Mineração de Dados Educacionais Visando a Identificação da Evasão no Ensino Superior

Um Estudo de Caso dos Cursos de Engenharia de uma Universidade Comunitária

Guilherme A. R. Carminati
Universidade do Vale do Itajaí – UNIVALI
Santa Catarina, Brasil
guicarminati@edu.univali.br

Roberto G. Augusto Jr.
Universidade do Vale do Itajaí – UNIVALI
Santa Catarina, Brasil
betoaugusto@univali.br

Norberto Dallabrida
Universidade do Estado de Santa Catarina – UDESC
Santa Catarina, Brasil
norbertodallabrida@hotmail.com

Raimundo C. G. Teive
Universidade do Vale do Itajaí – UNIVALI
Santa Catarina, Brasil
rteive@univali.br

ABSTRACT

This paper tackles the problem of dropout of undergraduate students in a private university, by using Educational Data Mining (EDM) techniques. The EDM is an emerging area, concerned with developing methods for exploring the increasingly large-scale data that come from educational settings and using those methods to better understand students and the settings which they learn in. In this work, EDM is used to identify profiles of students who withdraw from their engineering courses. The considered dataset is composed of 53 attributes, involving financial and academic aspects of 2,925 engineering students. Preliminary results have identified some attributes that are related to the dropout in engineering courses, such as: the semester of the year (students are more prone to dropout in the first half of the year), attendance, grades (in this case median is more important than the mean value) and number of credits in the previous semester, and the current semester the student is enrolled (students bellow the 5th semester have a higher tendency to dropout).

KEYWORDS

Knowledge Discovery in Databases. Educational Data Mining. Dro-pout.

1 INTRODUÇÃO

Nas últimas décadas no Brasil, houve um forte crescimento no número de Instituições de Ensino Superior (IESs), matrículas e ingressos no ensino superior, além do número de concluintes. Segundo dados do INEP [7], de 1991 a 2017, o número de IESs cresceu 174%, o de concluintes, 407%, enquanto que o número de matrículas e ingressos cresceu 429% e 620%, respectivamente. Este crescimento foi devido principalmente à expansão do setor privado, que atualmente conta com 88% das IESs e 75% das matriculas. Um ponto importante a ser analisado é com respeito ao baixo percentual de concluintes, em relação ao número de matriculados e ingressantes. Para o ano de 2017, por exemplo, o número de concluintes foi de apenas 14,48% dos matriculados, e 37,19% dos ingressantes.

A evasão de alunos de IESs é um problema recorrente no Brasil, assim como em outros países [8]. Segundo [16] e [17], a taxa de evasão das IESs no Brasil vem se mantendo na faixa dos 22% ao ano, considerando IESs públicas e privadas. Entretanto, a evasão das IES se mostra um problema que não é comumente tratado de forma efetiva pelo governo e pelas próprias IESs.

Um estudante que evade o seu curso antes da conclusão pode ser um "desperdício social, acadêmico e econômico" [16]. Para IESs privadas, evasões significam uma importante perda no faturamento; enquanto que para IESs públicas, os recursos públicos investidos são desperdiçados em alunos que não terminam seus cursos [16].

Com o avanço da tecnologia e custos cada vez menores para a compra de dispositivos de armazenamento de dados, é comum para empresas e instituições possuírem grandes quantidades de dados acumuladas ao longo dos anos [21]. Esses dados comumente são usados apenas para geração de relatórios e estatísticas [21], [22].

Conforme [5], "A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas", sendo necessário, por exemplo, a utilização de algoritmos de mineração de dados (MD). Nesse contexto, a área de Descoberta de Conhecimento em Base de Dados (KDD, do inglês "knowledge-discovery in databases"), como o nome indica, proporciona a descoberta de conhecimento novo e útil, a partir da análise da base de dados, sendo que A MD é a principal etapa do processo de KDD [5], [6]. A área de Mineração de Dados Educacionais (EDM, do inglês "Educational Data Mining") busca gerar informações relevantes a partir de diferentes tipos de dados educacionais [20].

A IES estudada vem armazenando, há mais de vinte anos, dados dos alunos, envolvendo aspectos acadêmicos e financeiros. Porém, estes dados não são devidamente trabalhados para gerar conhecimento novo e útil para os gestores da instituição.

Tabela 1: Matrículas, Ingressos e Concluintes de IES

Ano	Número de			
	IES	Matrículas	Ingressos	Concluintes
1991	893	1.565.056	447.929	236.410
1996	922	1.868.529	539.975	260.224
2001	1.391	3.030.754	1.206.273	395.998
2006	2.270	4.676.646	1.753.068	736.829
2011	2.365	5.746.762	1.915.098	865.161
2016	2.407	6.554.283	2.142.463	938.732
2018	2.537	6.394.244	2.072.614	990.415

Fonte: INEP [7]

Neste sentido, este trabalho visa estudar o problema de evasão por meio da execução de técnicas de MD sobre uma base de dados educacional, proveniente dos sistemas acadêmico e financeiro da IES. Entende-se que os dados armazenados dos estudantes possam esconder informações relevantes que revelem alguns padrões apresentados por alunos que evadem, o que pode vir a contribuir na prevenção da evasão. Em especial, este estudo focará sobre a evasão nos cursos de engenharia da IES.

Este trabalho está organizado da seguinte maneira: na Seção 2, é apresentado um panorama do ensino superior no Brasil com foco na evasão; na Seção 3, é discutida a literatura relacionada; na Seção 4, os principais conceitos de MD são introduzidos; na Seção 5, são apresentados os resultados obtidos; e, por último, na Seção 6, são destacadas as conclusões e trabalhos futuros.

2 ENSINO SUPERIOR E EVASÃO

2.1 Evolução das IES e Matriculas

Segundo [7], no ano de 2018 o Brasil contava com um total de 2537 IES, sendo que destas, 299 eram públicas e 2238, privadas. É notável a superioridade numérica por parte das IES privadas, com 88,2% de todas as IES.

Neste trabalho, é adotado o conceito de ingressantes como a quantidade de estudantes que entraram para uma IES no ano considerado na pesquisa, enquanto que matrículas, se refere à quantidade de alunos ativos durante o devido ano.

Apesar do número de ingressantes ter tido um crescimento alto, o mesmo não pode se dizer a respeito do número de concluintes. Como é de se esperar, o número de concluintes também cresceu com o aumento de IES e ingressantes. Porém, como pode ser observado na Tabela 1, o número de concluintes continua bem aquém em relação ao número de ingressantes e matriculados.

Analisando a Tabela 1 fica evidenciado o forte crescimento que ocorreu em termos de número de IES, matrículas e ingressos no ensino superior, além do número de concluintes, considerando o período de 1991 a 2018. Neste período, o número de IES cresceu 184,1%, o número de matrículas e ingressos cresceu 308,6% e 362,7%, respectivamente. Houve também um crescimento expressivo no número de concluintes, chegando a 318,9%. Para o ano de 2018, o número de concluintes foi de apenas 15,49% dos matriculados, e 47,79% dos ingressantes.

2.2 A evasão no Ensino Superior

A evasão de alunos de instituições de ensino superior (IES) é um problema recorrente no Brasil assim como em outros países. Os autores [17], [9] afirmam que a evasão discente de cursos superiores é um "desperdício social, acadêmico e econômico". Além disso, matrículas desvinculadas geram uma importante perda de receitas, no caso de IES privadas, e, em IES públicas, o investimento feito em alunos que vêm a evadir não apresenta retorno.

De acordo com [15], os alunos evadidos encontram mais dificuldades no mercado de trabalho, o que afeta no crescimento da economia. Os autores acrescentam ainda que o impacto de uma evasão se dá também no âmbito pessoal, podendo causar sentimento de frustração e fracasso, desmotivando o evadido. Também, como colocado em [15], pessoas com ensino superior tendem a cometer menos crimes, são mais inclinadas a realização de serviços comunitários e utilizam menos serviços públicos, o que contribui para o desenvolvimento da sociedade.

Neste trabalho, adota-se o conceito de evasão conforme o apresentado em [13]:

> **evasão de curso**: quando o estudante desliga-se do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional; [...] **evasão da instituição**: quando o estudante desliga-se da instituição na qual está matriculado; [...] **evasão do sistema**: quando o estudante abandona de forma definitiva ou temporária o ensino superior.

O gráfico apresentado na Figura 1 mostra a taxa de evasão anual de IES públicas e privadas, no período de 1992 a 2018. O gráfico foi construído com base nos dados disponibilizados anualmente no site do INEP[1]. A taxa de evasão anual foi calculada de acordo com a equação apresentada em [17], considerando o caso EC.

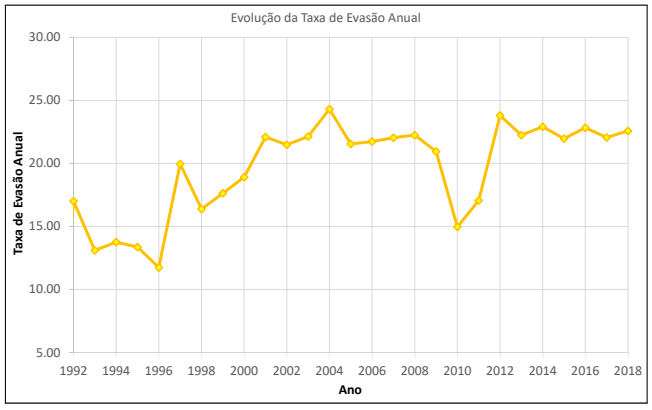

Figura 1: Evolução da Taxa de Evasão Anual no Brasil. Fonte: INEP [7].

Na Figura 1, é possível observar que a taxa de evasão anual, a partir de 1998, foi crescente até 2004, quando começou a cair,

[1]Disponível em: <http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior>. Acesso em: 05/06/2019.

até 2010. A partir de 2010, a taxa de evasão anual foi novamente crescente, se mantendo sempre acima de 20 %, chegando a 23,81%, em 2012. A grande diminuição da evasão no ano de 2010 pode estar vinculada com a queda no número ingressos no ano anterior.

2.3 Causas de Evasão

Este trabalho busca identificar possíveis causas de evasão da instituição (segundo definição do MEC [13]) com a aplicação de técnicas de mineração de dados. Porém, há diversos estudos existentes que visam responder à mesma questão com diferentes abordagens. Nesta seção são apresentados os principais motivos que levam à evasão de alunos do ensino superior encontrados na literatura.

Em [14], os autores realizaram uma pesquisa nos diversos trabalhos publicados entre os anos 2000 e 2011. Em sua pesquisa, os autores apontam alguns possíveis motivos para a evasão. Os principais fatores apontados pelos autores ficaram organizados da seguinte forma:

(a) Características individuais dos estudantes
- Questões financeiras relativas à vida pessoal ou familiar do estudante.
- Questões relacionadas à escolha do curso.
- Expectativas anteriores ao ingresso, nível de satisfação com o curso e com universidade.
- Aspectos interpessoais - dificuldades de relacionamento com colegas e docentes.
- Questões relacionadas aos desempenhos nas disciplinas.
- Questões familiares.

(b) Fatores internos às instituições
- Baixo prestígio social do curso ou da IES escolhida.
- Baixo nível de motivação e compromisso com o curso.

(c) Fatores externos às instituições
- Baixo prestígio social da profissão.
- Incompatibilidade entre os horários de estudos e trabalho.

Na mesma linha, o estudo realizado por [18] apontou também algumas das principais causas de evasão. Os autores observam que, segundo estudos pelo MEC, alunos bolsistas ProUni (Programa Universidade para Todos) possuem menores taxas de evasão em comparação a alunos sem bolsa [18]. Além dos aspectos acima, os autores apontam em seu estudo os seguintes fatores relevantes para a evasão:

(a) Características individuais dos estudantes
- Idade do aluno (quanto maior a idade, mais fácil do aluno evadir).
- Dificuldades na aprendizagem.

(b) Fatores internos às instituições
- Qualidade do curso escolhido,
- Localização da IES.
- Nota mínima para ingresso (quando a nota para ingresso é mais baixa, o índice de evasão é maior).
- Insatisfação com o projeto pedagógico, com os professores, com a infraestrutura e recursos disponíveis.

(c) Fatores externos às instituições
- Desemprego.

Os aspectos citados acima, corroboram com o que foi apresentado na pesquisa realizada pela Comissão Especial de Estudos Sobre Evasão em 1996 (composta pelo MEC, ANDIFES, ABRUEM e SESU). Além dos pontos apontados anteriormente, em [13], foram levantados algumas características complementares à evasão do ensino superior.

Em relação às características individuais dos estudantes: formação escolar anterior, escolha precoce da profissão e desmotivação. Com relação aos fatores internos, acrescenta-se: currículos desatualizados, alongados; rígida cadeia de pré-requisitos, além da falta de clareza sobre o próprio projeto pedagógico do curso; ausência ou ao pequeno número de programas institucionais para o estudante (como Iniciação Científica, Monitoria, etc.) e decorrentes de insuficiente estrutura de apoio ao ensino de graduação: laboratórios de ensino, equipamentos de informática, etc.

Como descrito nesta seção, a evasão discente no ensino superior é um tema complexo que pode ter múltiplos fatores causadores. Muitos desses fatores, especialmente àqueles vinculados às questões pessoais dos alunos e externos às IES, são difíceis de ser identificados a partir dos dados da IES. Entretanto, com auxílio de técnicas de mineração de dados, busca-se neste trabalho gerar algum conhecimento sobre perfis de alunos que evadem, considerando dados e informações de bases acadêmicas e financeiras da IES.

3 MAPEAMENTO SISTEMÁTICO DA LITERATURA - EVASÃO

3.1 Protocolo de Busca

A busca por artigos relacionados foi realizada por meio de uma única string de busca base. No processo de construção da *string* de busca, notou-se que a inclusão de termos completos ou muito comuns — como "KDD", "*educational data mining*", "*student dropout*", entre outros — se mostrou de pouca ajuda, seja por limitar a busca a contextos relacionados somente ao termo, ou por retornar um número muito grande de resultados com pouco relevância. Por esse motivo, optou-se por utilizar somente a parte principal de cada termo (somente "*mining*", por exemplo) em combinação com outras palavras chave específicas para a criação da *string* de busca.

A versão final da string de busca passou por diversão iterações entre ajustes da string e leitura de artigos identificados por esta. A combinação dos termos "*student*", "*dropout*" (e variações) e "*mining*" ou "*statistics*" mostrou-se abrangente o suficiente para o foco da pesquisa. Excluir termos como "*school*" e "online" também apresentou resultados significativos, filtrando muitos artigos que fugiam do foco buscado neste trabalho. Por fim, montou-se a seguinte *string* de busca base:

"student AND (dropout OR "drop out" OR attrition) AND (mining OR statistics) NOT school NOT online"

A pesquisa por trabalho relacionados limitou-se ao período de dez anos, sendo considerados trabalhos publicados de 2009 até a presente data de busca, 18 de outubro de 2019. Para a busca dos trabalhos foram utilizadas quatro bases de dados principais, são elas: IEEEXplore[2] (IEEE), ACM Digital Library[3] (ACM), ScienceDirect[4] (SD) e SpringerLink[5] (SL).

[2] http://ieeexplore.ieee.org
[3] https://dl.acm.org
[4] https://www.sciencedirect.com
[5] https://link.springer.com

Tabela 2: Seleção de Artigos

	Quantidade de artigos			
Base	**Inicial**	**Pré--seleção (título)**	**Separados para revisão (resumo)**	**Selecionados (texto)**
IEEE	61	26	16	2
ACM	13	4	2	1
SD	12	3	1	1
SL	18	3	1	0
Total	104	37	21	4

Como cada base possui um método de busca diferente, a string de busca base teve que ser adaptada para cada uma das ferramentas de busca. Deu-se preferência para realizar a busca somente no resumo (abstract) dos artigos, porém, quando essa opção não se fazia presente, a busca foi realizada somente pelo título. As ferramentas de busca que permitem a busca no abstract são IEEEXplore, ACM Digital Library e ScienceDirect. A Tabela 2 sumariza os resultados obtidos com a busca dos artigos nas bases de dados.

3.2 Trabalhos Correlatos

Em [10], os autores buscaram identificar e avaliar diversos fatores que ocorrem durante a trajetória dos acadêmicos da Universidade Federal do Rio de Janeiro (UFRJ). Os dados dos estudantes foram separados em três classes que representam a situação final: evasão, ainda cursando e egresso.

O *dataset* utilizado pelos autores inclui os ingressantes de todos os cursos da IES durante os anos de 2003 e 2004 e contém dados referentes a esses alunos 12 semestres, até o segundo semestre de 2010, totalizando 3808 instâncias.

Os autores utilizaram a ferramenta Weka e seus algoritmos de classificação para o treino dos modelos de conhecimento. A validação dos modelos foi realizada com a técnica de validação cruzada com 10 conjuntos (*10-fold cross-validation*). Os algoritmos utilizados e os atributos mais relevantes podem ser vistos na Tabela 9.

Os atributos presentes no *dataset* eram de dois tipos: relativos a cada semestre e relativos a todo o tempo de curso. Com os resultados obtidos, os autores destacaram algumas relações importantes com relação aos alunos que evadem:

- Eles reduziram a quantidade de disciplinas em cada semestre, até a evasão;
- Eles passaram em menos disciplinas conforme avançando nos semestres;
- Eles reprovaram por nota em pelo menos uma disciplina no primeiro semestre;
- Eles obtiveram média de disciplinas aprovadas inferior aos egressos no primeiro semestre.

Em [12], os autores aplicaram algoritmos de MD nos dados de alunos do curso de Graduação em Matemática da Universidade Federal Fluminense (UFF) para gerar um modelo para a previsão de alunos que possam vir a evadir.

O *dataset* utilizado pelos autores era composto de 1369 registros com dados referentes a notas do ENEM (Exame Nacional do Ensino Médio) do aluno, além de seu gênero, idade e cidade. O *dataset* contava ainda com o atributo "ação afirmativa", referente a IES, e o atributo situação do aluno, que podia assumir os valores "evasão" e "não evasão". Os autores utilizaram uma combinação da ferramenta estatística "R" e a plataforma de KDD "H2O" para a preparação e mineração dos dados. Para a mineração dos dados, os autores treinaram 321 modelos com diferentes valores de parâmetros, incluindo os algoritmos apresentados na Tabela 9. Os resultados obtidos não foram reportados na integra, sendo apresentados apenas as taxas de VP e VN obtidas com o que foi considerado o melhor modelo dos três métodos aplicados.

No estudo apresentado em [2], os autores utilizam a técnica estatística de Análise de Sobrevivência para identificar fatores que contribuem para a evasão de estudantes de uma universidade federal brasileira. O *dataset* utilizado pelos autores foi cedido pela Universidade Federal da Paraíba (UFPB) e era composto por 1202 registros de ingressantes do curso de Administração entre os anos 2004 a 2009.

Os atributos que faziam parte desse *dataset* incluem: ano de ingresso, ano de saída da IES, saída por graduação ou evasão, o tempo de currículo (8, 9 ou 10 semestres), o gênero do estudante, o estado civil na data de ingresso (casado ou não), a idade do estudante no ingresso, o tipo de escola que frequentou (pública ou não), a raça (considerado "branco" ou não), o coeficiente de notas (calculado pela média ponderada pela quantidade de créditos da disciplina), a informação sobre falha em alguma disciplina (considerado "falhou em pelo menos uma" ou não) e por último a informação sobre a evasão de uma disciplina (dada por "evadiu em pelo menos uma" ou não).

Os autores utilizaram os pacotes de ferramentas R e SPSS para auxiliar na aplicação da técnica de Análise de Sobrevivência, uma técnica utilizada em aplicações em que se deseja avaliar eventos que aconteçam com o decorrer do tempo.

Os resultados obtidos pelos autores mostram os impactos que os atributos possuem sobre o tempo de graduação e a evasão do curso. Foi evidenciado que currículos de nove semestres apresenta o menor risco de evasão. Em relação ao gênero, encontrou-se que mulheres possuem menor chance de evasão e levam menos tempo na graduação. O coeficiente calculado das notas mostrou que notas mais altas reduzem o risco de evasão e reduzem o tempo de graduação, enquanto que a reprovação em um disciplina, como esperado, exibiu um aumento no risco de evasão do curso. Os demais atributos não exibiram nenhuma influência na evasão ou no tempo de graduação.

Em [3] é apresentado um estudo sobre evasão nos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de informação da Universidade Federal de Sergipe. Os dados considerados são relativos aos anos de 2010 a 2018, considerando apenas registros de alunos do primeiro ao sexto período, totalizando 25690 estudantes. O objetivo da utilização de técnicas de mineração de dados foi identificar qual o semestre com maior risco de evasão. Como resultado obtido neste estudo, pode-se destacar que para os cursos de Engenharia de Computação e Sistemas de Informação, o semestre mais crítico para evasão é o quarto, enquanto que para o curso de Ciência da Computação, os alunos evadem com maior frequência no sexto período.

Na Tabela 9 deste artigo, é apresentada uma comparação destes trabalhos com os resultados obtidos neste estudo.

4 MINERAÇÃO DE DADOS

4.1 Metodologia Crisp-DM

Neste trabalho foi utilizada a metodologia CRISP-DM, como método para aplicação das técnicas de mineração de dados. A metodologia CRISP-DM foi desenvolvida na década de 1990 por um conjunto de organizações envolvidas em atividade de MD, com o objetivo de padronizar a execução do processo de KDD e torná-lo mais acessível e independente de ferramenta ou área de aplicação. Essa é considerada a metodologia padrão [1], [11].

A metodologia CRISP-DM é composta de seis fases que podem acontecer repetidas vezes durante a aplicação da metodologia, são elas: entendimento do negócio; entendimento dos dados; preparação dos dados; modelagem; avaliação e desenvolvimento [11].

4.2 Pré-processamento de Dados

A etapa de pré-processamento é um importante estágio da KDD. As tarefas de seleção, limpeza, codificação e enriquecimento dos dados compõem a etapa de pré-processamento, que antecede a MD (Mineração de Dados) [5], [6].

O *dataset* completo disponibilizado pela IES conta com 37.212 instâncias e 53 atributos de 10.960 alunos. Cada registro é referente a um aluno (atributo "id_Aluno", código anônimo de identificação do aluno) e sua situação atual no semestre registrado. Cada aluno pode aparecer mais de uma vez no *dataset*. Esse *dataset* abrange todos os alunos dos cursos de direito e de engenharias no período de 2016-1 a 2018-2, totalizando seis semestres.

Após a remoção de atributos irrelevantes e instâncias inconsistentes do *dataset*, restaram ainda 34.578 instâncias e 44 atributos de 10.397 alunos, sendo 2925 alunos de engenharia (EMCT). Foram criadas mais 150 variações do *dataset*, cada versão utilizada para encontrar as combinações que gerem modelos de classificação que melhor explicam os motivos de evasão para a amostra de população estudada. Porém, aqui será apresentada apenas a versão do *dataset* que apresentou melhores resultados nos experimentos realizados, o *dataset* "SU5".

Além da remoção de atributos, também foram criados os atributos "_Ant", derivados dos valores referentes aos semestres anteriores dos atributos. Um exemplo é o atributo "PerFaltas" que virou "PerFaltas_Ant" no semestre seguinte. Esses atributos foram utilizados no *dataset* intitulado "SU5", composto por apenas o último registro de cada aluno no *dataset*. A Tabela 3 apresenta os 22 atributos finais utilizados nesse *dataset*.

Tabela 3: Atributos do *dataset* SU5

Atributo	Descrição
Curso	Direito ou Engenharia (Civil, Computação, Elétrica, Mecânica, Produção ou Química)
Escola	ECJS ou EMCT (escola do direito ou das engenharias)
Ano	Ano do registro
Sem	Semestre do ano (1:Jan-Jun,2:Jul-Dez)
Genero	[F/M]
NecEspecial	[S/N]
Idade	Idade do aluno no semestre
CredAcademicosAnt	Quantidade de créditos que o aluno fez no semestre anterior
PercIntegralizado_Ant	Percentual completo do curso no semestre anterior
Periodo	[1-10] Período atual do aluno
Calouro_Ant	[S/N] Se o aluno era calouro no semestre anterior
Formando	[S/N] Se o aluno é formando no semestre atual
Formando_Ant	[S/N] Se o aluno era formando no semestre anterior
MediaDasMFAnt	Média de todas as MFs (Médias Finais) do aluno no semestre anterior
MedianaDasMFAnt	Mediana de todas as MFs (Médias Finais) do aluno no semestre anterior
PerFaltas_Ant	Percentual de faltas no semestre anterior
TitVenMais30_Ant	Quantidade de títulos vencidos a mais de 30 dias no semestre anterior
ResFinNeg_Ant	Quantidade de renegociações realizadas pelo responsável financeiro no semestre anterior
ResProtCred_Ant	[S/N] Se responsável financeiro já foi enviado aos serviços de proteção ao crédito
EgressoIES	[S/N] Se é um aluno que já concluiu algum outro curso de graduação na IES
EvasaoAntIES	[S/N] Se é um aluno que já evadiu algum outro curso de graduação na IES
Evasao	[S/N] Indica se o aluno se evadiu do curso (Atributo classe)

4.3 Algoritmos de MD

Neste trabalho foi utilizada a ferramenta de mineração de dados Weka[6] (*Waikato Environment for Knowledge Analysis*). A ferramenta possui algoritmos para as diversas etapas da KDD que foram realizadas neste trabalho.

Para a seleção de atributos, foi utilizados o algoritmo "PrincipalComponents", que realiza a análise dos componentes principais (ACP) [4]. Foi também utilizada a ferramenta Tanagra[7] para a execução do algoritmo "Spv Assoc Rule". Esse algoritmo busca regras de associação visando um atributo alvo utilizando uma abordagem APRIORI [19].

Para a geração e avaliação do modelo de classificação foram utilizados os algoritmos "J48", "NaiveBayes", "IBk" e "MultilayerPerceptron". O algoritmo "J48" é o equivalente ao algoritmo de árvore

[6]https://www.cs.waikato.ac.nz/ml/weka/
[7]http://eric.univ-lyon2.fr/ ricco/tanagra/en/tanagra.html

Tabela 4: Saída do algoritmo Svp Assoc Rule — alvo: Evasao=S

n.	Regra	Suporte	Confi.
1	"Genero=M" & "Formando=N" & "MedianaDasMFAnt=(-inf-0,125]"	14,22%	90,04%
2	"Genero=M" & "Formando=N" & "CredAcademicoAnt=(-inf-9,5]"	13,61%	90,46%
4	"CredAcademicoAnt=(-inf-9,5]" & "MedianaDasMFAnt=(-inf-0,125]"	17,37%	90,07%
8	"Idade=(-inf-18,5]" & "CredAcademicosAnt=(-inf-9,5]"	7,97%	95,89%
9	"Idade=(-inf-18,5]" & "MediaDasMFAnt=(-inf-0,05]"	7,97%	95,89%

de decisão "C4.5" na ferramenta Weka. O classificador *Naive Bayes* tem base nos princípios estatísticos do Teorema de Bayes. O algoritmo "IBk" aplica os conceitos do algoritmo *K-Nearest Neighbors* (K-NN, "K-Vizinhos Mais Próximos", em português) para classificar as instâncias no *dataset*. O "MultilayerPerceptron" é uma rede neural que é treinada utilizando *back propagation* [4], [6], [21].

A qualidade dos modelos gerados pelos algoritmos de classificação é medida pelos valores apresentados na matriz de confusão gerada pela ferramenta. As medidas de desempenho consideradas neste trabalho são o recall, a acurácia e a estatística Kappa. A validação dos modelos foi realizada com a técnica de *10-fold cross-validation*.

5 RESULTADOS

5.1 Regras de Associação

Buscando encontrar os atributos que melhor definem a evasão de um aluno, utilizou-se o algoritmo "Spv Assoc Rule" da ferramenta Tanagra. O algoritmo possui uma limitação de que todos os atributos devem ser nominais, por isso os atributos numéricos do *dataset* foram discretizados utilizando a técnica de *Equal Frequency Binning*. Os parâmetros suporte e confiança foram definidos para 0,05 e 0,9, respectivamente. As principais regras geradas foram compiladas na Tabela 4.

As regras na Tabela 4 mostram que alunos com médias ou medianas muito baixas no semestre anterior, vem a evadir no semestre atual. Também alunos que fizeram menos de 9 créditos no semestre anterior, combinado a outros fatores, possuem a tendência a evadir. Os atributos de gênero, idade e formando também apareceram em diversas regras.

5.2 Classificação

Neste caso, foi utilizado o algoritmo J48 com um valor de 20 para o parâmetro "minNumObj" (quantidade mínima para uma folha ser criada na árvore de decisão) para o *dataset* "SU5_EMCT". Posteriormente, foi também testado com o valor 200 para geração de uma árvore mais simplificada.

A Figura 2 ilustra a árvore de decisão gerada. Na figura, é possível observar que o atributo do semestre (primeira ou segunda parte do ano) foi o principal na construção do modelo de classificação para esse dataset. A validação dos modelos mediu valores consideravelmente altos, conforme Tabela 5.

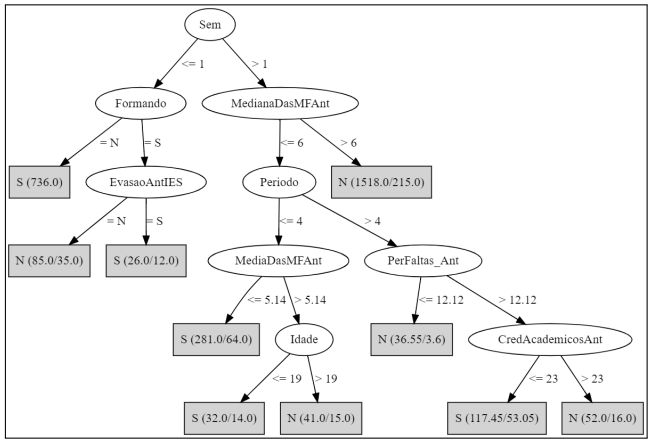

Figura 2: Árvore de Decisão — dataset SU5 EMCT.

Ainda utilizando o mesmo *dataset*, foi removido o atributo "Sem" e alterou-se o parâmetro "minNumObj" para 200, com o objetivo de simplificar a árvore de decisão. Os valores obtidos com a validação do modelo foram inferiores aos modelos anteriores, porém, ainda satisfatórios. A árvore de decisão gerada está ilustrada na Figura 3 e os resultados na Tabela 5.

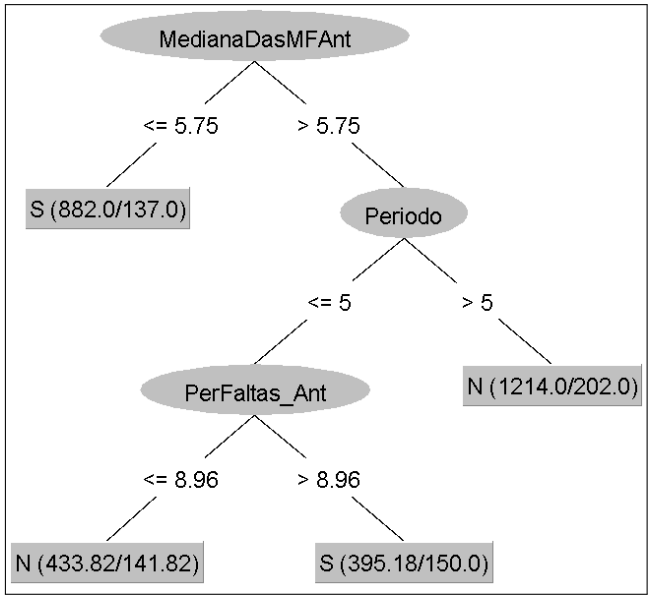

Figura 3: Árvore de Decisão reduzida — dataset SU5 EMCT sem atributo Semestre.

Com base na árvore da Figura 3, pode-se concluir que são evasores alunos que tiveram a mediana das MFs abaixo de 5,75 no

Tabela 5: Resultados dos *datasets* SU5_EMCT no J48

Dataset	SU5_EMCT	SU5_EMCT (reduzida)
Matriz de Confusão	1407 184 315 1019	1307 284 379 955
Acurácia	82,94%	77,33%
Kappa	0,6534	0,5405
Recall	76,40%	71,60%
Precisão	84,70%	77,10%

Tabela 6: Resultados Algoritmos de Classificação

Algor.	J48	NB	IBk	MLP
Matriz de Confusão	1407 184 315 1019	1450 141 340 994	1365 226 315 1019	1401 190 254 1080
Acurácia	82,94%	83,56%	81,50%	84,82%
Kappa	0,6534	0,6645	0,6252	0,6929
Recall	76,40%	74,50%	76,40%	81,00%
Precisão	84,70%	87,60%	81,80%	85,00%

semestre anterior. Também evadem aqueles que tiveram a mediana maior que 5,75, mas estão atualmente abaixo do 6º período e tiveram um percentual de faltas maior que 8,96%.

O mesmo *dataset* da Figura 2 também foi usado para gerar um modelo de classificação com os algoritmos Naive Bayes (NB), "IBk" e *Multilayer Perceptron* (MLP) da ferramenta Weka. Os resultados obtidos com esses modelos foram compilados na Tabela 6.

5.3 Análise Componentes Principais (ACP)

Visando melhorar os modelos de classificação obtidos, foi aplicado o algoritmo ACP ao *dataset*. A execução do algoritmo "PrincipalComponents" gerou um novo conjunto de atributos, organizados de forma a reduzir a dimensionalidade do *dataset* [6]. A Tabela 7 apresenta os primeiros três atributos gerados com execução do algoritmo. Pode-se observar que os principais atributos encontrados pelo algoritmo incluem a média e mediana das MF, assim como o período, a quantidade de créditos, entre outros.

Os três primeiro atributos desse novo *dataset* foram salvos para a criação de modelos de classificação. Os modelos gerados com a aplicação dos algoritmos de classificação sobre o novo *dataset* obtiveram melhores resultados, conforme relatado na Tabela **??**.

6 CONCLUSÕES

Nas últimas décadas observou-se no Brasil um forte crescimento de número de IES, impulsionado principalmente pelas IES privadas e comunitárias. Além disso, o número de matrículas e ingressos no ensino superior também cresceu fortemente, assim como o número de concluintes. Porém, o número de concluintes permanece muito abaixo do número de matrículas e ingressantes, sendo que a taxa de evasão se mantém acima de 22%; caracterizando um desperdício social, acadêmico e econômico.

A IES estudada tem armazenado dados dos estudantes nas últimas décadas, considerando aspectos acadêmicos e financeiros.

Tabela 7: Atributos ACP

Rank		Atributo	Peso
1º	0,8313	MediaDasMFAnt	-0,416
		MedianaDasMFAnt	-0,413
		Periodo	-0,413
		CredAcademicosAnt	-0,399
		Formando=S	-0,288
2º	0,7322	PercIntegralizado_Ant	-0,372
		Calouro_Ant=S	0,361
		ResFinNeg_Ant	-0,288
		TitVenMais30_Ant	-0,265
		ResProtCred_Ant=S	-0,256
3º	0,6501	ResProtCred_Ant=S	0,551
		TitVenMais30_Ant	0,466
		ResFinNeg_Ant	0,421
		PercIntegralizado_Ant	-0,265
		Calouro_Ant=S	0,229

Tabela 8: Resultados Algoritmos de Classificação — ACP

Algor.	J48	NB	IBk	MLP
Matriz de Confusão	1417 174 236 1098	1481 110 423 911	1364 227 228 1106	1402 189 232 1102
Acurácia	85,98%	81,78%	84,44%	85,61%
Kappa	0,7164	0,6256	0,6864	0,7091
Recall	82,30%	68,30%	82,90%	82,60%
Precisão	86,30%	89,20%	83,00%	82,60%

Entretanto, estes dados não são utilizados para gerar um conhecimento novo e útil para os gestores, particularmente em relação à evasão discente. Neste sentido, a utilização de técnicas de KDD, bem como Mineração de Dados Educacionais, mostra-se como ferramentas extremamente oportunas e viáveis para identificar possíveis fatores que podem levar os estudantes a evadirem o ensino superior.

Os resultados obtidos neste trabalho evidenciaram que alguns fatores podem apresentar maior influência na evasão de alunos, tais como: alunos são mais propensos a evadir no primeira metade do ano, mediana abaixo da nota mínima de aprovação no semestre anterior (isso implica que o aluno reprovou em pelo menos uma disciplina), alunos evadem mais até o quinto período. A frequência e a quantidade de créditos também apresentaram certo impacto.

Como trabalhos futuros, propõe-se a continuação deste estudo para outros cursos da IES, assim como, a partir dos conhecimentos obtidos, embasar a criação de uma estratégia de prevenção de evasão para a IES.

AGRADECIMENTOS

Os autores agradecem à Vice-reitoria de Graduação e Desenvolvimento Institucional, pela disponibilização dos dados acadêmicos e financeiros utilizados neste trabalho.

REFERÊNCIAS

[1] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. 2000. CRISP-DM 1.0: Step-by-step data mining guide. https://www.the-modeling-agency.com/crisp-dm.pdf

Tabela 9: Sumarização dos Resultados — Comparação com a Literatura

Autores	Principais Atributos	Validação	Métodos/Algoritmos Utilizados	Acurácia	Kappa	Recall
[10]	Nº de Disciplinas Nº de Disciplinas Aprovadas Nº de Disciplinas Reprovadas Média das Disciplinas Aprovadas (todos relativos a cada semestre)	10-fold cross- -validation	SimpleCart J48 (C4.5) Naive Bayes Support Vector Machine Multilayer Perceptron	83,90% 82,77% 79,59% 87,39% 85,34%	0,712 0,689 0,655 0,775 0,736	0,814 0,809 0,745 0,863 0,822
[12]	Notas no ENEM Idade Gênero Ação afirmativa	10-fold cross- -validation	Gradient Boosting Machine Distributed Random Forest Deep Learning	N/I N/I N/I	N/I N/I N/I	63,2% 55,4% 71,1%
[2]	Tempo de graduação Gênero do aluno Notas do aluno Trancamento ou reprovação	N/A	Análise de Sobrevivência	N/A	N/A	N/A
[3]	Período	10-fold cross- -validation	DecisionTree RandomForest SVM	66-72%	N/I	N/I
Este Trabalho	Mediana das MF Semestre Período Percentual de Faltas Quantidade de Créditos	10-fold cross- -validation	J48 (C4.5) Naive Bayes IBk (K-NN) Multilayer Perceptron Spv Assoc Rule ACP	82,94% 83,56% 81,50% 85,61% N/A N/A	0,6534 0,6645 0,6252 0,7091 N/A N/A	76,40% 74,50% 76,40% 82,60% N/A N/A

N/A: não se aplica. N/I: não informado.

[2] Francisco José da Costa, Marcelo de Souza Bispo, and Rita de Cássia de Faria Pereira. 2018. Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University. 53, 1 (2018), 74–85.

[3] Kelly J. de O. Santos, Angelo G. Menezes, Andre de Carvalho, and Carlos A. E. Montesco. 2019. Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, Vol. 2161-377X. 207–208.

[4] Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The WEKA Workbench. In *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (4 ed.). Morgan Kaufmann, 128. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

[5] Ronaldo Goldschmidt and Emmanuel Passos. 2005. *Data mining: um guia Prático.* Elsevier.

[6] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques* (3 ed.). Morgan Kaufmann.

[7] Inep. 2018. *Sinopse Estatística da Educação Superior 2018.* Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

[8] Edileusa Lima and Lucília Machado. 2014. A evasão discente nos cursos de licenciatura da Universidade Federal de Minas Gerais. 18, 2 (2014), 121–129. https://doi.org/10.4013/edu.2014.182.02

[9] Maria Beatriz de Carvalho Melo Lobo. 2011. Panorama da Evasão no Ensino Superior Brasileiro: Aspectos Gerais das Causas e Soluções. 25 (2011), 9–59. Cadernos nº 25.

[10] Laci Mary Barbosa Manhães, Sérgio Manuel Serra da Cruz, and Geraldo Zimbrão. 2014. The Impact of High Dropout Rates in a Large Public Brazilian University - A Quantitative Approach Using Educational Data Mining:. In *Proceedings of the 6th International Conference on Computer Supported Education.* SCITEPRESS - Science and and Technology Publications, 124–129.

[11] Gonzalo Mariscal, Óscar Marbán, and Covadonga Fernández. 2010-06. A survey of data mining and knowledge discovery process models and methodologies. 25, 2 (2010-06), 137–166. https://doi.org/10.1017/S0269888910000032

[12] Luiz Carlos Barbosa Martins, Rommel N. Carvalho, Ricardo S. Carvalho, Márcio C. Victorino, and Maristela Holanda. 2017-12. Early Prediction of College Attrition Using Data Mining. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).* IEEE, 1075–1078.

[13] MEC, SESu, ANDIFES, and ABRUEM. 1996. Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas. http://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf

[14] Marília Costa Morosini, Alam de Oliveira Casartelli, Ana Cristina Benso da Silva, Bettina Steren dos Santos, Rafael Eduardo Schmitt, and Rosana Maria Gessinger. 2012. A evasão na Educação Superior no Brasil: uma análise da produção de conhecimento nos periódicos Qualis entre 2000-2011. In *ICLABES. Primera Conferencia Latinoamericana sobre el Abandono en la Educación Superior.* E.U.I.T. de Telecomunicación, 65–73. https://doi.org/10923/8762 https://en.calameo.com/books/001171387403a4a0cab56.

[15] Nathália Prochnow Nagai and André Luís Janzkovski Cardoso. 2017. A Evasão Universitária: Uma Análise Além dos Números. 24, 1 (2017). https://doi.org/10.22410/issn.1983-036X.v24i1a2017.1271 Lajeado.

[16] Roberto Leal Lobo e Silva Filho. 2017-10-07. A Evasão No Ensino Superior Brasileiro - Novos Dados. https://educacao.estadao.com.br/blogs/roberto-lobo/497-2/

[17] Roberto Leal Lobo e Silva Filho, Maria Beatriz de Carvalho Melo Lobo, Paulo Roberto Motejunas, and Oscar Hipólito. 2007. A evasão no ensino superior brasileiro. 37, 132 (2007), 641–659. https://doi.org/10.1590/S0100-15742007000300007

[18] Clair Teresinha Souza, Caroline da Silva Petró, and Rosana Maria Gessinger. 2012. Um estudo sobre evasao no ensino superior do Brasil nos últimos dez anos. In *Conferencia Latinoamericana sobre el ABandono en la Educación Superior (CLABES).* 8.

[19] Tanagra. 2009. Predictive association rules. http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Predictive_AssocRules.pdf

[20] Evis Trandafili, Alban Allkoçi, Elinda Kajo, and Aleksandër Xhuvani. 2012. Discovery and evaluation of student's profiles with machine learning. In *Proceedings of the Fifth Balkan Conference in Informatics.* ACM Press, 174–179. https://doi.org/10.1145/2371316.2371350

[21] Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data mining: practical machine learning tools and techniques* (3 ed.). Morgan Kaufmann.

[22] Zhiyu Zhang. 2010. Study and analysis of data mining technology in college courses students failed. In *2010 International Conference on Intelligent Computing and Integrated Systems.* IEEE, 800–802. https://doi.org/10.1109/ICISS.2010.5657100