Descoberta de Conhecimento para Identificação de Fatores que Influenciam o Desempenho Discente

Knowledge Discovery to Identify the Factors which have Influence on the Academic Performance

Roberto Gonçalves Augusto Junior
Universidade do Vale do Itajaí

Guilherme Augusto Rosa Carminati
Universidade do Vale do Itajaí

Andre Luis Alice Raabe
Universidade do Vale do Itajaí

Raimundo Celeste Ghizoni Teive
Universidade do Vale do Itajaí

Resumo: A identificação de possíveis perfis de desempenho acadêmico, logo nas primeiras fases de um curso de graduação, pode ser um conhecimento útil para gestores de uma Instituição de Ensino Superior (IES). Conhecendo os fatores que contribuem para um baixo desempenho acadêmico, ações podem ser tomadas para melhorar este desempenho acadêmico ou, em alguns casos, até prevenir uma evasão indesejada. Neste contexto, este artigo apresenta a aplicação de algoritmos de mineração de dados em bases de dados de sistemas acadêmico e financeiro de alunos egressos dos cursos de Direito e Engenharia Civil, buscando identificar padrões de desempenho acadêmico e os fatores associados. Os cursos foram selecionados por serem de diferentes áreas e por terem maior número de egressos na IES estudada. Os resultados obtidos apontam para evidências interessantes sobre o impacto de alguns atributos no desempenho acadêmico, tais como: tipo de disciplina, tipo de ingresso, uso do Ambiente Virtual de Aprendizagem e biblioteca; além das notas e frequência nas primeiros semestres.

Palavras-chave: Descoberta de conhecimento. Mineração de dados. Desempenho acadêmico.

Abstract: The identification of academic profiles in the first stages of graduation might be a useful knowledge for the managers of a higher education institution (IES). Knowing the factors that contribute for the low performance of students, actions can be taken to improve students' performance, or even prevent their dropout. In this paper, in the looking for identify profiles of students with particular academic performances and the factors linked to them, several data mining algorithms were applied to databases of the courses of Law and Civil Engineering. These courses were chosen considering the fact they have the highest number of students in the institution in analysis. The results present interesting evidences that some attributes have a stronger impact in the academic performance, such as: type of class, type of ingress, student's usage of library and VLE; as well as the frequency and grades of the first semesters.

Keywords: Educational data mining. Knowledge discovery. Academic performance.

AUGUSTO JUNIOR, ROBERTO GONÇALVES; CARMINATI, GUILHERME AUGUSTO ROSA; RAABE, ANDRE LUIS ALICE; TEIVE, RAIMUNDO CELESTE GHIZONI. Descoberta de Conhecimento para Identificação de Fatores que influenciam o Desempenho Discente. *Informática na Educação*: teoria & prática, Porto Alegre, v. 22, n. 3, p. 58-82, set./dez. 2019.

1 Introdução

A gestão informatizada da vida acadêmica gera, para as instituições, um volume cada vez maior de dados, que muitas vezes são utilizados apenas em relatórios administrativos. A disponibilização aos gestores de Instituições de Ensino Superior (IES) de informações acadêmicas, obtidas a partir destes dados, pode ser considerada um grande desafio (TRANDAFILI et al., 2012).

Dentre as tecnologias com potencial de promover ganhos na área educacional está a Descoberta de Conhecimento em Base de Dados (*knowledge discovery in databases* - KDD), que possui, entre suas etapas a principal: a mineração de dados (MD), com o objetivo de buscar conhecimentos novos e úteis no processo de ensino e aprendizagem.

Neste contexto, o presente trabalho apresenta a aplicação de técnicas de MD para identificar aspectos ligados ao ambiente de ensino e de aprendizagem que podem impactar no desempenho acadêmico de estudantes de graduação de uma IES. Além disto, busca-se analisar a influência destes fatores em cursos de diferentes áreas, tais como Engenharias e Ciências Sociais Aplicadas. Adota-se o termo ambiente de ensino e de aprendizagem para representar um conjunto de variáveis referentes à vivência de um estudante em uma IES.

Para esta análise foram utilizados dados do sistema de gestão acadêmica da Universidade do Vale do Itajaí – Univali, uma IES localizada no litoral norte do estado de Santa Catarina. A Univali é uma universidade comunitária e possui cerca de 25 mil alunos distribuídos em mais de 50 cursos de Graduação, 9 mestrados e 6 doutorados.

Durante mais de duas décadas, a IES analisada vem construindo, gradativamente, sistemas de informação, capazes de armazenar dados de diversas áreas, contemplando, por exemplo, matrícula de alunos, frequência, uso da biblioteca, ambiente virtual de aprendizagem, dados financeiros e planos de ensino. Estes dados não são correlacionados em sua totalidade, e passam apenas por análises parciais, como a busca por eficiência na aquisição de livros e melhoria de matrizes curriculares, sempre focadas na solução de problemas específicos identificados pelos gestores de cursos da IES.

A literatura fornece evidências (GOTTARDO; KAESTNER; NORONHA, 2014; RIGO et al., 2014) de que estes dados podem esconder informações pedagogicamente relevantes. Como exemplo, a descoberta de características comuns em alunos com bom desempenho acadêmico, a qual pode auxiliar na tomada de decisões que visam a melhoria do desempenho acadêmico de outros alunos. Na mesma linha, conhecer fatores comuns em alunos com baixo desempenho também permite intervenções que tenham como objetivo tentar mitigar ou evitar tais fatores.

No contexto deste trabalho, com o intuito de classificar os alunos em termos do seu desempenho discente, foi utilizado o termo "Grupos de Desempenho Acadêmico" (GDA), o qual se refere à média das notas de todas as disciplinas cursadas ao final de sua graduação.

A identificação precoce do GDA a que o aluno se direciona durante sua graduação pode viabilizar a construção de um sistema de alerta que permita aos coordenadores de curso, auxiliarem alunos que estiverem caminhando para um baixo desempenho acadêmico, por

exemplo; ou eventualmente potencializar àqueles que possuem indicativos de que irão figurar no grupo de bom desempenho acadêmico. Um sistema de alerta precoce como este foi descrito por Macfadyen e Dawson (2010), com acurácia de até 70%. Neste caso, os autores consideraram apenas dados de um ambiente Virtual de Aprendizagem (AVA) e com foco específico na educação a distância.

2 Mineração de Dados Educacionais

A Mineração de dados educacionais (MDE) é definida em Dutt, Ismail e Heravan (2017) como uma área interdisciplinar, na qual são aplicadas técnicas de mineração de dados, estatística, aprendizado de máquina, recuperação da informação, entre outras; para resolver questões educacionais. O objetivo principal seria o de aplicar estas técnicas para compreender melhor os estudantes e seus perfis de aprendizado. Em Slater et al. (2017), são analisadas as principais ferramentas computacionais para realizar a MDE.

Devido à importância dessa aplicação de MD, em Baker, Isotani e Carvalho (2011) a mesma é apresentada como uma área de pesquisa, que vem sendo chamada de "Mineração de Dados Educacionais" (do inglês, "*Education Data Mining*", ou EDM), cuja comunidade, segundo estes autores, cresce rapidamente no mundo, e em ritmo um pouco mais lento no Brasil.

Na análise da literatura nota-se principalmente uma preocupação com a identificação precoce do risco de reprovação de alunos, como por exemplo nos trabalhos de Macfadyen e Dawson (2010), Zhang (2010) e Samaranayake e Caldera (2012), o que é justificável, pois são esses os alunos que se pretende acompanhar.

Entretanto, a análise do perfil dos alunos, independente do risco de reprovação, como apresentado nos trabalhos de Trandafili et al. (2012), Carmona et al. (2011) e Hoe et al. (2013), podem dar aos especialistas em educação informações importantes, auxiliando na melhoria do desempenho acadêmico de todos os alunos, independente se estão em risco de reprovação ou não.

Este tipo de análise pode ser feita como apresenta o trabalho Zeng e Zheng (2009), que busca diferenciar qual gênero se sai melhor em determinados assuntos. Podem ser realizadas análises como as desenvolvidas nos trabalhos de Macfadyen e Dawson (2010), Zhang (2010), Samaranayake e Caldera (2012), que se preocupam com alunos de baixo rendimento ou como Trandafili et al. (2012), Carmona et al. (2011), Hoe et al. (2013), que têm o intuito de entender o que leva um aluno a determinado desempenho acadêmico. Nos artigos citados o nome dado a um conjunto de notas correlatas (altas, baixas ou intermediárias) varia entre os autores.

Em Asif el al. (2017), busca-se identificar o desempenho de alunos de uma disciplina específica (Tecnologia da Informação), considerando o desempenho do aluno no ensino médio, além da notas em algumas disciplinas cursadas anteriormente na universidade, por estes alunos, as quais seriam "pré-requisitos" da disciplina de Tecnologia da Informação. Neste trabalho, utilizou-se algoritmos de classificação e clusterização com o *software Rapidminer* (RAPIDMINER, 2019).

Porém, constata-se que nenhum dos trabalhos analisados considera um grande número de variáveis e a correlação é feita basicamente entre disciplinas, notas e gênero do aluno. Nos trabalhos que analisam o uso de AVAs, variáveis sobre a utilização do sistema, como tempo de utilização do AVA e número de exercícios executados são analisadas, mas variáveis importantes como titulação do professor; carga horária teórica; carga horária prática e locação de livros na biblioteca não são consideradas.

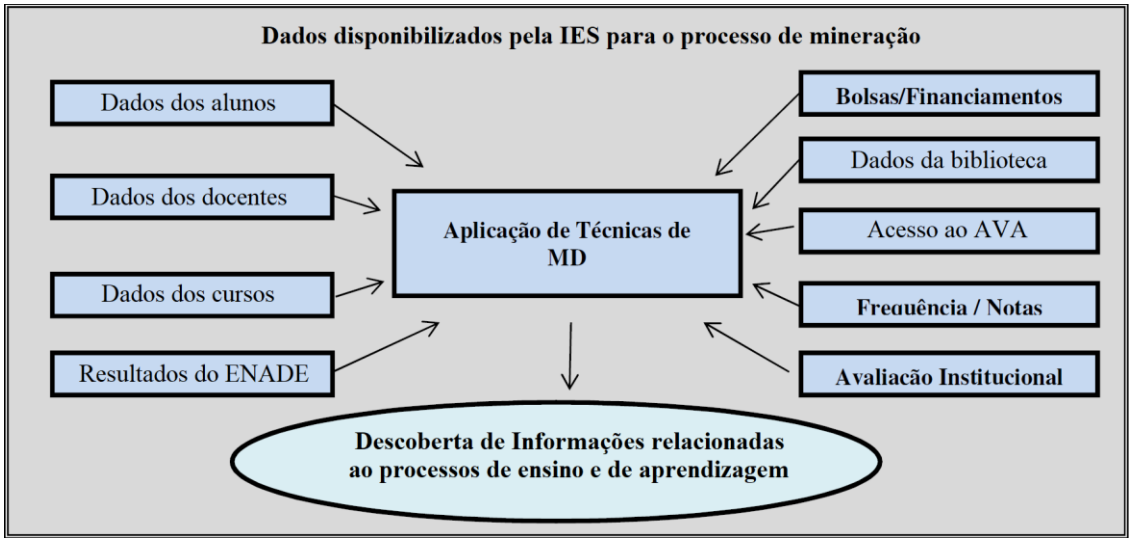
Além disto, dentre os trabalhos analisados não foi verificada nenhuma preocupação em analisar as características que podem levar alunos a um GDA e se há alguma diferença de desempenho entre cursos de áreas de conhecimento distintas.

Os motivos que levam à reprovação em dois cursos (Física e Biologia) são analisados em Samaranayake e Caldera (2012), sendo um exemplo de comparação de cursos de áreas distintas, mas com um enfoque diferente do pretendido neste trabalho. A análise destes trabalhos trouxe também a percepção de que existe uma lacuna em estudos que analisam dados na busca de padrões que levam alunos a obter um bom desempenho acadêmico ou não. Parte desta lacuna é composta pelo baixo número de variáveis encontradas nos estudos e a falta de correlação entre elas. A inclusão de outras variáveis também permite uma investigação do potencial destes dados para o desenvolvimento de um sistema de alerta precoce sobre o desempenho acadêmico do aluno, semelhante ao trabalho de Macfadyen e Dawson (2010), mas focado em cursos presenciais.

3 Dados Disponíveis na IES

No contexto deste trabalho, o ambiente de ensino e de aprendizagem resume as informações de alunos, professores, matriz curricular do curso, dados financeiros, dados de utilização de biblioteca, resultados do Exame Nacional de Desempenho de Estudantes (ENADE) e dados da Avaliação Institucional. Estes dados são apresentados de forma esquemática na Figura 1 e descritos em mais detalhes na sequência.

Figura 1 - Dados disponíveis para o processo de KDD



Fonte: Dos autores, 2019

Neste trabalho utilizou-se dados de alunos egressos dos cursos de Direito e do curso de Engenharia Civil. Os cursos foram selecionados por serem de diferentes áreas tendo como referência as áreas do Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq e por serem cursos com maior número de egressos em suas áreas para os anos de 2013 e 2014, considerando a IES foco deste estudo.

Foram considerados todos os turnos do curso de Direito (Matutino e Noturno) e também todos os turnos do curso de Engenharia Civil (Integral, Diurno e Vespertino/Noturno), e os alunos matriculados nos Campus de Itajaí. O número de alunos considerados neste estudo está apresentado na Tabela 1.

Tabela 1 - Resumo da população

Curso	**Nro. Egressos até 17/06/2014**	**Formados entre 2009 e 2014**
Direito	5410	1293
Engenharia Civil	454	285

Fonte: Dos autores, 2019

As variáveis analisadas foram:

- Aluno: idade; sexo; portador de necessidades especiais; procedente de escola pública ou privada; se o curso em questão é uma segunda graduação; frequência do aluno; notas dos alunos em disciplinas teóricas; notas dos alunos em disciplinas práticas; quantidade de reprovações; número de acessos ao ambiente virtual de aprendizagem

(AVA); índice de carência financeira do aluno (calculado com base na portaria número 37/2014 da Secretaria do Estado de Santa Catarina); se o aluno foi inadimplente (houve negociação financeira para parcelamento de débitos ao final de algum semestre)

- Docente: número de professores com título de doutor no curso; número de professores com título de mestre no curso; número de professores especialistas no curso; número de professores graduados no curso; número de professores com dedicação integral na instituição; número de professores com dedicação parcial na instituição; número de professores horistas (carga horária menor que 12 horas aula) na instituição e no curso;
- Curso: carga horária teórica do curso; carga horária prática do curso; resultado do ENADE para o curso;
- Bolsas/Financiamentos: se o aluno é ou não bolsista; se o aluno possui ou não financiamento estudantil;
- Biblioteca: quantidade de livros locados;
- Avaliação Institucional: satisfação do aluno com professores na avaliação institucional; satisfação de alunos com infraestrutura da IES na avaliação institucional; como o aluno avalia o nível de importância das disciplinas; como o aluno avalia o nível de exigência das disciplinas; como o aluno avalia o nível de exigência das disciplinas; como o aluno avalia os docentes com relação à articulação teoria e prática; como o aluno avalia os docentes com relação à aplicação do plano de ensino.

O banco de dados da IES **não** contempla informações importantes como número de horas de estudo extraclasse, trabalhos voluntários na área do curso, conhecimento de idiomas, quociente de inteligência do aluno, dentre outras. Por isso, estes aspectos não foram considerados na análise.

4 Metodologia

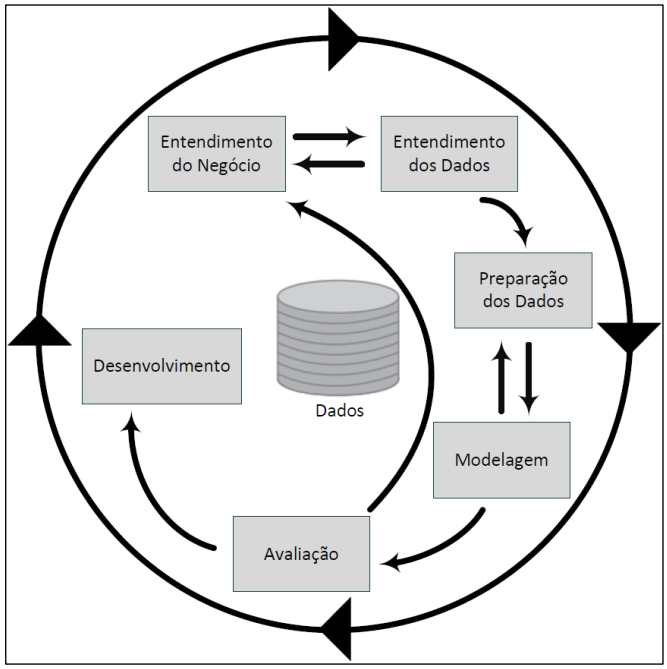
4.1 Metodologia KDD

Mariscal, Marbán e Fernández (2010) desenvolveram um trabalho com o intuito de descrever os principais processos e metodologias que são utilizadas para o processo de MD e KDD, dentre estes destaca-se o CRISP-DM, o qual foi utilizado como metodologia de KDD neste trabalho.

A metodologia CRISP-DM é acrônimo para *CRoss Industry Standard Process for Data*, a qual foi concebida pensando na independência do processo em relação à ferramenta e área de aplicação, sendo por isso adotada neste trabalho. É composta de seis etapas: entendimento do negócio; entendimento dos dados; preparação dos dados; modelagem; avaliação e desenvolvimento (MARISCAL; MARBÁN; FERNÁNDEZ, 2010), conforme descrito na Figura 2. Um manual desta metodologia pode ser encontrado em (IBM, 2011).

As fases da metodologia CRISP-DM estão no nível mais alto de abstração da metodologia. As iterações entre as fases podem ser vistas como um ciclo de vida da mineração de dados.

Figura 2 - Fases da metodologia CRISP-DM



Fonte: Adaptado de Chapman et al. (2000)

4.2 Escolha da Ferramenta

A metodologia CRISP-DM flexibiliza a escolha das ferramentas que podem ser utilizadas, flexibilidade esta que direcionou este trabalho ao desafio de escolher a ferramenta de mineração de dados mais adequada. Na literatura são encontradas aplicações de KDD com diversas ferramentas computacionais, não havendo consenso, nem tendência. Nos trabalhos de Trandafili et al. (2012), Zhang (2010), Carmona et al. (2011), Hoe et al. (2013), por exemplo, são descritas as várias ferramentas utilizadas em seus trabalhos, destacando-se Weka (WEKA, 2019). Porém, em nenhum destes trabalhos é apresentada uma justificativa para estas escolhas.

Assim, por não haver, nos trabalhos citados, uma ferramenta que se destaque, optou-se por realizar uma pesquisa e avaliação de possíveis opções. Neste sentido, foi realizada uma busca por trabalhos realizados pela comunidade científica que pudessem nortear a escolha da ferramenta a ser utilizada. Nesta busca foi localizado o trabalho realizado por Mikut e Reischl (2011).

A partir da relação de 89 ferramentas relacionadas por Mikut e Reischl (2011), 53 comerciais e 36 com uso livre, chegou-se a quatro ferramentas de uso livre, que atendiam aos requisitos necessários para a elaboração deste estudo; são elas: *Knime* (KNIME, 2019), *Rapidminer* (RAPIDMINER, 2019), *Tanagra* (RAKOTOMALALA, 2005) e *Weka* (WEKA, 2019)

Destas quatro ferramentas, o *Rapidminer* foi escolhido para ser utilizado neste trabalho, principalmente em função dos algoritmos disponíveis para cada tarefa de MD necessária para a realização desta pesquisa, além da facilidade de trabalhar com banco de dados.

4.3 Medidas de Desempenho

Nos trabalhos de Macfadyen e Dawson (2010), Samaranayake e Caldera (2012) e Hoe et al. (2013), a acurácia é a medida de desempenho utilizada, não sendo citada nenhuma outra. Assim, esta foi a medida utilizada para otimização dos algoritmos e modelos de classificação das características que se repetem em cada grupo de desempenho acadêmico.

Já no modelo de classificação do desempenho acadêmico do egresso, baseado nos semestres iniciais, optou-se por utilizar duas medidas de desempenho: acurácia e estatística *Kappa*. Esta última métrica foi selecionada por se tratar de uma medida de desempenho que retira da taxa de acerto da classificação, a probabilidade de acerto aleatório; sendo assim uma medida de desempenho mais crítica que a acurácia (GWET, 2012; POWERS, 2012).

4.4 Algoritmos Utilizados

Optou-se por utilizar apenas algoritmos que tivessem como saídas regras de associação ou árvores de decisão, pois são mais facilmente interpretáveis do que modelos de redes reurais e regressão linear, por exemplo; facilitando assim a compreensão dos gestores da IES.

As saídas de algoritmos que geram árvores de decisão foram ainda convertidas para regras de associação, deixando homogênea a documentação dos resultados.

Os algoritmos que foram utilizados neste trabalho são: Decision Tree, Randon Tree, CHAID, Decision Stump, ID3, Rule Induction e Single Rule Induction.

4.5 Dados de treinamento e teste

Os experimentos que efetivamente buscam gerar um modelo e classificação de alunos, foram realizados a partir de dois conjuntos de dados para cada curso analisado, seguindo a técnica *holdout* (GOLDSCHMIDT; PASSOS, 2005).

No primeiro conjunto de dados, foram selecionados aleatoriamente 80% dos registros disponíveis no curso. No segundo conjunto de dados, foram agrupados os 20% restantes. O conjunto de dados com 80% dos registros foi utilizado para gerar o modelo de classificação. O conjunto de dados com os 20% dos registros restantes foi utilizado para testes do modelo. A partir destes testes obtiveram-se as medidas de desempenho descritas nas seções 5.5, 5.6 e 5.7.

5 Resultados

Para facilitar a compreensão os resultados apresentados, foi elaborada a Tabela 2 com as siglas utilizadas como nomes de atributos e suas respectivas descrições.

Tabela 2 - Siglas utilizadas nos resultados

Sigla	Descrição
ALU_DIS_APR_QTD4	Quantidade de disciplinas aprovadas nos quatro períodos iniciais do curso.
ALU_DIS_FALTAS	Número de faltas
ALU_DIS_FALTAS4	Número de faltas nos 4 períodos iniciais do curso.
BADA	Baixo Desempenho Acadêmico
BDA	Bom Desempenho Acadêmico
DIS_HOR_LAB	Carga horária de laboratório (prática)
DIS_HOR_PRATICAS	Carga horária prática da disciplina
DIS_HOR_TEORICAS	Carga horária teórica da disciplina
DIS_OBR	Disciplina Obrigatória ou não obrigatória (optativa). Apresentada com **S**:Sim; **N**:Não.
DIS_TIP	Tipo de disciplina, sendo **N**:Normal; **E**:Estágio; **P**:Projetuais; **T**: TCC.
GDA	Grupo de Desempenho Acadêmico (BADA, MDA e BDA)
GDA_PER4	GDA nos 4 períodos iniciais do curso
GDA_SUJA_PER4	GDA nos 4 períodos iniciais do curso, considerando reprovações.

GRU_INGR	Grupo de ingresso do aluno, sendo: **TEX**: Transferência Externa; **TIN**: Transferência Interna; **UNI**: ProUni; **REI**: Reingresso; **SEL**: Processo Seletivo; **VES**: Vestibular; **OUT**: Outros;
MDA	Médio do Desempenho Acadêmico
PROF_TITUL	Titulação do professor, sendo E:Especialista; M:Mestre; D:Doutor;
TIP_DEF	Tipo de deficiência, com NI representando "Não Informada"

Fonte: Dos autores, 2019

5.1 Faixas de notas em cada GDA

A partir das médias dos alunos foi realizado um estudo para transformar a variável NOTA, com valores contínuos entre 0 e 10, em uma variável discreta chamada grupo de desempenho acadêmico (GDA).

Optou-se pela discretização por frequência (*equal frequency binning*), método de discretização não supervisionada que permite transformar uma variável com valores contínuos em um número pré-determinado de intervalos (faixas). A escolha se deu para garantir que os grupos de desempenho tivessem um número de indivíduos que viabilizasse a comparação entre os cursos. O limite superior e inferior de cada grupo é definido pelo algoritmo, sempre com o objetivo de manter todos os grupos com o mesmo número de registros. Caso não seja possível manter o mesmo número de registros em cada grupo, o algoritmo calcula os limites superior e inferior para que os grupos fiquem com o número de registros o mais semelhante possível (DOUGHERTY; KOHAVI; SAHAMI, 1995; GOLDSCHMIDT; PASSOS, 2005; KOTSIANTIS; KANELLOPOULOS, 2006).

Assim, foram definidas três faixas de notas, representando três grupos de desempenho acadêmico: Bom Desempenho Acadêmico (BDA), Médio Desempenho Acadêmico (MDA) e Baixo Desempenho Acadêmico (BADA).

A discretização por frequência foi realizada separadamente para os cursos de Direito e Engenharia Civil, o que possibilitou intervalos de nota com limites distintos para cada GDA, conforme mostram os dados da 3. A definição dos intervalos de cada GDA em função dos cursos possibilitou identificar que alunos cujas notas se destacam têm média acima de 8,8 no

curso de Direito, enquanto na Engenharia Civil, os alunos cujas notas se destacam têm média a partir de 8,3.

As médias foram analisadas isoladamente em cada disciplina em que o aluno foi aprovado. Desta forma, uma média é referente à uma ou mais avaliações realizadas por um Aluno/Disciplina/Ano/Semestre. Foram utilizadas apenas notas em que o aluno foi aprovado, pois essas são as notas que entram no histórico oficial do aluno. As notas de disciplinas onde o aluno foi reprovado não foram descartadas. Essas notas foram utilizadas para calcular o que foi chamado de "GDA com média suja", GDA_SUJA_PER4, conforme descrito na Tabela 2, o "GDA com média suja" foi útil para calcular o modelo de classificação do aluno, apresentado no modelo de classificação do curso de Direito, seção 5.6.

Tabela 3 - Faixas de notas em cada GDA

GDA	Direito	Engenharia Civil
BADA	0,0 – 7,7	0,0 – 7,2
MDA	7,8 – 8,7	7,3 – 8,2
BDA	8,8 – 10	8,3 – 10

Fonte: Dos autores, 2019

Nota-se na Tabela 3, que no curso de Engenharia Civil, a discretização por frequência colocou médias meio ponto menores no grupo BDA. Em Direito, o grupo BDA inicia com a média 8,8 e, em Engenharia Civil, médias acima de 8,3 já figuram no grupo BDA. Nas análises exploratórias realizadas notou-se, ainda, que esta diferença seria ainda maior, subindo de meio para um ponto, caso a discretização por frequência considerasse também as disciplinas em que os alunos reprovaram, o que ocorreu na definição do atributo GDA_SUJA_PER4.

A taxa de reprovações no curso de Engenharia Civil é maior que à do curso de Direito na IES analisada. Entre 2005/1 e 2014/1, no curso de Engenharia Civil, houve reprovação em 26,80% das disciplinas cursadas, o que representa 10,20 pontos percentuais maior que o percentual de reprovações em disciplinas cursadas no curso de Direito. Se fossem consideradas a discretização por frequência também sobre as disciplinas com alunos reprovados, o intervalo de definição dos GDA não deixaria nenhum aluno egresso de Engenharia Civil no grupo BADA.

5.2 Impacto do tipo de disciplina na nota

Na Tabela 4 são apresentados os resultados obtidos da aplicação do algoritmo "*Single Rule Induction (Single Attribute)*", o qual gera regras a partir de um único atributo de entrada. Este algoritmo é recomendado por Witten, Frank e Hall (2011).

Nota-se na Tabela 4, que tanto para o curso de Direito, quanto para o curso de Engenharia Civil, existe uma tendência de que as notas estejam no grupo BADA, exceto para as disciplinas de TCC (Trabalho de Conclusão de Curso) e Estágio.

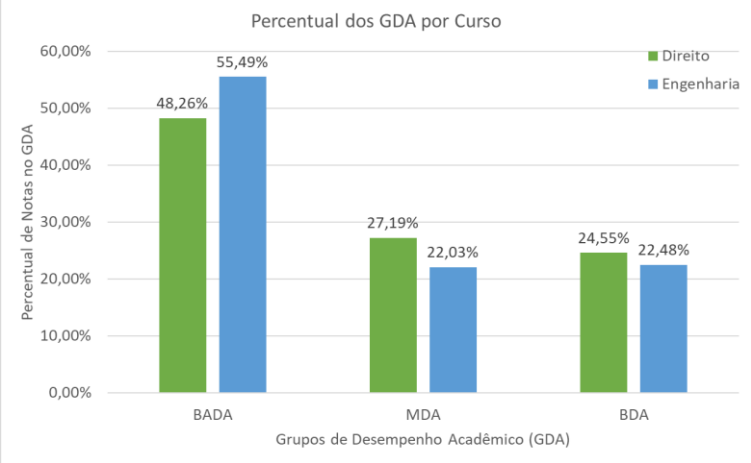
Tabela 4 - Regras do Algoritmo *Single Rule Induction (Single Attribute)*

Curso	Regra	Suporte	Confiança
Direito	if DIS_TIP = N then BADA	92,46%	50,90%
Direito	if DIS_TIP = E then BDA	5,31%	45,02%
Direito	if DIS_TIP = T then BDA	2,23%	61,48%
Direito	if DIS_TIP = P then MDA	0,00%	50,00%
Eng. Civil	if DIS_TIP = N then BADA	98,22%	56,45%
Eng. Civil	if DIS_TIP = T then BDA	0,89%	91,48%
Eng. Civil	if DIS_TIP = E then BDA	0,89%	89,56%
Eng. Civil	if DIS_TIP = P then BADA	0,01%	66,67%

Fonte: Dos autores, 2019

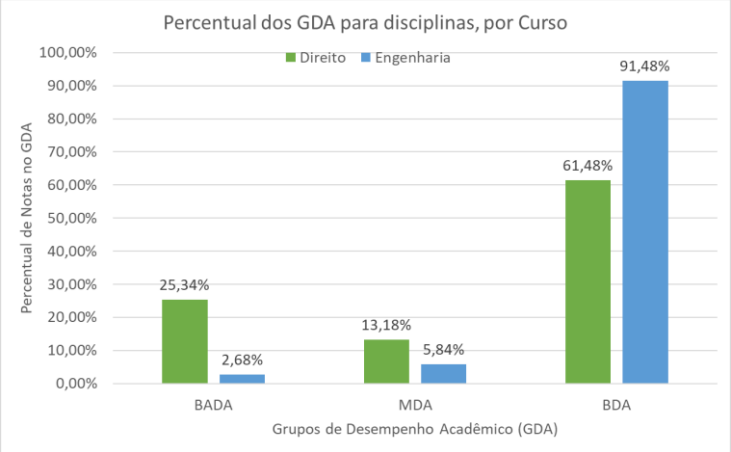
Nota-se ainda na 4, que as notas de disciplinas do tipo TCC (DIS_TIP=T) tendem a ficar no grupo BDA, sendo esta característica mais forte no curso de Engenharia Civil (confiança de 91,48%) do que no curso de Direito (confiança de 61,48%). A tendência de notas abaixo da média, na maioria das disciplinas cursadas em Direito e Engenharia Civil é notada na Figura 3, e o contraponto de notas de disciplinas de TCC acima da média é demonstrado na Figura 4.

Figura 3 - Percentual de notas nas disciplinas cursadas



Fonte: Dos autores, 2019

Figura 4 - Percentual de notas nas disciplinas de TCC



Fonte: Dos autores, 2019

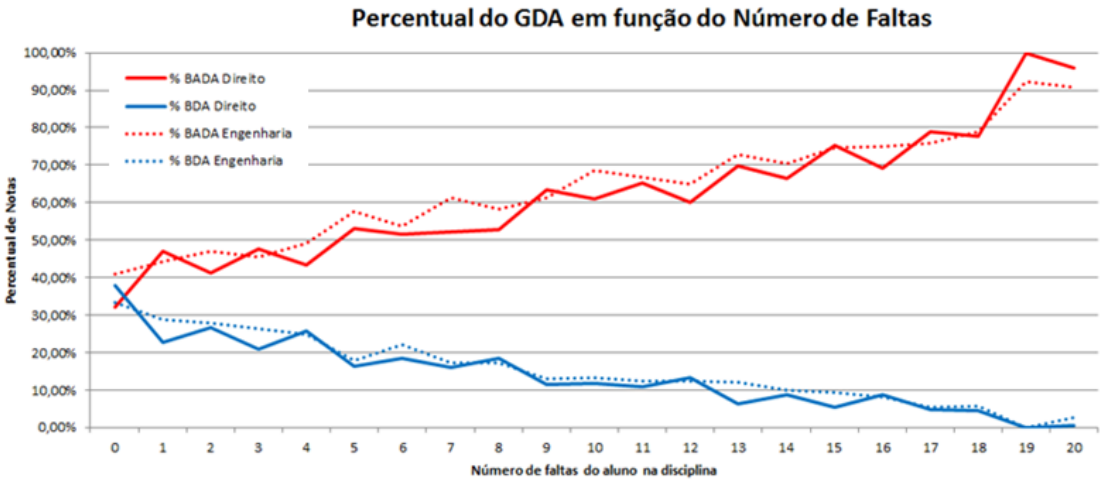
5.3 Impacto do número de faltas na nota

Pela análise das regras geradas pelos algoritmos "*Decision Tree*", "CHAID", "*Rule Induction*", "ID3" e "*Decision Tree* (*Weight-Based*)", foi possível observar que o atributo de número de faltas dos alunos (ALU_DIS_FALTAS) ocorria em regras com confiança acima de 60% nas análises do conjunto de dados referentes ao curso de Direito. Um exemplo de regra deste tipo gerada, com confiança de 67,59%, é: "if DIS_TIP = N and DIS_HOR_TEORICAS = range1 [-∞ - 75] and ALU_DIS_FALTAS = range1 [-∞ - 4.500] and DIS_OBR = N and DIS_HOR_PRATICAS = range1 [-∞ - 7.500] and PROF_TITUL = D and DIS_HOR_LAB = range1 [-∞ - 2.500] then BDA".

Neste exemplo, as duas primeiras variáveis indicam disciplinas normais (disciplinas que não são de Estágio, nem de TCC) (DIS_TIP=N) e carga horária teórica até 75 horas/aula (DIS_HOR_TEORICAS = range [-∞ - 75]). Já a ocorrência de até 4 faltas (ALU_DIS_FALTAS = range1 [-∞ - 4.500]) é menos frequente, e foi comum encontrar regras que indicavam acima de quatro faltas associadas ao BADA.

Foi construído o gráfico apresentado na Figura 5, no qual é possível notar que o aumento do número de faltas se correlaciona com uma diminuição no número de notas dentro do grupo BDA e aumento no grupo BADA.

Figura 5 - Percentual do GDA em função do número de faltas



Fonte: Dos autores, 2019

Uma regra do curso de Direito que chamou ainda mais atenção foi: "if ALU_DIS_FALTAS = range2 [4.500 - ∞] and DIS_TIP = T then BADA". Verificou-se, na Seção 5.2, que disciplinas de TCC (DIS_TIP=T) tendem a ter alunos com notas no grupo BDA, como apresentado na Figura 4. Entretanto, é possível inferir nesta regra, com 98,60% de confiança, que alunos das disciplinas de TCC com 5 ou mais faltas obtem notas pertencentes ao grupo BADA. Cabe comentar que a IES estudada tem uma disciplina de TCC, na qual os alunos têm aula de metodologia científica e constroem sua pré-proposta.

Dentre as disciplinas de TCC, o percentual de alunos que tiveram mais de quatro faltas é de apenas 5,57% no curso de Direito. Na Engenharia Civil 0,48% (apenas duas disciplinas) foram cursadas por alunos que tiveram mais de quatro faltas registradas. Apesar do pequeno percentual, quatro faltas ou mais coloca a nota do aluno, em praticamente 100% das vezes, no grupo de notas BADA.

5.4 Impacto do tipo de ingresso na nota

O algoritmo "*Decision Tree*", com critério acurácia, executado sobre o conjunto de dados do curso de Direito gerou, dentre outras regras, a regra "if GRU_INGR = REI and TIP_DEF = NI and DIS_TIP = N then BADA", com confiança de 61,04% e motivou uma análise mais detalhada do atributo grupo de ingresso (GRU_ING).

Conforme apresentado na Figura 3, o curso de Direito tem 48,26% das notas no grupo BADA. Entretanto, pela regra citada, disciplinas cursadas por alunos com ingresso por reingresso, sem deficiência física informada, cursando disciplinas do tipo normal, figurariam 61,04% no grupo BADA 12,78 pontos percentuais a mais. A Tabela 6 detalha o percentual de notas que figuram em cada GDA, em função da modalidade de ingresso do aluno.

Analisando a Tabela 5, curso de Direito, nota-se que o grupo BDA varia entre 23,04% e 26,04%, muito próximo da distribuição do curso de Direito apresentada na Figura 5, de

24,55%, exceto para os ingressos via Programa Universidade para Todos (ProUni) e Reingresso. Nos ingressos do tipo ProUni, existem uma tendência de notas do grupo BDA, com 8,76 pontos percentuais maior que na análise da distribuição natural do curso. Já no Reingresso, o percentual de notas do grupo BDA é de 18,33%, menor que a distribuição apresentada na Figura 3 em 6,22 pontos.

Ainda analisando os dados da Tabela 5, mas desta vez nas colunas BDA do curso de Engenharia Civil, nota-se que o percentual de alunos do ProUni também é maior que a distribuição apresentada na Figura 3, que é de 22,48%, gerando uma diferença de 10,27 pontos percentuais.

Tabela 5 - Percentual das notas em função do curso e ingresso

Tipo de Ingresso	Direito				Engenharia Civil		
	BDA	MDA	BADA		BDA	MDA	BADA
Outros (ex. diplomado ou egresso)	25,66%	28,16%	46,19%		28,03%	25,66%	46,30%
Processo Seletivo	26,04%	26,73%	47,23%		22,30%	21,08%	56,62%
ProUni	33,31%	30,96%	35,72%		32,75%	25,33%	41,92%
Reingresso	18,33%	24,08%	57,59%		14,31%	18,63%	67,05%
Transferência Externa	22,96%	27,55%	49,49%		19,55%	20,70%	59,74%
Transferência Interna	23,04%	27,37%	49,59%		18,66%	19,63%	61,71%

Fonte: Dos autores, 2019

Nota-se, também, que os alunos bolsistas do PROUNI tendem a ter um desempenho similar nos cursos de Direito e Engenharia Civil dentro do grupo BDA, com participação de 33,31% e 32,75%, respectivamente. Entretanto, quando comparados no mesmo curso, em relação aos alunos provenientes de Processo Seletivo (ingresso por análise currícular) e Vestibular, a diferença de desempenho (BDA) é maior na Engenharia Civil.

5.5 Modelo de Classificação do Egresso

Nesta seção, analisaram-se apenas dados de alunos que efetivamente tornaram-se egressos, com o objetivo de gerar um modelo de predição do GDA dos alunos, usando dados presentes nos quatro semestres iniciais do curso.

A Tabela 6 apresenta o total de egressos, incluindo a quantidade de alunos utilizados para treinamento e testes, do modelo de previsão da evasão.

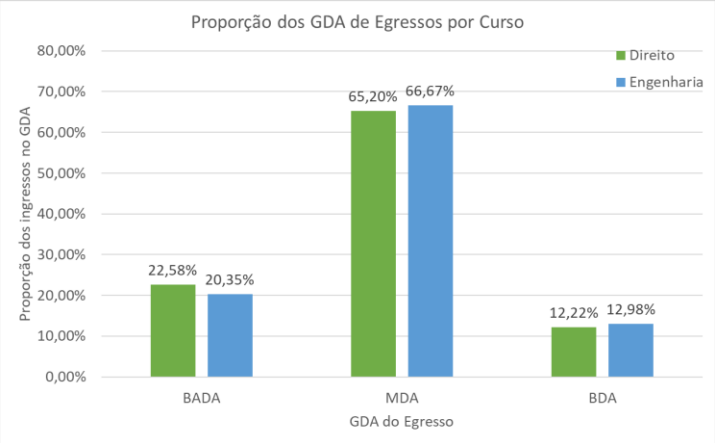
Tabela 6 - Quantidade de egressos por curso

Curso	Treinamento	Testes	Egressos
Direito	1034	259	1293
Eng. Civil	228	57	285

Fonte: Dos autores, 2019

Na Figura 6 é possível identificar a distribuição de egressos por GDA alcançado pela média de suas notas ao longo do curso, considerando apenas as aprovações, que são as notas que efetivamente entram em seu histórico escolar.

Figura 6 - Percentual dos GDA de Egressos por curso



Fonte: Dos autores, 2019

A Tabela 7 apresenta uma relação do GDA dos quatro primeiros semestres versus o GDA do Egresso, para o curso de Direito. A mesma relação para o curso de Engenharia Civil pode ser vista na Tabela 9.

É possível identificar na Tabela 7 que apenas 0,08% dos 36,79% de alunos do curso de Direito, os quais obtiveram uma média no grupo BADA nos semestres iniciais, figuraram como egressos de média BDA. Destes mesmos 36,79% que iniciaram no grupo BADA, 16,19% conseguiram concluir o curso com média dentro do grupo MDA. Dos 15,34% de egressos que começaram com média BDA, 5,27% não conseguiram manter a média e figurar como egressos no grupo BDA, mas a maior parte destes, 10,07% mantiveram notas que os fizeram egressos no grupo BDA. O grupo de desempenho acadêmico com maior percentual de redução é de alunos que iniciam como BADA (36,79%) e terminam como BADA (22,62%), ou seja, existe uma maior probabilidade de que um aluno que possui desempenho BADA nos semestres iniciais termine o curso com desempenho MDA.

Tabela 7 - Relação do percentual de GDA semestres iniciais versus egresso curso de Direito

		GDA egresso			Total
		BADA	MDA	BDA	
GDA Início	BADA	20,53	16,19	0,08	36,79
	MDA	2,01	43,76	2,09	47,87
	BDA	0,08	5,19	10,07	15,34
	Total	22,62	65,14	12,24	100%

Fonte: Dos autores, 2019

Na Tabela 8, é possível notar que, no curso de Engenharia Civil, não houve percentual representativo de alunos que tiveram os semestres iniciais com média no grupo BADA, os quais conseguiram se recuperar e figurar com média que os colocassem como egressos no grupo de BDA. Dos 44,21% de alunos que iniciaram com médias no grupo BADA, 23,86% conseguiram médias que os colocassem como egressos no grupo MDA. O grupo de desempenho acadêmico de maior percentual de redução, assim como no curso de Direito, é de alunos que iniciam como BADA (44,21%) e terminam como BADA (20,35%), ou seja, na Engenharia Civil também existe uma maior probabilidade de que um aluno BADA dos semestres iniciais termine o curso neste mesmo grupo.

Tabela 8 - Relação do percentual de GDA semestres iniciais versus egresso curso de Engenharia Civil

		GDA egresso			Total
		BADA	MDA	BDA	
GDA Início	BADA	20,35	23,86	0,00	44,21
	MDA	0,00	42,11	6,67	48,77
	BDA	0,00	0,70	6,32	7,02
	Total	20,35	66,67	12,98	100%

Fonte: Dos autores, 2019

Analisando a distribuição dos 7,02% que iniciaram o curso de Engenharia Civil no grupo BDA, não houve percentual significativo que indique que estes alunos se tornem egressos com médias no grupo BADA. Destes 7,02% que iniciaram no grupo BDA, 6,32% conseguiram concluir o curso com médias que os levaram ao grupo de egressos BDA.

5.6 Modelo de Classificação do curso de Direito

O modelo mais adequado, pelos critérios descritos na Seção 4.1, sobre o conjunto de dados dos alunos egressos de Direito foi gerado pelo algoritmo "*Decision Tree*", utilizando Coeficiente de Gini como critério de relevância. Este modelo obteve nível de aceitação Kappa "substancial" e Acurácia de 78,38%. Na Tabela 9 é apresentada a matriz de confusão do modelo.

Tabela 9 - Matriz de confusão do modelo para o curso de Direito

	Verdadeiro BADA	**Verdadeiro MDA**	**Verdadeiro BDA**	**Precision**
Classificado BADA	**46**	24	0	65,71%
Classificado MDA	12	**131**	6	87,92%
Classificado BDA	0	14	**26**	65,00%
Recall	79,31%	77,71%	81,25%	

Fonte: Dos autores, 2019

É possível notar, pelo indicador *Recall,* que o modelo segmenta corretamente 81,25% dos alunos que seriam BDA. Dos segmentados, a taxa de acerto, apresentada pelo indicador *Precision* é de 65,71%, bem acima da probabilidade dos alunos de Direito em figurar no grupo BDA, que é de 24,55%, conforme apresentado na Figura 3. Análise semelhante pode ser feita para o grupo BADA, que apresenta *Recall* de 79,31% e *Precision* de 65,71%.

Pelas regras do modelo de previsão do GDA do curso de Direito, listadas abaixo, é possível notar que apenas três atributos foram efetivamente utilizados: média dos quatro períodos iniciais (GDA_PER4), média suja (GDA_SUJA_PER4) e número de faltas nos quatro semestres iniciais (ALU_DIS_FALTAS4). Pelas regras (i) e (ii), é possível inferir que o aluno que, nos semestres iniciais tem média para pertencer ao grupo BADA tende a sair ou não deste grupo em função do número de faltas.

Nota-se que na regra (ii) 9,91% dos alunos (suporte da regra) vão para o grupo MDA caso tenham menos de 51 faltas. Esta regra é verdadeira para 69,64% dos alunos à que se aplica (confiança)

(i) if GDA_PER4 = BADA and GDA_SUJA_PER4 = BADA and ALU_DIS_FALTAS4 > 51 then BADA (Sup.23,81%, Conf.69,64%)

(ii) if GDA_PER4 = BADA and GDA_SUJA_PER4 = BADA and ALU_DIS_FALTAS4 ≤ 51 then MDA (Sup.9,91%, Conf.65,00%)

(iii) if GDA_PER4 = BDA and ALU_DIS_APR_QTD4 > 13.500 then BDA (Sup.31,00%, Conf.78,57%)

(iv) if GDA_PER4 = BDA and ALU_DIS_APR_QTD4 ≤ 13.500 and ALU_DIS_FALTAS4 > 3 then MDA (Sup.4,12%, Conf.80,00%)

(v) if GDA_PER4 = BDA and ALU_DIS_APR_QTD4 ≤ 13.500 and ALU_DIS_FALTAS4 ≤ 3 then BDA (Sup.3,35%, Conf.0,00%)

(vi) if GDA_PER4 = MDA then MDA (Sup.45,82%, Conf.94,69%)

Pelas regras (iii), (iv) e (v), nota-se que o aluno que inicia no grupo BDA tende a ficar neste grupo, apenas se ele obteve mais de 13 aprovações. A julgar pela matriz curricular do curso de Direito, a qual apresenta 23 disciplinas nos quatro primeiros períodos, tem-se uma evidência de que alunos do grupo BDA em disciplinas dos semestres iniciais, que cursaram o restante das disciplinas fora da UNIVALI (aproveitamento de créditos) tendem a não figurar como egressos pertencentes ao grupo BDA.

A regra (vi) indica que os alunos egressos analisados, que iniciaram o curso no grupo MDA, tendem a terminar no grupo MDA com confiança de 94,69%.

5.7 Modelo de Classificação do curso de Engenharia Civil

O modelo mais adequado, pelos critérios de descritos na Seção 4.1, sobre o conjunto de dados dos alunos egressos de Engenharia Civil, foi gerado pelo algoritmo "*Decision Tree*", utilizando Coeficiente de Gini como critério de relevância. Este modelo obteve nível de aceitação Kappa "justa" e Acurácia de 68,42%. Na Tabela 10 é apresentada a matriz de confusão do modelo.

Pelas regras (i) e (ii) do modelo de previsão do GDA do curso de Engenharia Civil, listadas abaixo, é possível observar que, iniciando com média no grupo de desempenho BADA, o algoritmo apresentou como condição para concluir o curso com média no grupo MDA, ter tido até 234 faltas, referindo-se aos quatro semestre iniciais do aluno. Acima disso, o modelo entende que o aluno será um egresso com média no grupo BADA. Ambas as regras têm uma confiança inferior à 60%, deixando-as muito próxima do acaso e não podendo ser consideradas uma evidência de padrão.

Tabela 10 - Matriz de confusão do modelo para o curso de Engenharia Civil

	Verdadeiro BADA	**Verdadeiro MDA**	**Verdadeiro BDA**	**Precision**
Classificado BADA	**4**	3	0	51,14%
Classificado MDA	8	**32**	4	72,73%
Classificado BDA	0	3	**3**	50%
Recall	33,33%	84,22%	42,86%	

Fonte: Dos autores, 2019

Regras do modelo de previsão do GDA do curso de Engenharia Civil:

(i) if GDA_PER4 = BADA and GDA_SUJA_PER4 = BADA and ALU_DIS_FALTAS4 > 234.500 then BADA (Sup.12,28%, Conf.57,14%)

(ii) if GDA_PER4 = BADA and GDA_SUJA_PER4 = BADA and

ALU_DIS_FALTAS4 ≤ 234.500 then MDA (Sup.13,58%, Conf.55,56%)

(iii) if GDA_PER4 = BDA then BDA (Sup.5,26%, Conf.66,67%)

(iv) if GDA_PER4 = MDA and GDA_SUJA_PER4 = BADA then MDA (Sup.10,53%, Conf.83,33%)

(v) if GDA_PER4 = MDA and GDA_SUJA_PER4 = MDA and ALU_DIS_FALTAS4 > 30.500 then MDA (Sup.53,09%, Conf.85,00%)

(vi) if GDA_PER4 = MDA and GDA_SUJA_PER4 = MDA and ALU_DIS_FALTAS4 ≤ 30.500 then BDA (Sup.5,26%, Conf.33,33%)

As regras (iv) e (v) indicam que os alunos do curso de Engenharia Civil, que têm uma média nos quatro semestres iniciais dentro do grupo MDA, tendem a concluir o curso no grupo MDA, caso a média suja dos quatro semestres iniciais seja BADA, ou se o aluno tiver tido acima de 30 faltas nos quatro primeiros semestres. As confianças de 83,33% apresentadas na regra (iv), e de 85% na regra (v), corroboram a tendência de se manter no grupo MDA, já apresentada na Tabela 9.

Iniciando no grupo MDA, o modelo considera que o aluno pode concluir o curso no grupo BDA, caso ele tenha menos de 30 faltas nos semestres iniciais, mas a confiança desta regra é de apenas 33,33%, não podendo ser consideradas uma evidência de padrão.

6 Considerações Finais

A identificação de possíveis perfis de desempenho acadêmico, logo nas primeiras fases de um curso de graduação, pode ser um conhecimento útil para gestores destas IES, no sentido de possibilitar a atuação em fatores que porventura estejam contribuindo para um baixo desempenho acadêmico, por exemplo, evitando que este baixo desempenho continue até o final do curso, ou que gere uma evasão indesejada.

Vários algoritmos foram aplicados em bases de dados de sistemas acadêmico e financeiro de alunos egressos dos cursos de Direito e do curso de Engenharia Civil, buscando-se identificar padrões de desempenho acadêmico e os fatores associados. Os cursos foram selecionados por serem de diferentes áreas e por terem maior número de egressos em suas áreas na IES estudada.

Analisando-se os resultados obtidos, ficou evidenciado um impacto direto do número de faltas no grupo de desempenho acadêmico em que a nota é enquadrada. Também se observou uma correlação do aumento no número de faltas com o aumento do percentual de notas no grupo BADA.

Observou-se, também, que alunos ingressantes pelo ProUni tendem a atingir um desempenho superior aos alunos de outros tipos de ingresso, tanto no curso de Direito como no curso de Engenharia Civil. Ainda com relação ao tipo de ingresso, alunos oriundos de Reingresso representaram o menor percentual no grupo BDA.

Alunos ingressantes via ProUni têm uma exigência de aprovação de, no mínimo, 75% das disciplinas cursadas (SESU, 2015), sob pena de perderem a bolsa, o que poderia explicar o melhor desempenho destes alunos. Por outro lado, parece compreensível que alunos que se

afastaram do curso tenham maior dificuldade que seus colegas, explicando assim o maior percentual de notas no grupo BADA existente entre alunos com ingresso via Reingresso.

Outros fatores como os dados provenientes das interações dos estudantes com o ambiente virtual de aprendizagem (AVA) e a locação de livros da biblioteca não foram identificados como significativos no desempenho acadêmico do aluno.

A análise dos dados permite afirmar, ainda, que alunos do curso de Direito, que iniciam com média no grupo BADA, têm uma tendência maior de que a média de suas notas, quando concluem o curso, estejam no grupo BADA. Já no curso de Engenharia Civil, alunos que iniciaram com média no grupo BADA tendem que suas notas migrem para o grupo MDA. Em ambos os cursos, alunos que iniciaram o curso com notas no grupo BDA tendem a ser egressos com média de notas dentro do grupo BDA.

Para desenvolver o modelo de predição do GDA do egresso, baseado na análise dos quatro semestres iniciais, foi possível perceber que os fatores que melhor descrevem o GDA do aluno ao final do curso foram: média, média suja (inclui as reprovações), número de faltas e o número de aprovações.

Nos modelos de classificação de ambos os cursos, os atributos média e faltas nos semestres inicias são mais relevantes para classificação do aluno. No curso de Direito o modelo também utiliza o número de aprovações nos semestres iniciais para classificação. A utilização deste atributo permite excluir alunos que ingressam no curso a partir de transferências de outras instituições, além de melhorar a acurácia da classificação.

Quando comparado aos trabalhos correlatos, esta pesquisa apresenta um número mais amplo de variáveis analisadas. As variáveis Tipo de Disciplina (tais como as disciplinas de TCC) e Tipo de Ingresso (Reingresso e ProUni) forneceram informações relevantes sobre o desempenho acadêmico do aluno. Entretanto, na construção do modelo de desempenho, apenas as variáveis notas e faltas foram consideradas nos modelos de classificação de egressos dos cursos de Direito e Engenharia Civil. Desta forma, não foi possível afirmar que usar um maior número de variáveis foi benéfico ao modelo de classificação.

O presente trabalho apresentou à área de informática na educação uma análise empírica, envolvendo a aplicação de técnicas de mineração de dados, buscando a identificação de perfis de desempenho discente de dois cursos de áreas diferentes de uma IES. Toda a análise dos resultados obtidos está delimitada pelas populações de alunos utilizadas, áreas e cursos envolvidos e IES considerada no estudo, não devendo-se fazer generalizações. Entretanto, estes resultados apontam caminhos bem interessantes para os gestores dos cursos estudados, os quais se traçados, podem trazer aprimoramentos para os processos de ensino e de aprendizagem.

Como trabalhos futuros os autores pretendem aplicar a técnica de Análise de Componentes Principais (ACP) no conjunto de dados original, buscando-se uma melhor seleção das variáveis para o modelo de predição. Além disto, buscar-se-á um melhor balanceamento do conjunto de dados e aplicação desta metodologia para os dados de outros cursos da instituição.

Referências

ASIF, R.; MERCERON, A.; ALI, S. A.; HAIDER, N. G. Analyzing Undergraduate Students' Performance Using Educational Data Mining. Computer & Education. Elsevier. Vol. 113.. Pp 177-194. 2017.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, 2011.

CARMONA, C. J. et al. Subgroup discovery in an e-learning usage study based on Moodle. In: 2011 7TH INTERNATIONAL CONFERENCE ON NEXT GENERATION WEB SERVICES PRACTICES 2011, Salamanca, Espanha. *Anais...* Salamanca, Espanha: IEEE, 2011.

CHAPMAN, P. et al. *CRISP-DM 1.0*: Step-by-step data mining guide, 2010. Disponível em: <https://www.the-modeling-agency.com/crisp-dm.pdf>. Acesso em: 1 jun. 2019

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and Unsupervised Discretization of Continuous Features. In: MACHINE LEARNING PROCEEDINGS 1995 1995, San Francisco, EUA. *Anais...* San Francisco, EUA: Elsevier, 1995.

DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A Systematic Review on Educational Data Mining. IEEE Open Access Journal. Vol 5, Pp 15991-16005. 2017.

GOLDSCHMIDT, R.; PASSOS, E. *Data mining*: um guia Prático. Rio de Janeiro: Elsevier, 2005.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. Estimativa de Desempenho Acadêmico de Estudantes: Análise da Aplicação de Técnicas de Mineração de Dados em Cursos a Distância. *Revista Brasileira de Informática na Educação*, v. 22, n. 01, p. 45–55, 2014.

GWET, K. L. The Kappa Coefficient: A Review. In: *Handbook of Inter-Rater Reliability*: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters. 3. ed ed. Gaithersburg, EUA: Advanced Analytics, LLC, 2012. p. 15–25.

HOE, A. C. K. et al. Analyzing students records to identify patterns of students' performance. In: 2013 INTERNATIONAL CONFERENCE ON RESEARCH AND INNOVATION IN INFORMATION SYSTEMS 2013, Kuala Lumpur, Malásia. *Anais...* Kuala Lumpur, Malásia: IEEE, 2013.

IBM. *IBM SPSS Modeler CRISP-DM Guide*, 2011. Disponível em: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf>. Acesso em: 10 mar. 2019.

KNIME: Open for Innovation. Open for Innovation. 2019. Disponível em: <https://www.knime.com/>. Acesso em: 23 nov. 2019.

KOTSIANTIS, S.; KANELLOPOULOS, D. Discretization Techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, v. 32, n. 1, p. 47–58, 2006.

MACFADYEN, L. P.; DAWSON, S. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, v. 54, n. 2, p. 588–599, 2010.

MARISCAL, G.; MARBÁN, Ó.; FERNÁNDEZ, C. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, v. 25, n. 02, p. 137–166, 2010.

MIKUT, R.; REISCHL, M. Data mining tools. *Wiley Interdisciplinary Reviews*: Data Mining and Knowledge Discovery, v. 1, n. 5, p. 431–443, 2011.

POWERS, D. M. W. The problem with kappa. In: PROCEEDINGS OF THE 13TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 2012, Avignon, França. *Anais...* Avignon, França: Association for Computational Linguistics, 2012.

RAKOTOMALALA, Ricco. TANAGRA : un logiciel gratuit pour l'enseignement et la recherche. in Actes de EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005.

RAPIDMINER. *Rapidminer*, 2019. Disponível em: <https://rapidminer.com/>. Acesso em: 21 nov. 2019.

RIGO, S. J. et al. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, v. 22, n. 01, p. 132–146, 2014.

SAMARANAYAKE, C. P.; CALDERA, H. A. A data mining solution on high failure rate in Physical Science stream at the university entrance examination. In: 2012 TENTH INTERNATIONAL CONFERENCE ON ICT AND KNOWLEDGE ENGINEERING 2012, Bangkok, Tailândia. *Anais...* Bangkok, Tailândia: IEEE, 2012.

SESU. *Manual do Bolsista ProUni*, Ministério da Educação, 2015. Disponível em: <http://prouniportal.mec.gov.br/images/pdf/manual_bolsista_prouni.pdf>. Acesso em: 10 mar. 2019.

SLATER, S.; JOKSIMOVIC, S.; KOVANOVIC, V.; BAKER, R. S.; GASEVIC, D. Tools for Educational Data Mining: A Review. Journal of Educational and Behavioral Statistics. Vol. 42. No 1. 2017.

TRANDAFILI, E. et al. Discovery and evaluation of student's profiles with machine learning. In: PROCEEDINGS OF THE FIFTH BALKAN CONFERENCE IN INFORMATICS 2012, Novi Sad, Servia. *Anais...* Novi Sad, Servia: ACM Press, 2012.

WEKA 3: Machine Learning Software in Java. Machine Learning Software in Java. 2019. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 23 nov. 2019.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data mining*: practical machine learning tools and techniques. 3. ed. Burlington, EUA: Morgan Kaufmann, 2011.

ZENG, X.; ZHENG, H. Genders Differentials in Computer Sciences Education: Analysis and Proposal. In: 2009 FIRST INTERNATIONAL WORKSHOP ON EDUCATION TECHNOLOGY AND COMPUTER SCIENCE 2009, Wuhan, Hubei, China. *Anais...* Wuhan, Hubei, China: IEEE, 2009.

ZHANG, Z. Study and analysis of data mining technology in college courses students failed. 2010 International Conference on Intelligent Computing and Integrated Systems. Anais...Guilin, China: IEEE, out. 2010.

Recebido em maio de 2019.

Aprovado para publicação em novembro de 2019.

Roberto Gonçalves Augusto Junior
Laboratório de Inteligência Aplicada, Universidade do Vale do Itajaí - Univali, betoaugusto@univali.br

Guilherme Augusto Rosa Carminati
Laboratório de Inteligência Aplicada, Universidade do Vale do Itajaí - Univali, guicarminati@edu.univali.br

Andre Luis Alice Raabe
Laboratório de Inovação Tecnológica na Educação, Mestrado em Computação Aplicada, Universidade do Vale do Itajaí – Univali. raabe@univali.br

Raimundo Celeste Ghizoni Teive
Laboratório de Inteligência Aplicada, Universidade do Vale do Itajaí – Univali, Mestrado em Computação Aplicada, rteive@univali.br