dados armazenados em repositórios, usando tecnologias de reconhecimento de padrões, bem como técnicas estatísticas e matemáticas (Gartner, 2018; Larose, 2005). Na realidade, trata-se de um passo num processo mais abrangente que é a descoberta de conhecimento em bases de dados (DCBD). O processo de descoberta de conhecimento tem sido cada vez mais utilizado nas mais diversas atividades económicas, como na área financeira (Ngai, Hu, Wong, Chen, & Sun, 2011), na área seguradora (Azadmanesh & Tarokh, 2012), na área do alojamento, restauração e similares (Loureiro, Lourenço, Costa, & Belfo, 2014), na área médica (Cios & Moore, 2002) e em muitas outras áreas, contribuindo com novo conhecimento e ajudando as suas organizações a definir estratégias que lhe permitam aumentar o seu desempenho. Um projeto de DCBD é normalmente

desencadeado por desafios ou oportunidades do lado do negócio e, para que seja verdadeiramente eficaz, o negócio deve adequadamente suportar e ser suportado por o lado das tecnologias de informação, o princípio fundamental do alinhamento do negócio com as tecnologias de informação (Belfo & Sousa, 2013; Reich & Benbasat, 1996). Na realidade, os projetos de DCBD são bons exemplos de iniciativas que devem ser alicerçadas num bom alinhamento do negócio com as tecnologias de informação, uma das maiores preocupações dos gestores de tecnologias de informação das principais empresas nos últimos anos (Kappelman, Nguyen, McLean, Maurer, Johnson, Snyder, & Torres, 2017).

Embora existam várias metodologias para implementar projetos de descoberta de conhecimento em bases de dados, a metodologia que tem ganho mais popularidade em iniciativas de DCBD tem sido a metodologia CRISP-DM ("Cross-Industry Standard Process for Data Mining") (Jackson, 2002). A metodologia CRISP-DM disponibiliza o ciclo de vida dum projeto de mineração de dados. Contém as fases dum projeto, as suas respetivas tarefas e as relações entre elas. Em teoria, é possível estabelecerem-se relações entre quaisquer tarefas de mineração de dados, dependendo dos objetivos, dos pressupostos, do interesse do utilizador e obviamente dos dados. Segundo a metodologia CRISP-DM, o ciclo de vida de um projeto de DCBD consiste em 6 fases, respetivamente: entendimento do negócio, compreensão dos dados, preparação dos dados, modelação (que corresponde à mineração dos dados ou "data mining"), avaliação e desenvolvimento. A sequência entre fases não é rígida, permitindo "andar para a frente" e "para trás" entre fases, dependendo tal percurso dos resultados de cada fase e de qual a fase ou tarefa seguinte (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Wirth, 2000).

No que diz respeito às tarefas da mineração dos dados, estas podem-se dividir em atividades preditivas, também designadas por supervisionadas, e em atividades descritivas, também designadas por não supervisionadas. As atividades preditivas aprendem critérios de decisão para se ser capaz de classificar casos desconhecidos (tendências futuras) a partir do conhecimento adquirido de um conjunto de amostras com classes conhecidas. As atividades descritivas trabalham com um conjunto de dados que não possuem uma classe de saída determinada procurando identificar padrões desconhecidos comuns existentes nestes dados. De entre as atividades preditivas destacam-se as tarefas de classificação, de previsão ou de análise de tendências. De entre as atividades descritivas, destacam-se as tarefas de agrupamento, associação ou de sumarização (Berry & Linoff, 1997; Gama, Carvalho, Faceli, Lorena, & Oliveira, 2017). O estudo aqui apresentado utilizou as tarefas de classificação e de agrupamento.

A classificação é uma das tarefas mais populares da mineração de dados. Consiste na predição de uma variável categórica, ou seja, descobrir uma função que mapeie um conjunto de registos em um conjunto de variáveis predefinidas, denominadas classes. A tarefa de classificação é possível através de uma definição bem definida das classes e um conjunto de treino que consiste em

exemplos pré-classificados. Tal função pode ser aplicada em novos registos, de forma a prever a classe em que se enquadram. Vários métodos são aplicados na tarefa de classificação, mas as que mais se destacam são as árvores de decisão, as redes neurais artificiais, os classificadores Bayesianos, os conjuntos Fuzzy e os algoritmos genéticos (Berry & Linoff, 1997; Fu, 1997; Galvão & Marin, 2009).

O agrupamento separa os registos duma base de dados em subconjuntos ou agrupamentos, de tal forma que os registos dum agrupamento partilhem propriedades comuns, distinguindo-os de outros agrupamentos, maximizando a similaridade *intra-cluster* e minimizando a similaridade *inter-cluster*. Contrariamente à classificação, em que as variáveis são predefinidas, o agrupamento identifica automaticamente os grupos de dados e as suas respetivas variáveis. O método de análise de agrupamentos pode utilizar uma grande variedade de algoritmos, onde alguns dos mais utilizados são o K-Means, o KModes, o K-Protopypes, o K-Medoids e o Kohonem (Galvão & Marin, 2009; Gama et al., 2017).

3 Metodologia

De acordo com a metodologia seguida, apresentam-se em seguida os principais aspetos relativos ao entendimento do tema, à compreensão dos dados, à preparação dos dados, à modelação ("data mining") e à avaliação de resultados. A fase de desenvolvimento não foi efetuada até ao momento da elaboração deste artigo, pelo facto de necessitar da aprovação e o envolvimento efetivo da gestão do Instituto. Espera-se que tal fase possa vir a acontecer num futuro próximo.

3.1 Entendimento do tema

Vários estudos anteriores têm-se centrado na tentativa de entendimento das razões que levam ao sucesso escolar dos alunos. A motivação na realização, a autoeficácia académica, os objetivos académicos, o compromisso institucional, o apoio social percebido, as aptidões académicas ou as influências contextuais são alguns dos fatores que têm sido analisados por estudos anteriores (Robbins, Lauver, Le, Davis, Langley, & Carlstrom, 2004). Embora existam muitos fatores que possam influenciar o sucesso escolar, este estudo pretendeu analisar um conjunto deles, como a assiduidade nas aulas, o tempo de estudo, a frequência de estudo em grupo, a distância à localidade da residência, pode ter no sucesso escolar dos alunos de uma instituição de ensino superior. Para além disso, pretendeu ainda evidenciar as principais características dos grupos de alunos dessa instituição e como é que essas características se podem eventualmente relacionar.

Relativamente aos fatores assiduidade nas aulas ou tempo de estudo, era espectável que eles tivessem uma influência positiva no sucesso escolar. Estes fatores estão intimamente ligados com a motivação para alcançar o sucesso e o impulso associado a alcançá-lo, aspetos confirmados em estudos anteriores como essenciais nesse sucesso anteriores (Robbins et al., 2004). Um pouco

pelas mesmas razões, a espectativa quanto à frequência de estudo em grupo, também foi no sentido de uma sua influência positiva no sucesso escolar. No entanto, também se pode contra-argumentar que quanto mais frequente for esse tipo de estudo, menos frequente pode eventualmente ser o estudo individual. A distância à localidade da residência parece ser um fator inexplorado em estudos anteriores e que merece atenção. Se por um lado se pode argumentar que uma maior distância poderá estar associada a um eventual menor apoio familiar, por outro lado, também se pode argumentar que uma maior distância pode reduzir o número de viagens a casa, a intensidade e o tempo gasto na relação familiar, permitindo ao aluno aumentar o seu foco nas suas atividades académicas.

A instituição que serviu de base ao estudo apresentado é uma instituição de ensino superior portuguesa. O seu nome não será identificado, sendo designada por "Instituto". O Instituto está vocacionado para as áreas das ciências empresariais, caraterizando-se por oferecer diversos cursos breves, licenciaturas, mestrados e pós-graduações da área da gestão ao marketing, passando por várias outras áreas do conhecimento complementares.

O Instituto tem tido nos últimos anos um excelente nível de aceitação por parte dos candidatos às suas licenciaturas, preenchendo, por regra, as vagas disponíveis. O nível de emprego dos seus licenciados também pode ser considerado bom, não existindo um nível de desempregados significativo. No entanto, a nota média final dos licenciados é considerada baixa. Este facto, ou o nível de aprendizagem escolar médio que lhe possa ter dado origem, poderá eventualmente justificar o facto do nível salarial dos graduados do Instituto também ser relativamente baixo.

O trabalho desenvolvido pretendeu descobrir alguns dos principais fatores que levam um aluno a obter um maior ou menor sucesso escolar. O sucesso escolar pode-se entender como sendo o desenvolvimento de aprendizagens significativas em contexto escolar relativas a conhecimentos selecionados historicamente como relevantes para a vida na sociedade contemporânea. O processo de aprendizagem escolar não pode ser traduzido apenas por desempenhos em nível cognitivo e assim, é redutor avaliar o sucesso ou insucesso escolar com apenas as classificações obtidas pelo estudante (Gatti, 2010). Contudo, embora o método de classificação do desempenho dos alunos seja muitas vezes questionável, quer quanto à sua confiabilidade, quer quanto à sua validade, o modo mais frequente como esse sucesso é medido é precisamente através das classificações obtidas pelos estudantes (Eurydice, 1994; Saavedra, 2001). Este estudo irá igualmente usar a classificação do desempenho como o método de avaliação do sucesso escolar de um aluno.

3.2 Compreensão dos dados

Os dados resultaram de um inquérito feito pela Web a 86 alunos durante o ano de 2018. Durante este processo foi garantido a confidencialidade dos respondentes, não existindo nenhuma associação entre quem respondeu e a sua respetiva resposta. Teve-se em atenção algumas das

melhores práticas relativas à implementação de inquéritos baseado na Web, por forma a obter taxas de resposta satisfatórias, designadamente aspetos associados à seleção da ferramenta de *software*, ao *design* de questionário e às fases de administração de pesquisa (Belfo & Sousa, 2011). Os respondentes eram alunos do segundo e terceiro ano das respetivas licenciaturas. As questões do questionário (ver em anexo) contemplaram aspetos relativos a possíveis fatores explicativos do sucesso escolar como assiduidade nas aulas, tempo de estudo, frequência de estudo em grupo e distância à localidade da residência.

Para além de questões relativas aos referidos quatro fatores, o inquérito também colocou uma questão relativa ao sucesso escolar. A forma de medir o sucesso escolar foi feita com base na classificação média do estudante até ao momento da sua resposta. A questão colocada foi "De momento, qual é a tua média?".

Classe	Limite mínimo	Limite máximo
1	0%	24%
2	25%	49%
3	50%	74%
4	75%	100%

Tabela 1 – Classes relativas ao fator assiduidade nas aulas

Outra questão colocada foi "Qual é a tua assiduidade nas aulas?". As respostas possíveis para esta questão estão apresentadas na Tabela 1. As várias possibilidades de respostas dadas a escolher aos respondentes relativamente à assiduidade variaram desde a classe 1, a qual representa uma assiduidade até 24% das aulas, que caracterizará alunos muito pouco assíduos, até à classe 4, a qual retrata os alunos com maior assiduidade, correspondendo aos alunos que no mínimo frequentam 3/4 do número total de aulas.

Classe	Tipo de antecedência	Limite mínimo	Limite máximo
1	Estudo antes de uma avaliação	0 dias	2 dias
2		2 dias	4 dias
3		4 dias	-
4	Estudo diário	0 horas	2 horas
5		2 horas	4 horas
6		4 horas	-

Tabela 2 – Classes relativas ao fator tempo de estudo

A Tabela 2 representa as possibilidades de resposta quanto à questão "Geralmente, és um aluno que estuda", a qual recolheu o tempo de estudo diário ou antes de uma avaliação. Foi feita uma divisão em 6 classes, as quais correspondem aos diferentes tempos que o aluno investe no seu estudo. As possibilidades vão desde a classe 1, associada a uma frequência de estudo "apenas" até 2 dias antes de uma avaliação, até à classe 6 correspondente a um estudo diário de pelo menos 4 horas.

Classe	Frequência
1	Não, e não tiro dúvidas
2	Não, e tiro dúvidas
3	Sim, poucas vezes
4	Sim, algumas vezes
5	Sim, muitas vezes

Tabela 3 – Classes relativas ao fator frequência de estudo em grupo

Outra questão colocada foi "Estudas em grupo?" e as respostas possíveis foram as que se apresentam na Tabela 3, correspondentes aos diferentes níveis de estudo em grupo. Estas variam da classe 1, representando alunos que não estudam em grupo e nem tiram dúvidas, até à classe 5, representando alunos que estudam frequentemente em grupo.

Classe	Limite mínimo	Limite máximo
1	0 km	49 km
2	50 km	99 km
3	100 km	149 km
4	150 km	199 km
5	200 km	249 km
6	250 km	-

Tabela 4 – Classes relativas ao fator à distância da localidade de residência

A Tabela 4 apresenta as 6 classes do fator correspondente à distância entre a localidade em que os alunos vivem e o Instituto. As respostas resultaram da pergunta "Qual é a distância da tua localidade ao ISCAC?", a qual está dependente da pergunta "Em tempo de aulas, moras com os teus pais?". Todos os alunos, à exceção de 2, que disseram que em tempo de aulas moravam com os pais, indicaram que essa distância era inferior a 49 quilómetros. Na classe 1 estão representados

os alunos que vivem até 49 km. Optou-se por distribuir as restantes classes de 50 km em km até à classe 6, onde estão apresentados os alunos cuja residência se situa a uma distância igual ou superior a 250 km.

3.3 Preparação dos dados

Esta fase contemplou a organização e a inspeção dos dados. Foram preparados todos dados, definidos os formatos necessários para a análise e ajustadas demais questões técnicas. Das 87 respostas obtidas no inquérito foram aceites 86 respostas e rejeitada uma dessas respostas.

O inquérito contemplou questões relativas a outros possíveis fatores explicativos em ciclos anteriores do processo de descoberta de conhecimento. Esses outros fatores corresponderam às possibilidades de ser trabalhador estudante, de ter explicações ou o grau de satisfação com a licenciatura. No entanto, após adequada modelação, esses fatores não se revelaram influentes quanto ao grau de explicação que tinham sobre o sucesso escolar e foram abandonados nos modelos finais aqui apresentados. Este artigo apresenta apenas os modelos da última iteração deste estudo, os quais apenas consideram como possíveis explicações do sucesso escolar, os fatores assiduidade nas aulas, tempo de estudo, frequência de estudo em grupo e distância à localidade da residência.

3.4 Modelação (*Data Mining*)

Esta fase centra a sua atenção na escolha da técnica que será usada na modelação. Esta investigação pretendeu ser abrangente e, por isso, contemplou atividades preditivas e descritivas. Nas atividades preditivas, utilizou-se a técnica de indução de árvores de decisão com vista à classificação dos alunos em função dos diversos fatores exógenos considerados. No que diz respeito a atividades descritivas, utilizou-se a técnica de análise de agrupamento com vista à segmentação dos alunos. O software usado para executar os respetivos algoritmos foi o Weka, na sua versão 3.6.8. É um conhecido pacote de software de aprendizagem automática, escrito na linguagem Java e desenvolvido na Universidade de Waikato, na Nova Zelândia.

A técnica de agrupamento (*clustering*) mostrou resultados bastante interessantes, apresentando agrupamentos automáticos segundo o próprio grau de semelhança. O primeiro modelo de agrupamento testado traduziu-se na representação da similaridade de entre todos os alunos do Instituto que responderam ao questionário. Outros dois modelos foram testados com 2 e 3 agrupamentos, respetivamente. O algoritmo usado foi o *Simple EM* (*expectation maximisation*).

Attribute	Cluster 0 (1)
assiduidade	
mean	3.3488
std. dev.	0.8459
tempo_estudo	
mean	3.186
std. dev.	1.1565
estudo_grupo	
mean	3.2209
std. dev.	1.1753
distancia	
mean	2.1395
std. dev.	1.5032
media	
mean	12.7674
std. dev.	1.0419

Tabela 5 – Resultados do modelo com agrupamento único

O modelo com agrupamento único, resultado da utilização do algoritmo escolhido, é apresentado na Tabela 5 e representa o perfil do aluno típico do Instituto.

	Cluster	
Attribute	0	1
	(0.44)	(0.56)
assiduidade		
mean	2.5263	4
std. dev.	0.6381	0.8508
tempo_estudo		
mean	2.7368	3.5417
std. dev.	0.9917	1.154
estudo_grupo		
mean	3.6053	2.9167
std. dev.	1.2469	1.0172
distancia		
mean	2.5263	1.8333
std. dev.	1.5851	1.3591
media		
mean	12.4211	13.0417
std. dev.	0.9356	1.04

Tabela 6 – Resultados do modelo com dois agrupamentos

As caraterísticas do modelo com dois agrupamentos são apresentadas na Tabela 6.

	Cluster		
Attribute	0	1	2
	(0.46)	(0.25)	(0.3)
assiduidade			
mean	4	3.122	2.5263
std. dev.	0.8508	0.8739	0.6394
tempo_estudo			
mean	3.5924	2.913	2.7818
std. dev.	1.1578	1.0655	1.0095
estudo_grupo			
mean	2.8852	3.3903	3.6015
std. dev.	0.9933	1.2039	1.2626
distancia			
mean	1.2773	4.4295	1.5778
std. dev.	0.5245	1.1324	0.5747
media			
mean	12.9867	12.7638	12.4301
std. dev.	1.0966	0.9972	0.8893

Tabela 7 – Resultados do modelo com três agrupamentos

A Tabela 7 apresenta as caraterísticas de modelo com 3 agrupamentos.

Quanto à técnica de classificação, esta pode ser usada para dividir uma grande coleção de registos em conjuntos sucessivamente menores de registos através da aplicação de uma sequência de regras de decisão simples. Com cada divisão sucessiva, os membros dos conjuntos resultantes tornam-se mais e mais semelhantes uns aos outros. Deste modo, a utilização de árvores de decisões funciona com constantes testes, retratados na árvore como nós, de uma forma sucessiva, até que esta leve a uma conclusão, representada na árvore como uma das possíveis folhas, utilizando as regras que assim foram ditadas dentro da árvore. O algoritmo usado na modelação foi o *Random Tree*.

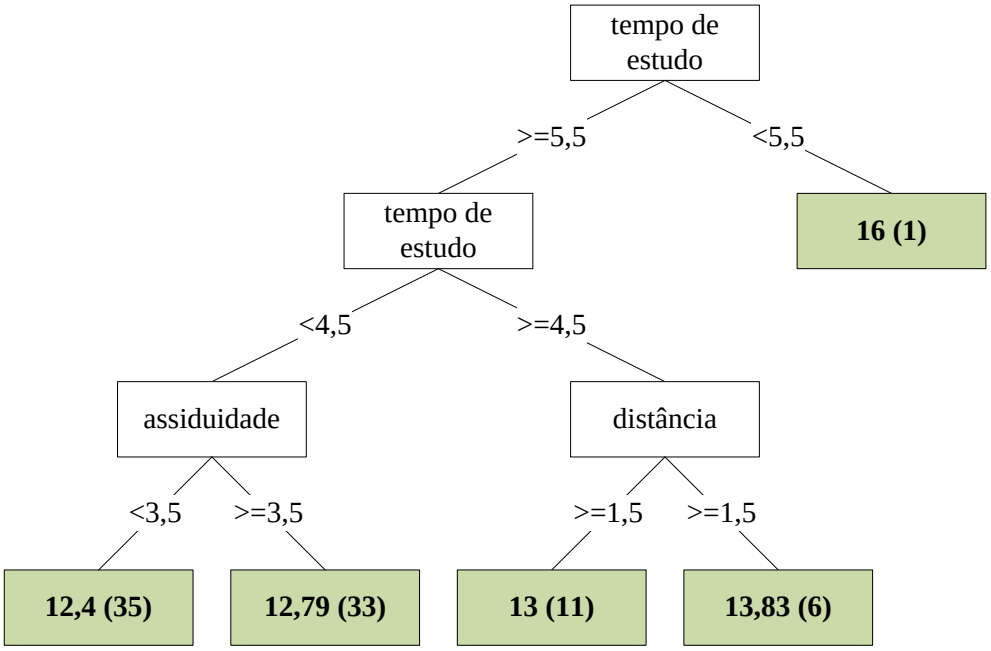

Figura 1– Resultados do modelo de árvore de decisão binária

A Figura 1 apresenta os resultados do modelo de classificação com base em árvore de decisão binária induzida a partir do algoritmo escolhido.

3.5 Avaliação de resultados

Esta secção apresenta a interpretação dos resultados dos três modelos construídos com a técnica de agrupamento e do modelo de classificação com recurso a uma árvore de decisão.

Os algoritmos de agrupamento encontram grupos homogéneos, evidenciando as suas características. O primeiro modelo de agrupamento, apresentado na Tabela 5, carateriza o aluno típico do Instituto. Esse aluno tem uma assiduidade média de 3,3488, o que significa que a sua frequência às aulas se situa entre a classe 3 (entre 50% a 74%) e a classe 4 (entre 75% a 100%). Se cada classe for associada à média aritmética dos seus limites, a média da classe 3 é de 62%, tal valor poderá ser somado ao valor diferencial da assiduidade média face a essa classe, o que dará um nível de aproximadamente 71%. Quanto ao tempo de estudo do aluno típico, a sua média é de

3,186 o que significa que se encontra entre uma situação de 4 dias de estudo antes de uma avaliação e um estudo diário, embora mínimo (inferior a 2 horas). Relativamente ao estudo em grupo, o perfil do aluno típico corresponde a uma classe média de 3,2209, o que significa que estuda em grupo, embora poucas ou algumas vezes. Ainda, no que diz respeito à distância a que vive da sua localidade, o aluno típico do Instituto tem uma média de 2,1395. Isso significa que se encontra aproximadamente na categoria que corresponde a uma distância entre 50 a 99 km, o que não lhe permitirá viver na sua residência em tempo de aulas. Por fim, apresenta uma classificação média de aproximadamente de 12,8 valores.

O segundo modelo de agrupamento, apresentado na Tabela 6, contempla dois grupos de alunos. Embora exista muita homogeneidade nos alunos do Instituto, o grupo identificado como 1, com 55 alunos, pode-se caraterizar como aquele que tem alunos um pouco melhores (classificação média de aproximadamente 13,0 valores) e o grupo 0, com 31 alunos, como aquele constituído por alunos que têm uma classificação um pouco mais baixa (classificação média de aproximadamente 12,4 valores). Apesar da diferença na classificação média entre os dois grupos não ser significativa, estes grupos apresentam diferenças significativas nos fatores que possam ditar o bom sucesso escolar. Esta técnica mostra diferenças enormes entre estes dois grupos relativamente ao fator assiduidade. Podemos observar que o grupo 0 tem uma assiduidade muito mais baixa que o grupo 1. Enquanto os alunos do grupo 0 apenas frequentam tipicamente entre 25% a 49% das aulas, o grupo 1 denominado, como o dos melhores alunos, frequentam 75% a 100% das aulas. Por outro lado, analisando o atributo tempo de estudo, o modelo carateriza o grupo 0 como sendo o correspondente ao dos alunos que estudam apenas entre 2 a 4 dias antes de uma avaliação e o grupo 1 como o daquele que estudam pelo menos 4 dias antes de uma avaliação, tendo tendência a estudar todos os dias. Em relação ao estudo em grupo também se observam diferenças entre os dois grupos. Enquanto o grupo 0 se carateriza por alunos que estudam em grupo, embora que poucas vezes, o grupo 1 não tem o hábito de estudar em grupo, tirando dúvidas posteriormente com alguém. Por último, no que se refere à distância a que cada aluno vive da sua localidade, verificam-se diferenças pouco significativas entre os dois grupos. Enquanto o grupo 0 vive entre 50 km a 99 km, o grupo 1 vive mais perto, num máximo até 49 km da sua localidade.

A Tabela 7 apresenta o terceiro modelo de agrupamento com três principais grupos de alunos típicos. No grupo 0 estão 41 alunos dos 86, representando praticamente metade dos alunos inquiridos. Embora as diferenças entre estes 3 grupos sejam pequenas quanto à classificação média obtida, podemos identificar o grupo 0 como aquele que corresponde aos melhores alunos, com uma média de aproximadamente 13,0 valores, podendo ser considerados os alunos que alcançam um melhor sucesso escolar. Relativamente ao grupo 1, fazem parte deste grupo 23 alunos,

alcançando uma média mais baixa que o grupo 0, conseguindo aproximadamente 12,8 valores. Por fim, o grupo 2 é constituído por 22 alunos e é o que tem menor classificação nesta amostra, com uma média aproximada de 12,4 valores.

Uma primeira questão é saber quais os fatores que mais ditam esta discrepância na média escolar, embora não muito significativa, entre os grupos. Analisando a Tabela 7, pode-se observar que um dos fatores que afeta o aproveitamento escolar do aluno é a sua assiduidade nas aulas. Observamos que os alunos com a melhor média frequentam as aulas entre 75% a 100%, ao contrário do pior perfil, onde apenas frequentam as aulas apenas entre 25% a 50%. Relativamente ao grupo de alunos intermédios, estes frequentam as aulas entre 50% a 75%, mais que o grupo dos piores e menos que o grupo dos melhores. Constatamos deste modo que o facto de os alunos irem frequentemente às aulas, com uma elevada assiduidade é um passo muito importante para conseguirem atingir um bom resultado escolar e assim um maior sucesso no seu percurso académico. Por outro lado, analisando o tempo que cada aluno despende para estudar, observamos que este é um dos fatores que mais influencia o sucesso escolar dos alunos. O grupo que mais tempo dedica a estudar é o dos melhores alunos. A diferença entre os três grupos ainda é significativa. O típico melhor aluno dedica até 2 horas de estudo diárias, enquanto os alunos dos outros dois grupos estudam apenas entre 2 a 4 dias antes de uma avaliação. Este resultado está em linha com o que seria previsível, já que um maior tempo de estudo ajuda numa melhor aquisição e consolidação do conhecimento e no consequente melhor sucesso escolar. Quanto ao fator de estudo em grupo, poderiam ser avançados argumentos de que alunos que o fazem com frequência poderiam conseguir alcançar melhores notas. No entanto, após pode-se constatar que afinal estudar em grupo, não é um hábito daqueles que conseguem uma melhor nota. O perfil dos melhores alunos mostra que estes não costumam estudar em grupo, embora posteriormente tirem as suas dúvidas com alguém. Em contrapartida, o perfil dos alunos intermédios bem como o perfil dos piores alunos mostra que estes costumam estar algumas vezes em grupo. Este resultado poderá justificar o argumento de que um maior tempo de estudo em grupo poderá reduzir o tempo disponível para o estudo individual e a eficiência do processo aquisitivo global de conhecimento por parte do aluno. Por último, neste modelo, analisando a distância a que cada aluno vive da sua localidade, observamos que é um fator que aparentemente não se relaciona significativamente com o sucesso escolar. As distâncias médias são idênticas para os melhores alunos e para o perfil dos piores alunos, mostrando que aparentemente a distância a que o aluno vive da sua localidade não parece influenciar significativamente o seu sucesso escolar.

O modelo que foi resultado da aplicação de técnica de indução de árvore de decisão está apresentado na Figura 1. A dimensão da árvore gerada inicialmente foi enorme, tendo-se optado pela sua poda e obtendo-se este modelo mais simplificado. Este modelo mostra que a raiz da árvore é o fator tempo de estudo, representando assim o fator mais importante no que diz respeito

ao sucesso escolar. Ou seja, este fator é o que separa melhor todos os casos. Este resultado é coerente com os modelos anteriormente apresentados. Num dos primeiros ramos da árvore, aquele em que o tempo disponível tem classe >= 5.5, verifica-se que quando um aluno dedica ao seu estudo pelo menos 3 horas diárias, este fator é determinante para o seu sucesso escolar. Neste caso, é encontrada logo uma folha que corresponde a 1 aluno, cuja média é de 16 valores.

Se por outro lado, o tempo de estudo representar menos do que 3 horas diárias (classe <5.5), este fator é ainda dividido em duas possibilidades: quando estudam menos de 1 hora por dia (<4.5) e quando estudam 1 hora ou mais por dia (>=4.5). Analisando os alunos que dedicam 1 hora ou mais por dia, isto leva ainda a um outro fator para determinar o seu sucesso escolar: a distância. É relevante perceber se o aluno mora a menos de 25 quilómetros (classe <1.5) ou pelo menos a 25 quilómetros (>=1.5) da sua localidade. Pode-se concluir que, por um lado, o modelo indica que os alunos que moram mais perto da instituição têm tendência a alcançar uma média de 13 valores. Por outro lado, os alunos que moram mais longe têm tendência a alcançar uma classificação média superior (14 valores). Este resultado parece dar consistência ao argumento de que uma maior distância pode reduzir o número de viagens a casa, a intensidade e o tempo gasto na relação familiar, permitindo ao aluno aumentar o seu foco nas suas atividades académicas e assim, o seu sucesso.

Por último, estudando agora os casos dos alunos que estudam menos de 1 hora por dia, o fator que condiciona e que divide os casos é o fator assiduidade. Se um aluno frequenta menos de 60% das aulas a média é de aproximadamente 12 valores. Nos casos dos alunos que têm uma assiduidade de pelo menos 60%, estes têm uma média de 13 valores. Isto comprova que, embora o tempo de estudo médio seja mais importante, a assiduidade é também um fator muito determinante para o sucesso escolar dos alunos.

4 Conclusões e trabalhos futuros

Este artigo descreveu um projeto de descoberta de conhecimento em bases de dados numa instituição do ensino superior portuguesa. As etapas contemplaram tarefas como entendimento do tema, compreensão dos dados, preparação dos dados, modelação e avaliação dos resultados. Os dados resultaram de um inquérito feito a uma amostra de 86 alunos de várias licenciaturas da instituição.

Os objetivos de descoberta de conhecimento em bases de dados que foram traçados para este projeto foram atingidos. Este estudo contribui para um melhor entendimento sobre o que condiciona o sucesso escolar de um aluno que está a tirar uma licenciatura no Instituto. Embora os alunos inquiridos sejam muito homogéneos quanto ao seu sucesso escolar, foi possível entender quais são os fatores mais relevantes que ditam as diferenças entre os melhores e os piores alunos.

Se por um lado a técnica de agrupamento utilizada permitiu perceber as diferenças entre os grupos dos melhores e dos piores alunos, a técnica de classificação permitiu dar a entender como é que um aluno consegue alcançar melhor sucesso escolar. Os resultados relativos às árvores de indução induzidas revelaram que o fator mais importante para justificar o sucesso escolar é o tempo de estudo. Por outro lado, a árvore de decisão permitiu ainda classificar o sucesso escolar em função dos diversos fatores explicativos. Os resultados relativos ao agrupamento permitiram caraterizar o aluno típico e ainda os dois ou três principais grupos de alunos, esclarecendo as principais diferenças entre esses agrupamentos.

A avaliação dos resultados dos modelos de agrupamento e de classificação leva-nos a concluir que ambos os tipos de modelos mostram conclusões relativamente semelhantes quanto aos fatores que mais determinam o sucesso escolar dos alunos. Ficou evidente que, mesmo não necessariamente com a mesma importância, os fatores tempo de estudo, distância à localidade de residência e a assiduidade nas aulas são fatores que influenciam significativamente o sucesso escolar. O mesmo já não se verificou quanto ao estudo em grupo.

O processo de descoberta de conhecimento de dados e o *data mining* podem representar uma mais-valia para as empresas e organizações que procuram conhecer melhor o seu negócio, aumentando o conhecimento que lhes permita obter vantagens competitivas face aos seus adversários. A interpretação efetuada aos resultados dos modelos de agrupamento e classificação poderá permitir à instituição em questão a definição das bases para estratégias que permitam melhorar o sucesso escolar dos seus alunos. Ao perceber-se quais aos fatores que mais influenciam o sucesso escolar, a instituição pode desenvolver iniciativas que influenciem esses mesmos fatores e, dessa forma, combater o indesejável insucesso escolar.

Como trabalhos futuros sugere-se que este estudo passe à fase de desenvolvimento, definindo-se e implementando-se estratégias que, tendo por base o conhecimento adquirido neste estudo, promovam a melhoria do sucesso escolar no Instituto. Por outro lado, este tipo de estudo poderá ser feito ao nível de cada licenciatura. Dever-se-á garantir nesse caso que o número de estudantes inquiridos por curso tenha um número mínimo de respondentes. A realidade para cada licenciatura não é necessariamente igual e poderá necessitar de estratégias diferentes para aumentar o sucesso escolar. Por outro lado, outras variáveis explicativas poderão ser consideradas em futuros estudos.

5 Referências

Azadmanesh, S., & Tarokh, M. J. (2012). Labeling Customers using Discovered Knowledge Case Study: Automobile Industry. *International Journal of Managing Value and Supply Chains (IJMVSC), 3*(3), 13-24.

Belfo, F., & Sousa, R. D. (2011). A web survey implementation framework: evidence-based design practices. Proceedings of *MCIS 2011 - 6th Mediterranean Conference on*

Information Systems held in Cyprus, 2011, 3-5 Sep (pp. Paper 43). Association for Information Systems (AIS).

Belfo, F., & Sousa, R. D. (2013). Reviewing Business-IT Alignment Instruments Under SAM Dimensions. *International Journal of Information Communication Technologies and Human Development, 5*(3), 18-40. doi: 10.4018/jicthd.2013070102

Berry, M. J., & Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York, USA: John Wiley & Sons, Inc.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. 2000. *CRISP-DM 1.0: Step-By-Step Data Mining Guide*. U.S.A.: SPSS, CRISP-DM Consortium.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine, 26*(1-2), 1-24.

Eurostat. (2017). *Europe 2020 Indicators - Education, Statistics Explained*. Eurostat.

Eurydice. 1994. *Measures to Combat Failure at School: A Challenge for the Construction of Europe*. Bruxelas: Comissão Europeia.

Fu, Y. (1997). Data Mining: Tasks, Techniques, and Applications. *IEEE Potentials, 16*(4), 18-20.

Galvão, N. D., & Marin, H. d. F. (2009). Data Mining: A Literature Review. *Acta Paulista de Enfermagem, 22*(5), 686-690.

Gama, J., Carvalho, A. C. P. d. L., Faceli, K., Lorena, A. C., & Oliveira, M. (2017). *Extração de Conhecimento de Dados: Data Mining* (3 ed.). Lisboa: Edições Silabo.

Gartner. (2018). Data Mining. *IT Glossary.* Gartner, Inc. Retrieved 2018, May 20, from https://www.gartner.com/it-glossary/data-mining

Gatti, B. A. (2010). Sucesso Escolar. In D.A. Oliveira, A.M.C. Duarte & L.M.F. Vieira (Eds.), *Dicionário: Trabalho, Profissão e Condição Docente*. Belo Horizonte: Universidade Federal de Minas Gerais.

Jackson, J. (2002). Data Mining: A conceptual overview. Proceedings of *Communications of the Association for Information Systems* held 267-296). 8.

Kappelman, L., Nguyen, Q., McLean, E., Maurer, C., Johnson, V., Snyder, M., et al. (2017). The 2016 SIM IT Issues and Trends Study. *MIS Quarterly Executive, 16*(1), 47-80.

Larose, D. T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Loureiro, A., Lourenço, J., Costa, E., & Belfo, F. (2014). *Indução de Árvores de Decisão na Descoberta de Conhecimento: Caso de Empresa de Organização de Eventos*. Paper presented at the VI Congresso Internacional de Casos Docentes em Marketing Público e Não Lucrativo, ISCAC Business School, Coimbra, Portugal.

Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559-569.

Reich, B. H., & Benbasat, I. (1996). Measuring the linkage between business and information technology objectives. *MIS Quarterly, 20*(1), 55-81.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261.

Saavedra, L. (2001). Sucesso-insucesso escolar: A importância do nível socioeconómico e do género. *Psicologia, XV*(1), 67-92.

Anexo: Inquérito Utilizado

1. **Qual o curso que frequentas?** *
   Marcar apenas uma oval.
   - ◯ Informática de Gestão
   - ◯ Gestão de Empresas
   - ◯ Contabilidade e Gestão Pública
   - ◯ Contabilidade e Auditoria
   - ◯ Marketing e Negócios Internacionais
   - ◯ Secretariado de Direção e Administração
   - ◯ Solicitadoria e Administração

2. **Em que ano escolar estás?** *
   Marcar apenas uma oval.
   - ◯ 2º Ano
   - ◯ 3º Ano

3. **Qual o teu número de matriculas?** *
   _____

4. **Qual é a tua assiduidade nas aulas?** *
   Marcar apenas uma oval.
   - ◯ 0-24%
   - ◯ 25-49%
   - ◯ 50-74%
   - ◯ 75-100%

5. **Geralmente, és um aluno que estuda:** *
   Marcar apenas uma oval.
   - ◯ Diariamente, até 2h
   - ◯ Diariamente, entre 2h a 4h
   - ◯ Diariamente, mais de 4h
   - ◯ Antes de uma avaliação, até 2 dias;
   - ◯ Antes de uma avaliação, entre 2 a 4 dias;
   - ◯ Antes de uma avaliação, mais de 4 dias.

6. **Estudas em grupo?** *
Marcar apenas uma oval.

- ◯ Sim, na maior parte das vezes
- ◯ Sim, com alguma frequência
- ◯ Sim, embora seja raro estudar em grupo
- ◯ Não, embora tire as dúvidas com alguém
- ◯ Não, e não costumo tirar dúvidas com ninguém

7. **És trabalhador-Estudante?** *
Marcar apenas uma oval.

- ◯ Sim, sou trabalhador-estudante a menos de 25% do meu tempo
- ◯ Sim, sou trabalhador-estudante entre 25% a 50% do meu tempo
- ◯ Sim, sou trabalhador-estudante entre 50% a 75% do meu tempo
- ◯ Sim, a full-time
- ◯ Não

8. **Durante a licenciatura, já tiveste explicações?** *
Marcar apenas uma oval.

- ◯ Sim, tenho explicações frequentemente
- ◯ Sim, já recorri significativas vezes a explicações
- ◯ Sim, já recorri esporadicamente a explicações
- ◯ Não, nunca tive explicações

9. **Gostas do teu curso?** *
Marcar apenas uma oval.

- ◯ Não gosto nada
- ◯ Baixo
- ◯ Intermédio
- ◯ Gosto
- ◯ Gosto muito

10. **Em tempo de aulas, moras com os teus pais?** *
Marcar apenas uma oval.

- ◯ Sim
- ◯ Não, moro noutro local, embora esteja com eles com muita frequência
- ◯ Não, moro noutro local, embora ainda esteja com eles algumas vezes
- ◯ Não, moro noutro local, e raramente estou com eles

11. **Qual é a distância da tua localidade ao ISCAC?** *
Marcar apenas uma oval.

- ◯ Até 49km;
- ◯ De 50km até 99km;
- ◯ De 100km até 149km;
- ◯ De 150km até 199km;
- ◯ De 200 até 249km;
- ◯ + De 250km.

12. **De momento, qual é a tua média?** *

13. **Achas que a tua nota final de curso vai influenciar o teu futuro profissional?** *
Marcar apenas uma oval.

- ◯ Sim, penso que a minha nota vai influenciar decisivamente o meu futuro profissional
- ◯ Sim, penso que a minha nota vai influenciar razoavelmente o meu futuro profissional
- ◯ Sim, penso que a minha nota vai influenciar pouco o meu futuro profissional
- ◯ Não

14. **Porquê?** *

Com tecnologia
Google Forms