Pró-Reitoria Acadêmica
Escola de Educação, Tecnologia e Comunicação
Programa de Pós-Graduação Stricto Sensu em Gestão do Conhecimento e da Tecnologia da Informação

PREDIÇÃO DE EVASÃO NA EDUCAÇÃO A DISTÂNCIA COMO SUBSÍDIO À TOMADA DE DECISÃO

Autor: Weslley Rodrigues Sepúlvida
Orientador: Prof. Dr. Edilson Ferneda
Coorientador: Dr. Fernando Antonio Hello

Brasília - DF
2016

WESLLEY RODRIGUES SEPÚLVIDA

PREDIÇÃO DE EVASÃO NA EDUCAÇÃO A DISTÂNCIA COMO SUBSÍDIO À TOMADA DE DECISÃO

Dissertação apresentada ao Programa de Pós-Graduação *Stricto Sensu* em Gestão do Conhecimento e da Tecnologia da Informação da Universidade Católica de Brasília, como requisito parcial para obtenção do título de Mestre em Gestão do Conhecimento e da Tecnologia da Informação.

Orientador: Prof. Dr. Edilson Ferneda

Coorientador: Dr. Fernando Antonio Hello

Brasília-DF
2016

Dissertação de autoria de Weslley Rodrigues Sepúlvida, intitulada "PREDIÇÃO DE EVASÃO NA EDUCAÇÃO A DISTÂNCIA COMO SUBSÍDIO À TOMADA DE DECISÃO", apresentada como requisito para obtenção do grau de Mestre em Gestão do Conhecimento e da Tecnologia da Informação, defendida e aprovada, em 29 de Agosto de 2016, pela banca examinadora constituída por:

_____

Prof. Dr. Edilson Ferneda
Orientador
Gestão do Conhecimento e da Tecnologia da Informação – UCB

_____

Prof. Dr. Hércules Antônio do Prado
Examinador Interno
Gestão do Conhecimento e da Tecnologia da Informação – UCB

_____

Profª. Dra. Ana Paula Bernardi da Silva
Examinador Interno
Gestão do Conhecimento e da Tecnologia da Informação – UCB

_____

Prof. Dr. Fernando Antonio Hello
Examinador Externo
Empresa Brasileira de Pesquisa Agropecuária Brasília

_____

Profa. Dra. Maria de Fátima Ramos Brandão
Examinador Externo
Universidade de Brasília

Brasília
2016

Aos meus pais, irmãos, filhos e esposa, pela compreensão e sabedoria de me deixar partir, sem data e hora para retornar...

Senhor, fazei de mim um instrumento de vossa paz;
Onde houver ódio, que eu leve o amor;
Onde houver discórdia, que eu leve a união;
Onde houver dúvidas, que eu leve a fé;
Onde houver erros, que eu leve a verdade;
Onde houver ofensa, que eu leve o perdão;
Onde houver desespero, que eu leve a esperança;
Onde houver tristeza, que eu leve a alegria;
Onde houver trevas, que eu leve a luz.

Ó Mestre, fazei com que eu procure mais consolar,
que ser consolado;
Compreender, que ser compreendido;
Amar, que ser amado;
Pois é dando que se recebe;
É perdoando, que se é perdoado;
E é morrendo que se vive para a vida eterna.

Oração de São Francisco de Assis

RESUMO

Referência: SEPÚLVIDA, Weslley. **PREDIÇÃO DE EVASÃO NA EDUCAÇÃO A DISTÂNCIA COMO SUBSÍDIO À TOMADA DE DECISÃO**. 2016. P 113. Dissertação do Mestrado em Gestão do Conhecimento e da Tecnologia da Informação – Universidade Católica de Brasília (UCB), Brasília – DF, 2016.

A educação a distância (EAD) tem crescido nos últimos anos. Várias instituições de ensino têm ofertado cursos que vão desde aperfeiçoamentos e capacitações internas até cursos de extensão, graduação e pós-graduação. Com o crescimento da oferta de cursos e o aumento significativo dos estudantes, as instituições educacionais se colocam frente a novos desafios, entre eles o combate das altas taxas de evasão, comum em cursos na modalidade EAD. Nesse sentido, a Mineração de Dados é uma das abordagens que vem sendo explorada para o desenvolvimento de métodos preditivos de evasão. O presente trabalho propõe uma análise da evasão no contexto da EAD de uma tradicional universidade do Centro Oeste. O estudo busca identificar comportamentos dos estudantes que abandonaram cursos de graduação nessa modalidade, de maneira a fornecer subsídios preditivos para os atores envolvidos no processo de ensino-aprendizagem de forma a apoiar a tomada de decisões preventivas a respeito da evasão. Parte-se do pressuposto de que intervenções junto aos estudantes propensos a evadir podem acarretar mudança no comportamento via Ambiente Virtual de Aprendizagem (AVA) que contribuem para minimizar a evasão. Este trabalho é dividido em quatro partes: *(i)* uma revisão da literatura, para embasamento teórico, sobre evasão no âmbito da EAD, ambientes virtuais de aprendizagem (AVA), *knowledge Discovery in Database* (KDD) e *Education Data Mining*; *(ii)* análise da base de dados do AVA utilizado pela instituição aplicando técnicas de KDD, para identificação do comportamento de abandono do curso; *(iii)* geração e validação de um modelo para identificação preditiva de estudantes propensos a evadir; (*iv*) Ações Gerenciais propostas para mitigação do problema. Os resultados mostram que ao aplicar o KDD nos dados variantes no tempo, com 30 dias após o início das aulas, é possível predizer evasão com precisão significativa. A partir dos resultados obtidos, gerou-se um modelo de predição de evasão bem como um modelo de tomada de decisão e ações de combate à evasão.

Palavras-chave: Educação a Distância, KDD, Mineração de Dados, EDM, Mineração de Dados Educacionais, Predição de Evasão.

ABSTRACT

Distance Education has grown along the years. Several educational institutions have been offering courses of internal improvement and qualification, as well as extension, undergraduate and postgraduate courses. The increase in the number of courses and the significant raise in the number of students result in new challenges to the educational institutions. The high dropout rates, common in Distance Education courses, is one of the crucial problems the institutions have to deal with. In this context, Data Mining is one of the main approaches for the development of predictive methods of evasion. The present paper aims an analysis of the Distance Education evasion in a traditional Midwest Brazilian University. The study intends to identify the behavior of the students who have dropped out undergraduate courses, in order to provide subsidies for the subjects involved in the teaching-learning process. As a preventive solution to the evasion issue, it is understood that proper communication with the students who are about to evade can lead to changes that contribute to minimize the problem. This study is presented in four parts: (*i*) literature review based on theoretical framework on Distance Education evasion, Virtual Learning Environments, Knowledge Discovery in Database (KDD) and Education Data Mining; (*ii*) an analysis of the Virtual Learning Environment institution database applying KDD techniques to identify the course abandonment behavior; (*iii*) development and validation of a model for predictive identification of students prone to evade; (*iv*) management actions to mitigate the problem. The results show that, when applying KDD to the variant data in time, 30 days after the beginning of the classes, it is possible to significantly predict evasion. From the results, an evasion prediction model was developed, as well as an evasion combat model.

Keywords: Distance Education, KDD, Data Mining, EDM, Educational Data Mining, Dropout Prediction.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS

ABED	Associação Brasileira de Educação a Distância
AVA	Ambiente Virtual de Aprendizagem
BD	Banco de Dados
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CRISP-DM	*CRoss Industry Standard Process for Data Mining*
DM	*Data Mining*
EAD	Educação a Distância
EDM	*Education Data Mining*
FSF	*Free Software Fundation*
GPL	*General Public Licence*
IA	Inteligência Artificial
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	*Knowledge Discovery in Databases*
MEC	Ministério da Educação e Cultura
PROUNI	Programa Universidade para Todos
SGA	Sistema de Gestão da Aprendizagem
TI	Tecnologia da Informação
UAB	Universidade Aberta do Brasil
WEKA	*Waikato Environment for Knowledge Analysis*

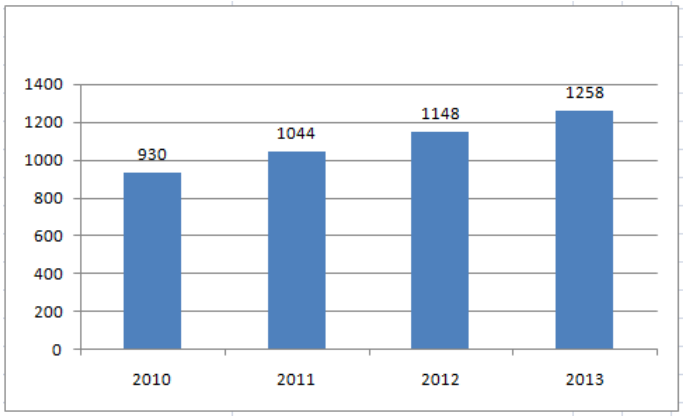
SUMÁRIO

1. INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Embora a educação a distância corresponda a uma pequena parcela do total de cursos de graduação, cada vez mais é parte do cotidiano brasileiro. Atingindo o maior percentual de crescimento, de 35,3% considerando o período de 2010 a 2013 (INEP, 2013). As novas mídias e as novas formas de interação entre estudante e docente fizeram com que os cursos ofertados nessa modalidade saltassem para 1.258 no ano de 2013, conforme mostrado no Gráfico 1.

Gráfico 1 – Crescimento da oferta de cursos na modalidade a distância



Fonte: adaptado de INEP (2013)

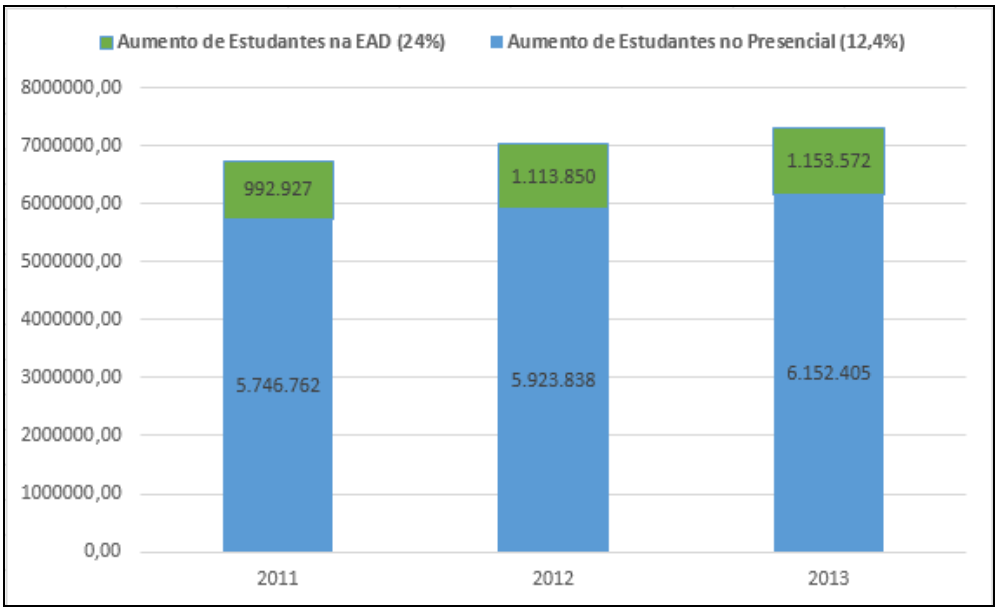
O Censo EAD, da Associação Brasileira de Educação a Distância (ABED, 2013) revela que existe uma expectativa de crescimento do número de matrículas. Na Tabela 1 mostra-se que, de 2013 para 2014, 64% das instituições participantes do CENSO tinham a expectativa de crescimento do número de estudantes, enquanto apenas 14% esperavam uma diminuição desse quantitativo. O índice era ainda mais promissor de 2014 para 2015, onde 82% das instituições tinham uma expectativa de aumento do número de estudantes, enquanto apenas 5% esperavam uma diminuição. Essa expectativa baseia-se, entre outras coisas, nos índices de crescimento do quantitativo de estudantes na modalidade EAD apresentados pelo Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP, 2013), que aponta crescimento das modalidades presencial e EAD, no período de 2011 a 2013, conforme mostrado no Gráfico 2.

Tabela 1 – Expectativa de crescimento do número de estudantes na EAD

Situação	% Média											
	% Média Totalmente a Distância		Semipresencial		Disciplinas EAD		Livre não Corporativos		Livres Corporativos		Total	
	2013/ 2014	2014/ 2015	2013/ 2014	2014/ 2015	2013/ 2014	2014/ 2015	2013/ 2014	2014/ 2015	2013/ 2014	2014/ 2015	2013/ 2014	2014/ 2015
Aumento	39 (59%)	55 (83%)	20 (64%)	27 (87%)	17 (58%)	23 (82%)	53 (70%)	59 (78%)	34 (66%)	43 (82%)	163 (64%)	207 (82%)
Diminuição	14 (21%)	4 (6%)	5 (16%)	2 (6%)	5 (17%)	1 (3%)	5 (6%)	3 (4%)	7 (14%)	3 (6%)	36 (14%)	13 (5%)
Manutenção	13 (20%)	7 (11%)	6 (19%)	2 (6%)	7 (24%)	4 (14%)	17 (22%)	13 (17%)	10 (19%)	6 (11%)	53 (21%)	32 (12%)
Total	66	66	31	31	29	28	75	75	51	51	252	252

Fonte: ABED (2013)

Gráfico 2 – Crescimento dos estudantes matriculados na modalidade a distância e presencial



Fonte: Adaptado de INEP (2013)

Comparando os números absolutos, a quantidade de matrículas de cursos de graduação presencial no período de 2010 a 2013, foi 3 vezes maior, somando 703.285 (setecentos e três mil duzentos e oitenta e cinto) matrículas, enquanto a educação a distância teve um aumento de 223.393 (duzentos e vinte e três mil trezentos e noventa e três) matrículas no mesmo período. Em termos percentuais, a maior elevação ocorreu nos cursos a distância, com crescimento registrado de 24% de 2010 a 2013, com uma média de crescimento de aproximadamente 8% ao ano. Já as matrículas de cursos presenciais aumentaram 12,4% no mesmo período, apresentando uma média de crescimento de aproximadamente 4% ao ano. Em valores absolutos, o

número de estudantes da modalidade a distância está longe de alcançar o número de estudantes da modalidade presencial. Em 2013, enquanto na modalidade presencial existiam 5.152.405 estudantes, na modalidade a distância existia 1.153.572.

Apesar de toda expectativa e crescimento do número de estudantes matriculados nos cursos na modalidade a distância, o Censo EaD.br disponível em ABED (2013) revela que a evasão[1] de estudantes é apontada pelas instituições pesquisadas como o maior obstáculo enfrentado na gestão de cursos de EAD. Os índices de evasão variam de acordo com o tipo de EAD praticadas (cursos de graduação *on-line* autorizados pelo MEC, cursos livres não corporativos, cursos livres corporativos e disciplinas virtuais de cursos presenciais).

Entre os tipos de EAD, os dados Censo EaD.br, ABED (2013) mostra que os menores índices de evasão são das disciplinas EAD e os maiores, em , 2010, 2011, 2013, apontam para os cursos Livres não corporativos, ou seja, cursos em que não há compromisso com a instituição, como pode ser observado na Tabela 2.

Tabela 2 – Índices de evasão registrados por ano/tipo de curso

Tipo de cursos	2010	2011	2012	2013
Autorizados pelo MEC	18,60%	20,50%	11,74%	16,94%
Livres não corporativos	22,30%	23,60%	10,05%	17,08%
Livres corporativos	7,60%	20,00%	3,00%	14,62%
Disciplinas EAD	–	17,60%	3,10%	10,49%

Fonte: ABED (2013)

Possivelmente os índices variam por se tratarem de públicos diferentes. No primeiro tipo de curso, autorizados pelo MEC, os estudantes estudam inteiramente à distância e de certa forma são expostos a muito mais estímulos concorrenciais em suas residências, no trabalho ou em qualquer outro ambiente no qual escolham estudar. No segundo tipo, os Livres não Corporativos, onde ocorre o maior registro de evasão a maioria dos cursos não têm interação, são gratuitos e o estudante não tem compromisso financeiro ou institucional de concluí-lo. O terceiro tipo de curso, os Livres Corporativos, os estudantes geralmente são funcionários da empresa ofertante do curso e o fazem para capacitação, treinamento, desenvolvimento e/ou aperfeiçoamento de novas técnicas de gestão e/ou trabalho de relevância administrativa/operacional para a organização. Já o quarto tipo de curso, Disciplinas EAD, são disciplinas de um curso presencial que são ministradas *on-line.* Elas fazem parte da grade curricular dos estudantes presenciais que, pelo contato do estudante com a instituição, apresenta um menor

[1] Santos, et al (2008) que definem a evasão como sendo a desistência definitiva do estudante em qualquer etapa do curso. Portanto, consideraremos o abandono e cancelamento no decorrer do curso para registro do índice de evasão.

índice de evasão.

A evasão na educação a distância é um fenômeno que vem sendo objeto de pesquisa há muito tempo e diferentes índices são apresentados em trabalhos correlatos a esse no decorrer deste mesmo período.

Segundo Bourdages (1996 apud LANGUARDIA; PORTELA, 2009) as cifras sobre evasão situam-se no intervalo de 20% a 85% e variam conforme a definição adotada, a população estudada, a estratégia educacional, por exemplo: semi-presencial, totalmente *on-line*; o desenho, tipo e duração do curso.

Conforme Maia et al. (2004), a evasão é maior em instituições privadas, chegando a 30%, entretanto, os autores ressaltam que esse fator não está relacionado ao custo do curso. Os autores apontam que a evasão na EAD está relacionada à forma de interação com o uso das tecnologias de informação e comunicação (e-mail, chat, fórum e videoconferência) e a forma de ensino (semipresencial ou totalmente a distância). Ambos fatores estão diretamente relacionados ao sentido de pertencimento à comunidade acadêmica. Com a utilização efetiva das ferramentas de informação comunicação e bem como encontros presenciais regulares, motivam os estudantes a aprender e interagir devido ao sentimento de inclusão acadêmico.

Dados do Anuário Brasileiro Estatístico de Educação Aberta e a Distância (Instituto Monitor, 2005) mostram que entre as instituições pesquisadas, 55% têm taxas de evasão em torno de 30%, e dois terços registram índices de evasão superiores a 70%.

A evasão, portanto, é um fenômeno complexo e seus índices podem variar de acordo com tipo de curso: totalmente a *on-line*[2], semi-presencial[3], cursos livres[4], cursos corporativos[5], disciplinas EAD[6], de instituições públicas ou privadas e ainda podem estar relacionadas à quantidade de vagas/candidatos por curso.

Conforme Cunha e Morosini (2014):

> Alguns autores procuraram estudar a evasão no ensino superior brasileiro a partir de dados disponíveis no INEP – Instituto Nacional de Estudos e Pesquisas Anísio Teixeira e relativamente aos anos de 2000 a 2005. Concluíram que os índices de evasão variam significativamente segundo a categoria administrativa[7], sendo que no período

[2] Cursos em que os encontros presencial são apenas para realização de avaliações presenciais em atendimento às exigências legais do Ministério da Educação e Cultura (MEC).

[3] Cursos que são ministrados parte *on-line* e parte presencial. Geralmente vinculados aos cursos que requerem práticas laboratoriais indispensáveis para formação completa dos estudantes.

[4] Cursos disponibilizados por instituições de ensino ou corporações de forma gratuita.

[5] Cursos ofertados por instituições ou corporações para seus funcionários.

[6] Disciplinas da grade curricular de um curso presencial realizada de forma *on-line*.

[7] Termo utilizando pelo INEP – Instituto Nacional de Estudos e Pesquisas Anísio Teixeira ao se referir a instituições publicas ou privadas.

> estudado a evasão média nas instituições públicas foi de 12%, enquanto que, nas instituições particulares, essa taxa chegou a 26%. Quando procuraram correlacionar os índices de evasão com a relação candidato/vaga em alguns cursos, concluíram que quanto maior a densidade candidato/vaga nos processos de ingresso na Educação Superior, menor são os índices de evasão/abandono.

Atualmente presencia-se um aumento significativo do número de vagas no ensino superior devido aos incentivos governamentais, como ProUni, UAB, FIES e Reuni, e à concorrência das instituições de ensino superior privadas.

Por outro lado, as universidades privadas têm tornado seus preços cada vez mais atrativos para fomentar o ingresso dos estudantes no nível superior, mas todo esse esforço pode fracassar parcialmente caso não exista uma política de combate à evasão.

No sentido de compreender esse fenômeno, diversas pesquisas vêm sendo realizadas a fim de mitigar o fenômeno da evasão em cursos na modalidade EAD. Uma das linhas de pesquisa utiliza Técnicas de Extração do Conhecimento em Base de Dados (do Inlgês *knowledge Discovery in Databases* – KDD[8]) procurando extrair da base de dados conhecimento sobre o comportamento do estudante no Ambiente Virtual de Aprendizagem (AVA). Com isso, espera-se identificar o comportamento dos discentes propensos a evadir, para que se possa tomar medidas preventivas para mitigação do problema.

1.2 REVISÃO DA LITERATURA

Para essa dissertação foi realizado um levantamento das publicações sobre evasão em educação superior na modalidade EAD até o presente momento. Utilizou-se para pesquisa os motores de busca do portal de periódicos da CAPES[9], Google Acadêmico[10] e Scielo[11] com as combinações de expressões de busca mostradas no Quadro 1. Os resultados obtidos foram divididos nos Gráfico 3 e 4, que representam, respectivamente, os resultados dos motores de busca no Portal da CAPES e Scielo e do Google Acadêmico. A opção pela separação dos gráficos deu-se para melhorar a visualização e identificação dos resultados devido ao grande número de resultados do Google Acadêmico, já que esse mecanismo de busca e reúne resumos, artigos, dissertações e teses de diversas fontes ao mesmo tempo.

---

[8] KDD – Processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados. (FAYYAD ET AL. 1996)

[9] http://www.periodicos.capes.gov.br

[10] http://scholar.google.com

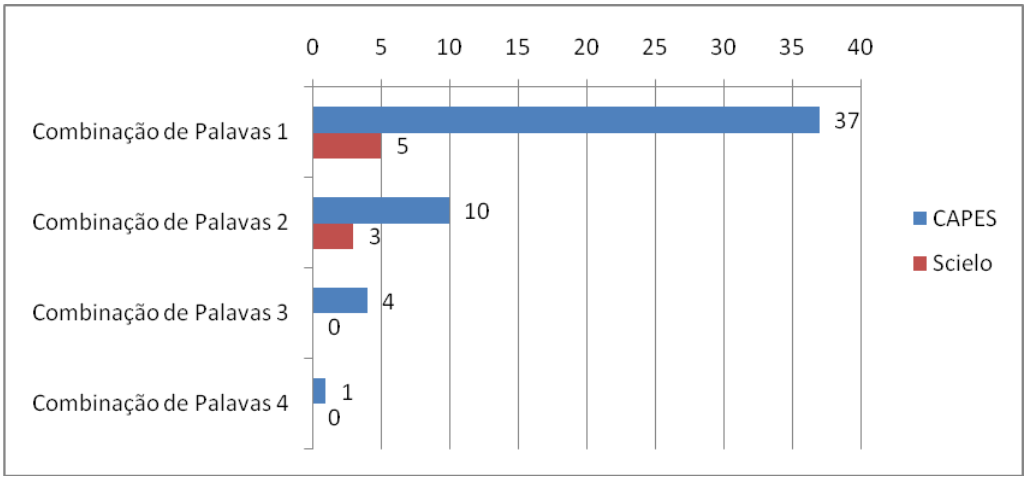
[11] http://www.scielo.org

Quadro 1 – Combinações de palavras para pesquisa nos portais

Expressão 1	Evasão AND (EAD OR "Educação a Distância" OR "Educação *On-line*")
Expressão 2	Evasão AND (EAD OR "Educação a Distância" OR "Educação *On-line*") AND ("Educação Superior")
Expressão 3	Evasão AND (EAD OR "Educação a Distância" OR "Educação *On-line*") AND ("Mineração de Dados" OR "Data Mining")
Expressão 4	Evasão AND (EAD OR "Educação a Distância" OR "Educação *On-line*") AND ("Educação Superior") AND ("Mineração de dados" OR "Data Mining")

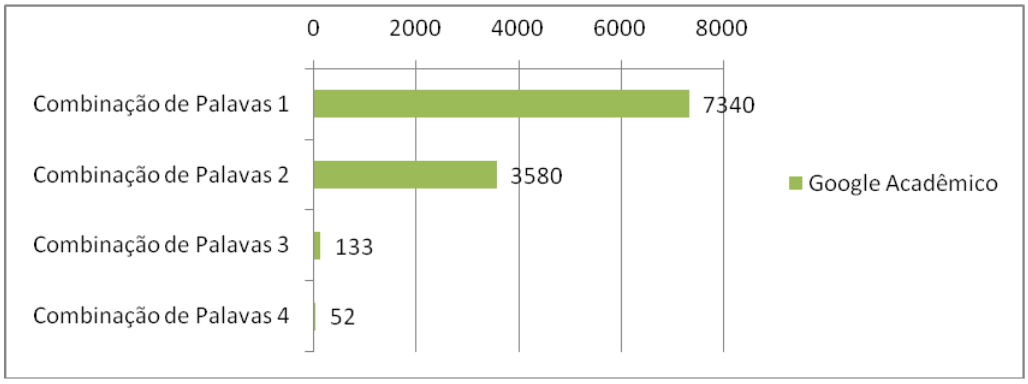
Fonte: do Autor

Gráfico 3 – Resultados das buscas no portal da CAPES e SCIELO



Fonte: do Autor

Gráfico 4 – Resultados das buscas no portal Google Acadêmico



Fonte: do Autor

Após a análise dos títulos e resumos do portal da CAPES com as combinações de palavras 1, 2, 3 e 4, do portal Scielo com a combinação 1 e 2, já que as combinações 3 e 4 não retornaram resultados e do Google Acadêmico com as combinações 3 e 4, selecionou-se os artigos listados no Quadro 2, do períododos últimos 10 anos. Os artigos foram organizados por data de publicação e ao todo foram selecionados 16 artigos, três dissertações e três teses, com maior aderência ao tema proposto.

Quadro 2 – Trabalhos científicos relacionados

Tipo*	Título	Autor(es)
2006		
D	Variáveis preditoras de evasão em dois cursos a distância	A. M. Walter
2009		
Artigo	Evasão na educação a distância	J. Laguardia
T	Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente	A. J. C. Kampff
2010		
D	Análise da evasão no curso de administração a distância: projeto-piloto UAB : um enfoque sobre a gestão	A. F. A. de Andrade
2012		
T	Adesão e permanência discente na educação a distância: investigação de motivos e análise de preditores sociodemográficos, motivacionais e de personalidade para o desempenho na modalidade	P. J. Fiuza
A	Mineração de Dados na Descoberta do Padrão de Usuários de um Sistema de Educação à Distância	J. R. Penedo, E. P Capra
A	Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais	E. Gottardo, C. Kaestner, R. V. Noronha
A	Avaliação de Desempenho de Estudantes em Cursos de Educação a Distância Utilizando Mineração de Dados	E. Gottardo, C. Kaestner, R. V. Noronha
A	Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades	S. J. Rigo, S. C. Cazella, W. Cambruzzi
D	Estimativa de desempenho acadêmico de estudantes em um AVA utilizando técnicas de mineração de dados	E. Gottardo
2013		
A	O perfil dos estudantes de administração matriculados nas disciplinas da área contábil: o caso do Centro de Educação a Distância do Estado do Rio de Janeiro (CEDERJ)	R. A. D. Condé; R. S. Quintal; S. M. de Souza; S. S. da C. Vieira
A	Estratégias para detecção precoce de propensão à evasão	A, Mezzari
A	Mineração de Dados na Identificação de Grupos de Estudantes com Dificuldades de Aprendizagem no Ensino de Programação	R. S. de França, H. J. C. do Amaral
2014		
A	Identificação de Perfis de Evasão e Mau Desempenho para Geração de Alertas num Contexto de Educação a Distância	A. J. C. Kampff, V. H. Ferreira, E. Reategui, J. V. de Lima
A	Evasão nos cursos na modalidade de educação a distância: Estudo de caso do curso piloto de administração da UFAL/UAB	I. M. BittencourtI; L. P. L. MercadoII
A	Minerando dados sobre o desempenho de estudantes de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS	S. S. da Costa, S. Cazella, S. J. Rigo
A	Uma Proposta para Identificação de Causas da Evasão na Educação a Distância através de Mineração de Dados	W. Barbosa, D. Máximo, A. Jatobá, A. Leite, E. Soares
A	Educação em Engenharia e Mineração de Dados Educacionais: oportunidades para o tratamento da evasão	S. J. Rigo, J. Barbosa, W. Cambruzzi
A	Uma Abordagem Genérica de Identificação Precoce de Estudantes com Risco de Evasão em um AVA utilizando Técnicas de Mineração de Dados	R. N. dos Santos, C. de A. Siebra, E. D. S. Oliveira
A	Analisando Fatores que Afetam o Desempenho de Estudantes Iniciantes em um Curso a Distância	J. L. C. Ramos, R. L. Rodrigues, J. Sedraz, A. S. Gomes
A	Estudantes em Risco: como identificá-los por meio de um ambiente virtual de aprendizagem?	J. M. C. da Silva, F. G. Andrade, R. Tessari
T	Sistema inteligente para a predição de grupo de risco de evasão discente	V. R. de C. Martinho

***** A = Artigo; T = Tese de Doutorado; D = Dissertação de mestrado

Fonte: do Autor

1.3 DISCUSSÃO DA REVISÃO DA LITERATURA

Segundo Favero e Franco (2006), o problema da evasão é observado em quase todas as instituições que oferecem cursos na modalidade a distância.

O estudo de Walter (2006) evidencia altas taxas de evasão em cursos realizados na modalidade EAD, contudo, ele também evidencia que existe escassez de pesquisas sobre esse fenômeno. Seu trabalho foi composto por três estudos, o primeiro objetivando construir e validar uma Escala de Comportamentos e Atitudes do Estudante em Relação a Cursos a Distância; o segundo, revalidar a Escala de Valor Instrumental do Treinamento, no âmbito da EAD; e o terceiro, objetivou analisar o relacionamento entre características da clientela (idade, gênero, participação anterior em curso a distância, pagamento do curso pelo estudante e valor instrumental do treinamento), características do curso (semipresencial ou totalmente a distância) e comportamentos e atitudes do estudante em relação a cursos a distância com a variável critério evasão.

Languardia (2009) faz uma revisão dos trabalhos relacionados à evasão em EAD, onde discute os diferentes conceitos, modelos teóricos e as variáveis identificadas como preditoras de evasão, bem como as estratégias para reduzi-la. Segundo o autor, a adoção de estratégias para redução da evasão pode ocorrer em consonância com as etapas do percurso do estudante, quais sejam:

(i) A primeira etapa refere-se à demonstração do interesse pelo curso. Para evitar que os candidatos desistam do curso por uma escolha errada, o autor sugere que seja explicitado, logo nos primeiros contatos com os estudantes, os pré-requisitos formais do curso. Ainda é sugerido que ele faça uma navegação experimental nos conteúdos, no desenho do curso, nas funcionalidades do AVA e verifiquem se o conjunto atende suas expectativas;

(ii) A segunda etapa é a da matrícula. Ela é fortemente caracterizada para dar ao estudante o sentido de pertencimento à instituição com envio de mensagens individuais pelo tutor, envio de materiais de apoio, aproximação dos pares através de fóruns de apresentações e expectativas, encontros presenciais e utilização de ferramentas síncronas, no sentido de reduzir a sensação de isolamento que a modalidade provoca e, também, disponibilizando formalmente um prazo para desistência sem prejuízo financeiro;

(iii) A terceira etapa vai do envio da primeira atividade à avaliação final, na qual a evasão do estudante depende, em certa medida, da garantia de processos de aprendizagem de qualidade e uma equipe amigável, entusiasmada e profissional. O tutor deve ter, nos primei-

ros módulos do curso, uma atitude mais proativa na oferta sistemática de dicas e conselhos para o desenvolvimento de habilidades de estudos e gerenciamento do tempo. Nessa fase também deve ser aplicado um questionário de avaliação, que inclua os níveis de satisfação dos estudantes com o tutor, o ambiente e o conteúdo. Tais indicadores sinalizam para os tutores e coordenadores as medidas a serem tomadas, tais como a revisão do currículo e da carga de estudo;

(iv) A quarta e última etapa é o retorno do estudante. Quando ele novamente escolhe a mesma instituição para um novo curso e essa escolha depende da percepção da experiência prévia de estudo à distância e se essa experiência valeu o dinheiro e o tempo investidos. Nessa etapa é necessário a realização de contatos periódicos da instituição com os egressos ou com os estudantes que abandonaram o curso, convidando-os para os programas de egresso e valorizando o vínculo institucional.

Kampff (2009), analisa as possibilidades de utilização das diversas informações geradas e mantidas nas bases de dados dos AVAs para a identificação de perfis de estudantes que pudessem estar associados com evasão do curso e, além da identificação, o trabalho foi voltado para o desenvolvimento da geração de alertas para os professores para que esses pudessem agir preventivamente no combate à evasão.

Gottardo (2012) reforça que o acompanhamento efetivo dos estudantes em cursos na modalidade à distância é um grande desafio para os profissionais que atuam nessa área. Além disso, ele traz considerações sobre outros trabalhos desenvolvidos sobre EDM[12], mostrando a amplitude da área bem como suas possibilidades. O autor sintetiza na seção 2 (Trabalhos Relacionados) que o EDM está sendo utilizado para: identificar o grau de motivação dos estudantes, desenvolvimento de AVAs adaptativos, desenvolvimento de métodos de ensino que permitam melhorar as condições de aprendizagem, predizer o desempenho do estudante, predizer a evasão dos estudantes, para o estudo comportamental dos estudantes no AVA, na melhoria de Sistemas Tutores Inteligentes e para medir os níveis de relação entre os estudantes.

Gottardo (2012) propõe um conjunto de atributos (Quadro 3) para avaliar o desempenho dos estudantes EAD, destacando que os atributos são amplos e generalizáveis, entretanto, a seleção não é uma tarefa trivial.

Inicialmente, para a seleção dos atributos o autor ponderou, sobre as questões:

1. Como acompanhar efetivamente um estudante?
2. Como verificar se os estudantes estão interagindo entre si?

[12] EDM - Mineração de Dados Educacionais (do inglês, *Educational Data Mining*) surgiu como sinônimo de KDD para extração de conhecimento especificamente em base de dados educacionais. (PINHEIRO, 2009)

3. Como identificar estudantes com problemas de desempenho?
4. Como identificar estudantes desmotivados ou prestes a abandonar o curso?

Quadro 3 – Atributos para avaliação dos estudantes EAD

Dimensão	Atributo	Descrição
Perfil Geral de uso do AVA	nr_acessos	Número total de acesso ao AVA
	nr_posts_foruns	Número total de postagens realizadas em fóruns
	nr_post_resp_foruns	Número total de respostas postadas em fóruns referindo-se a postagens de outros participantes (estudantes, professores, tutores)
	nr_post_rev_foruns	Número total de revisões em postagens anteriores realizadas em fóruns
	nr_sessao_chat	Número de sessões de chat que o estudante participou
	nr_msg_env_chat	Número de mensagens enviadas ao chat
	nr_questoes_resp	Número de questões respondidas
	nr_questoes_acert	Número de questões respondidas corretamente
	freq_media_acesso	Frequência média em que o estudante acessa o AVA
	tempo_medio_acesso	Tempo médio de acesso ao sistema
	nr_dias_prim_acesso	Número de dias transcorridos entre o início do curso e o primeiro acesso do estudante no AVA
	tempo_total_acesso	Tempo total conectado no sistema
Interação Estudante-Estudante	nr_post_rec_foruns	Número de postagens do estudante que tiveram respostas feitas por outros estudantes.
	nr_post_resp_foruns	Número de respostas que o estudante realizou em postagens feitas por outros estudantes.
	nr_msg_rec	Número de mensagens recebidas de outros participantes durante a realização do curso.
	nr_msg_env	Número de mensagens enviadas a outros participantes durante a realização do curso.
Interação Estudante-Professor	nr_post_resp_prof_foruns	Número de postagens de estudantes que tiveram respostas feitas por professores ou tutores do curso
	nr_post_env_prof_foruns	Número de postagens de professores ou tutores que tiveram respostas feitas por estudantes
	nr_msg_env_prof	Número de mensagens enviadas ao professor/tutor durante a realização do curso.
	nr_msg_rec_prof	Número de mensagens recebidas do professor/tutor durante a realização do curso.
Objetivo da Previsão	Resultado_final	Resultado final obtido pelo estudante no curso. Representa a classe objetivo da técnica de classificação

Fonte: Gottardo 2012

Para responder essas questões, o autor propõe o agrupamento dos atributos em três dimensões:

- **Perfil geral de uso do AVA**: nesta dimensão o objetivo é identificar dados que representem aspectos de planejamento, organização e gestão do tempo do estudante para a realização do curso. Para isso definiu-se indicadores gerais de quantidade e tempo médio de acessos aos recursos do AVA. Foram incluídos também atributos que representem atividades rotineiras e regulares dos acessos dos aprendizes;
- **Interação Estudante-Estudante**: com esta dimensão pretende-se verificar se os estu-

dantes interagem entre si usando as ferramentas disponíveis, como fóruns, chats, envio ou recebimento de mensagens. Espera-se, com estes atributos, identificar a existência de colaboração e cooperação entre estudantes;

- **Interação Estudante-Professor**: nesta dimensão o objetivo é averiguar como professores ou tutores interagem com estudantes no contexto do AVA. O autor ressalta que os professores ou tutores têm um papel fundamental no sentido de facilitar e incentivar a colaboração entre estudantes.

A atividade de KDD utilizada no trabalho foi a de classificação que, com base em exemplos do passado, constrói um esquema de mapeamento dos descritores de um objeto em um conjunto de classes, de modo a predizer a classe de um exemplo novo. Como essa tarefa necessita de valores discretos para representar as classes, o autor utilizou o método de discretização *equal-width* dividindo os resultados em três subintervalos, conforme mostrado na Tabela 3.

Tabela 3 – Intervalos da discretização de Gottarto

Classe	Descrição	Número de Estudantes	Intervalo de Notas
A	Estudantes com desempenho superior	22	87-97
B	Estudantes com desempenho intermediário	109	77-87
C	Estudantes com desempenho inferior	24	67-77

Fonte: Gottardo (2012)

Chegou-se por meio das técnicas RandonForest e MultilayerPerceptron, a índices de acurácia[13] de 76,6%.

Santos et al. (2014) aborda em seu estudo que a evasão pode ser investigada por intermédio das técnicas de KDD fazendo uso de dados *(i)* não variantes no tempo em conjunto com dados variantes no tempo, ou somente com *(ii)* dados variantes no tempo. Os dados não variantes no tempo, tais como endereço, forma de ingresso no nível superior e estado civil, são advindos, por exemplo, de questionários socioeconômicos. Nesse tipo de dado, a variação não é inexistente, mas sua incidência é muito baixa. Já os dados variantes no tempo são advindos dos AVAs e são gerados na medida em que os estudantes vão percorrendo o ambiente e as incidências vão sendo registradas. Tomemos como exemplo os registros de acesso às disciplinas, atividades, materiais, fóruns, exercícios, entre outros; na medida em que os estudantes vão utilizando o AVA, suas ações vão sendo registradas.

[13] Proporção entre o número de estudantes corretamente classificados pelos algoritmos em sua respectiva classe e o número total de estudantes considerados no estudo. (GOTTARDO, 2012)

Santos et al. (2014) ainda nos traz que o problema da predição de evasão de um estudante pode ser analisado como um problema de predição de desempenho no qual são consideradas duas classes principais: evadido ou graduado. Em sua pesquisa, a classe “evadido” refere-se a um estudante que não conseguiu concluir o curso seja por abandono, por insuficiência no desempenho acadêmico ou uma solicitação formal. Já a classe “graduado” refere-se a um estudante que concluiu todas as etapas para o término do curso. Como resultado, a partir dos experimentos realizados foi constatado que já ao final do primeiro período é possível predizer os padrões de comportamento que levam à evasão com precisão média maior do que 80%.

Kampff et al. (2014), utilizam técnicas de KDD para identificar perfis de evasão e mau desempenho de estudantes no contexto da educação a distância. O trabalho visou desenvolver alertas para os docentes sobre a situação dos estudantes. Tais alertas visavam dar suporte à atuação do professor no acompanhamento dos processos de aprendizagem. O estudo envolveu 1780 estudantes e permitiram concluir que o sistema de alerta proposto pode contribuir com o aumento dos índices de aprovação e redução dos índices de evasão de disciplinas na modalidade à distância.

Existem dois tipos de alertas configuráveis pelo professor. O primeiro tipo de alerta é baseado em indicadores do ambiente virtual, chamados de alertas fixos, que estão relacionados a questões que o professor deseja acompanhar explicitamente, tais como acesso a materiais, avaliações, presença em sala de aula, acesso ao sistema, realização de atividades. O segundo tipo de alerta é baseado em padrões obtidos por meio de processos de mineração de dados, e são automaticamente gerados pelo sistema. O módulo de mineração de dados consulta os dados históricos das disciplinas ou cursos correspondentes, previamente organizados, e gera as regras de classificação para o período correspondente, classificando os estudantes em TENDÊNCIA DE EVADIDO, TENDÊNCIA DE APROVADO, TENDÊNCIA DE REPROVADO e SEM ACESSO.

Cada alerta, com base no evento ou regra específica, aponta o grupo de estudantes, notificando o professor e sugerindo o contato com os estudantes. A partir da análise do professor, outras ações puderam ser tomadas, como, por exemplo, a ampliação de prazos ou a oferta de materiais complementares. Com essas ações pode-se perceber melhoria nas aprovações e baixa nos índices de evasão e reprovação, conforme pode ser visto na Tabela 4.

Tabela 4 – Resultados alcançados com a utilização do sistema de alertas.

Resultado	Amostra histórica		Amostra Atual (apoiada pelo sistema de alerta)	
	Nº de Estudantes	Percentual	Nº de Estudantes	Percentual
Aprovado	1286	82,97	199	86,52
Reprovado	52	3,35	7	3,04
Evadido	212	13,68	24	10,43
Total	1550	100,00	230	100

Fonte: Kampff et al. (2014)

Além dos resultados obtidos sobre aprovação, reprovação e evasão, na pesquisa também foi possível observar que o sistema de alerta contribuiu de maneira significativa na gestão do ambiente virtual de aprendizagem. Com dois dias de antecedência do prazo para entrega de cada tarefa, o professor era notificado sobre os estudantes que ainda estavam com a tarefa pendente e também sobre grupos de estudantes com tendência à aprovação reprovação ou evasão. Assim, eles puderam tomar medidas para contornar a situação, tais como: lembrar dos prazos das atividades, ampliar os prazos de entrega das atividades ou disponibilizar material complementar.

Muitas das práticas docentes para engajamento dos estudantes em seus estudos foram realizadas a partir do encaminhamento de mensagens aos estudantes.

Costa. et al. (2014) buscam, "através da aplicação do processo de Descoberta de Conhecimento em Bases de Dados, explicitar padrões de evasão nos cursos de educação permanente em modalidade EAD para profissionais da saúde, promovidos pela Universidade Aberta do SUS (UNA-SUS)." Os autores afirmam ter chegado a uma acurácia de 97,6% aplicando a tarefa de regra de classificação com a técnica de árvore de decisão. Os dados foram fornecidos pela Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), para determinação do perfil do estudante propenso a evadir do curso.

Barbosa (2014) descreve uma proposta para uso de técnicas de Mineração de Dados para identificação das causas de evasão e de levantamento do perfil do estudante evadido num projeto de ensino de língua inglesa à distância denominado IngRede. Tal projeto visa a capacitação de estudantes desde o nível técnico ao doutorado e vem sendo realizado desde 2012 o qual já recebeu mais de 1000 inscrições, porém, somente 100 estudantes concluíram o curso, somando uma taxa de evasão de 90%. A proposta do autor é aplicar técnicas de EDM, com a metodologia CRISP-DM, na base de dados atual para predizer a evasão futura, ficando a parte da execução da metodologia selecionada para um próximo artigo, que infelizmente não estava disponível até fevereiro de 2016.

A evasão nos cursos de graduação a distância tem se mostrado um dos problemas que

mais preocupam os gestores de EAD, conforme relatos no Censo EAD (ABED, 2013). Nesse sentido, alguns trabalhos foram e estão sendo realizados para entender as razões que levam um estudante a evadir. Alguns trabalhos estão sendo direcionados para predizer quais estudantes estão propensos a evadir, e dessa forma viabilizar ações preditivas, preventivas e corretivas para mitigar esse fenômeno.

1.4 SINTESE DA REVISÃO DA LITERATURA

Com a revisão da literatura, pode-se perceber que a maioria das instituições fornecedoras de cursos na modalidade EAD apresentam altas taxas de evasão e que o acompanhamento efetivo dos estudantes é um grande desafio para os profissionais que atuam nessa modalidade. Tal fenômeno despertou o interesse de vários pesquisadores.

As pesquisas sobre evasão incluíram técnicas de KDD, utilizando informações geradas e mantidas nas bases de dados educacionais, a fim de identificar o comportamento dos estudantes com tendência a evadir, emitir alertas e/ou fornecer informações importantes para que medidas preventivas possam ser tomadas.

Observa-se que emissão de alertas preditivos e/ou geração de informações gerenciais torna-se uma importante ferramenta para o combate à evasão, auxiliando no incremento dos índices de conclusão dos cursos, e melhorando inclusive as práticas docentes para ampliar o grau de engajamento dos estudantes no curso.

Com a utilização específica das técnicas de KDD nas bases de dados educacionais, o termo *Education Data Mining* (EDM) passou a ser utilizado em vários artigos, dissertações e teses que investigam tal fenômeno. O EDM tem sido utilizado em várias pesquisas educacionais, dentre elas a identificação do comportamento dos estudantes propensos a evadir, bem como para medir o grau de motivação dos estudantes, fato que está relacionado à evasão ou permanência dos estudantes no curso.

As pesquisas mostram que entre as instituições de ensino existe uma preocupação com os altos índices de evasão, e várias pesquisas tem sido realizadas a fim de entender e combater esse fenômeno, alguns autores fazem uso de técnicas de KDD para dentre outras coisas, investigar os estudantes propensos e evadir. Os índices de acerto na identificação desses estudantes variam entre 76% até 97%.

1.5 PROBLEMATIZAÇÃO

Em ABED (2013) é possível identificar o crescimento do número de cursos ofertados e de estudantes da modalidade EAD, consolidando uma alternativa para a formação de nível superior. Porém, essa modalidade educacional enfrenta o problema da evasão escolar, que acarreta prejuízo para o indivíduo, para a instituição, bem como para a sociedade.

Considerando que: *(i)* a evasão de estudantes é apontada como o maior obstáculo enfrentado pelos cursos da modalidade EAD, *(ii)* partindo do pressuposto de que é possível identificar os estudantes com tendência a evadir preditivamente, *(iii)* a importância de gerar informações que forneçam insumo ao processo de tomada de decisão, colocam-se as questões relacionadas ao problema motivador dessa dissertação: como, através do uso do KDD, podemos identificar e predizer, os estudantes EAD propensos a evadir e gerar informações para que os gestores adotem medidas para mitigar o problema.

Em posse de um modelo de predição de evasão, os gestores poderão receber, em tempo hábil, informações relevantes para acompanhar e agir preventivamente no combate a evasão, podendo colocar-se à disposição dos estudantes com alguma dificuldade, solicitando maior empenho dos estudantes que não acessam com frequência o AVA, identificando estudantes com dificuldades técnicas, lembrando dos prazos das atividades e dos encontros presenciais avaliativos, estimulando a participação dos fóruns de discussão e/ou atividades propostas pelo professor e incentivando a interação entre os estudantes.

1.6 OBJETIVOS

O objetivo geral deste trabalho é propor um modelo de predição de evasão de estudantes EAD tendo por base técnicas de KDD aplicadas a dados gerados por um SGA/AVA.

Para isso, identificam-se os objetivos específicos:

- Compreender a categoria evasão em EAD;
- Estudar modelos preditivos capazes de detectar comportamentos indicativos de evasão;
- Propor ações para mitigar o problema da evasão;
- Propor inovações e relançar o problema para novas pesquisas.

1.7 JUSTIFICATIVA

O estudo torna-se relevante em relação à sua contribuição para gestão de cursos na modalidade EAD face ao problema de evasão. Instituições de ensino poderão adequar o Modelo de Predição de Evasão, bem como um conjunto de ações, à sua realidade.

O constante crescimento da EAD abre espaço para novas pesquisas e discussões. Portanto, este estudo visa cooperar com as pesquisas correlatas com essa temática, ressaltando o caráter de interdisciplinaridade onde áreas como Educação, Tecnologia da Informação e inteligência competitiva, são relacionadas para a solução do problema.

Dado a relevância do tema, a pesquisa pode auxiliar no combate da evasão na instituição pesquisada.

2. REFERENCIAL TEÓRICO

As pesquisas realizadas nos motores de busca abordados nesse item revelam a existência de vários artigos, dissertações e teses que de forma direta ou indireta se relacionam com essa pesquisa. Buscou-se neste capítulo articular e discutir os avanços teóricos e práticos relatados em tais publicações de modo a pavimentar o caminho para a proposição de uma abordagem para identificação e tratamento do problema da evasão em EAD.

2.1 EDUCAÇÃO A DISTÂNCIA

Segundo o Decreto nº 5.622, a EAD pode ser descrita como:

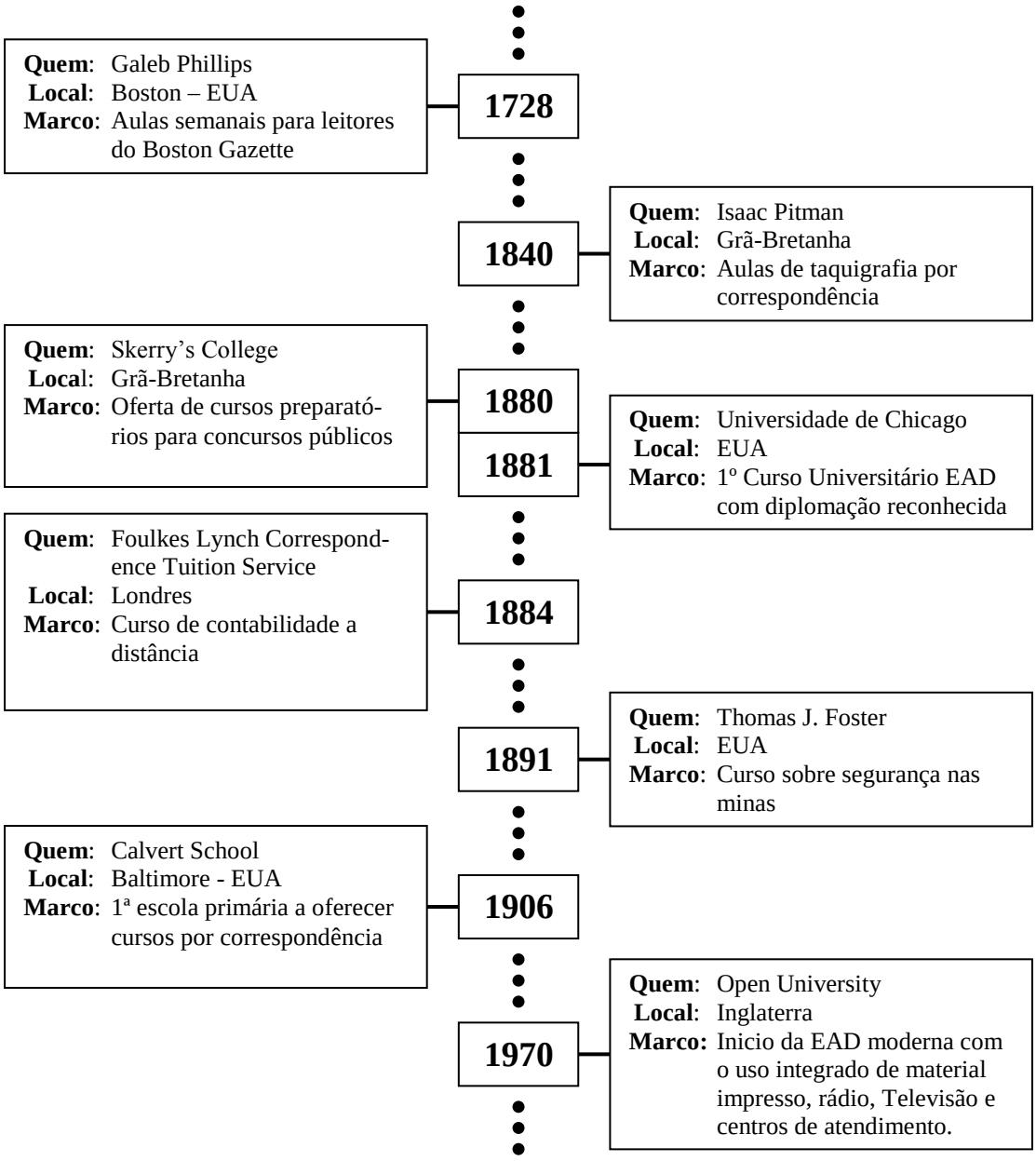
> [...] a modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre com a utilização de meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo atividades educativas em lugares ou tempos diversos. (BRASIL 2005, p. 1)

O Decreto acima é contemporâneo, datado de 2005, contudo, dentre os vários autores que se dedicam aos estudos de EAD não existe uma unanimidade quanto à classificação das gerações e da evolução histórica da EAD. Alguns autores fazem referência ao surgimento da EAD às civilizações antigas, considerando as mensagens escritas utilizadas para difusão do cristianismo como a primeira iniciativa educacional a distância. As Epístolas de São Paulo, são exemplos dessa ação, em que informação e conhecimento são levados a lugares distantes, assumindo assim uma nova dimensão de circulação e mobilidade do conhecimento (LANDIM,1997).

Mais recentemente, na visão de Jónasson (2001), a EAD foi marcada por vários grandes acontecimentos, dentre os quais destacam-se os apresentados na Figura 1.

A EAD moderna, de origem ainda mais recente, deu-se com o êxito da Open University da Inglaterra, que surgiu no final dos anos de 1960 com o início dos seus cursos em 1970, passando a ser referência estabelecida para vários autores como marco de EAD contemporânea (NUNES, 2009).

Figura 1 – Principais marcos da EAD



Fonte: Adaptado de Jónasson (2001)

Dentre os vários autores pesquisados, não existe uma unanimidade sobre quando vamos tratar das gerações da EAD, entretanto, a definição de Taylor (2009) torna-se interessante para esse trabalho por fazer referência à EAD enquanto cursos universitários, os quais são objeto de estudo dessa pesquisa. Portanto, tomando-a como referência, encontramos as seguintes gerações:

1ª geração: oferece cursos por correspondência com escassa ou nula interatividade entre as partes;

2ª geração: os cursos a distância deixam de se basear exclusivamente na remessa de materi-

ais por correspondência e passam a explorar outros caminhos para acessar e estimular o aprendiz. A partir desse modelo, a oferta de cursos a distância é enriquecida com reuniões, encontros e sessões periódicas de tutorias, emissões radiofônicas, remessas de leituras específicas, em material impresso complementar e titulação oficial;

3ª geração: fortemente caracterizada pelas tecnologias da comunicação e da informação (audioconferência, videoconferência, rádio e tv em rede);

4ª geração: aprendizagem flexível, caracterizada pela forte influência de processos interativos, tendo como recurso principal o computador, seja por meio de softwares, seja pelo uso da internet;

5ª geração: engloba as tecnologias da quarta geração aliadas à comunicação via computadores com sistema de respostas automatizadas, além de acesso via portal a processos institucionais. O diferencial nessa geração será a diminuição de custos e melhoria da qualidade da interação. A disseminação das redes sem fios e de portais que popularizem a interação também fazem parte dessa geração.

Percebe-se, portanto, que a EAD saiu do uso dos serviços postais, passou pela utilização do rádio, da TV, pelo uso de computadores com os softwares educacionais e utilização de CDs de vídeos-aula. Tais formatos foram decaindo em detrimento principalmente da utilização dos recursos da Internet e seus serviços, como e-mail, bate-papo, fórum e recursos multimídias. Com a utilização da Internet a EAD deu um grande salto qualitativo e de alcance. Tais recursos foram integrados em ambientes próprios para atividades de ensino-aprendizagem, os Ambientes Virtuais de Aprendizagem (AVA), propiciando à educação a distância *on-line*, tendo como meio a Internet.

Moran (2003) define educação on-line como o conjunto de ações de ensino-aprendizagem desenvolvidas através de meios telemáticos, como a Internet, a videoconferência e a teleconferência. Para esse autor, educar em ambientes virtuais exige mais dedicação do professor, mais apoio de uma equipe técnico-pedagógica, mais tempo de preparação ao menos na primeira fase e principalmente de acompanhamento, mas para os estudantes há um ganho grande de personalização da aprendizagem, de adaptação ao seu ritmo de vida, principalmente na fase adulta.

A EAD, especialmente na modalidade *on-line*, tem valor significativo para o atendimento de grandes quantidades de estudantes. Nesse quesito, ela é efetiva e melhor que a modalidade presencial. Sendo trabalhada corretamente não sofre riscos de reduzir a qualidade dos serviços oferecidos em decorrência da expansão do número de participantes (Alves, 2005). O autor ainda cita que existem várias vantagens nessa modalidade de ensino, particu-

larmente no que concerne à flexibilidade que o estudante tem para o seu estudo, entretanto, problemas oriundos da educação tradicional também encontraram seu espaço na educação a distância, tal como a evasão, que veremos mais a diante.

Atualmente para se valer do ensino-aprendizagem a distância *on-line*, professores e estudantes utilizam os Ambientes Virtuais de Aprendizagem (AVAs) que, segundo Dias (2008), podem ser definidos como um sistema que fornece suporte a qualquer tipo de atividade realizada pelo estudante, isto é, um conjunto de ferramentas de comunicação e interação que são utilizadas em diferentes situações do processo de aprendizagem.

Como pode ser observado nos Quadros 4 e 5, os AVAs agrupam uma série numerosa de ferramentas que podem dar suporte a diferentes abordagens epistemológicas. Gonzales (2005) agrupa as ferramentas em 4 grandes conjuntos, a saber: ferramentas de coordenação; ferramentas de comunicação; ferramentas de produção dos estudantes ou de cooperação e ferramentas de administração; enquanto Sabbatini (2007) agrupa em 3 conjuntos, a saber: ferramentas de comunicação; ferramentas de interação; e ferramentas de avaliação. Basicamente as ferramentas são as mesmas, sendo elas classificadas e agrupadas de diferentes formas pelos autores.

Quadro 4 - Grupos de ferramentas dos ambientes virtuais de aprendizagem

Grupo de ferramenta	**Descrição**
Ferramentas de Coordenação	Servem de suporte para a organização de um curso. São utilizadas pelo professor para disponibilizar informações aos estudantes, tanto informações das metodologias do curso (procedimento, duração, objetivos, expectativa, avaliação) e estrutura do ambiente (descrição dos recursos, dinâmica do curso, agenda, etc.), quanto informações pedagógicas: material de apoio (guias, tutoriais), material de leitura (textos de referência, links interessantes, bibliografia, etc.) e recurso de perguntas freqüentes (reúne as perguntas mais comuns dos estudantes e as respostas correspondentes do professor).
Ferramentas de Comunicação	Englobam fóruns de discussão, bate-papo, correio eletrônico e conferência entre os participantes do ambiente. Têm o objetivo de facilitar o processo de ensino-aprendizagem e estimular a colaboração e interação entre os participantes e o aprendizado contínuo.
Ferramentas de Produção dos Estudantes ou de Cooperação	Oferecem o espaço de publicação e organização do trabalho dos estudantes ou grupos, através do portfólio, diário, mural e perfil (de estudantes e/ou grupos).
Ferramentas de Administração	Oferecem recursos de gerenciamento, do curso (cronograma, ferramentas disponibilizadas, inscrições, etc.), de estudantes (relatórios de acesso, frequência no ambiente, utilização de ferramentas, etc.) e de apoio a tutoria (inserir material didático, atualizar agenda, habilitar ferramentas do ambiente, etc.). Através delas é possível fornecer ao professor formador informações sobre a participação e progresso dos estudantes no decorrer do curso, apoiando-os e motivando-os durante o processo de construção e compartilhamento do conhecimento.

Fonte: Gonzales (2005)

Quadro 5 - Recursos de TI utilizados em ambientes virtuais

Grupo de ferramenta	**Descrição**
Ferramentas de Comunicação	– Páginas simples de texto – Páginas em HTML – Acesso a arquivos em qualquer formato (PDF, DOC, PPT, Flash, áudio, vídeo, etc.) ou a links externos (URLs) – Acesso a diretórios (pastas de arquivos no servidor) – Rótulos – Lições interativas – Livros eletrônicos – Glossários (estático) – Perguntas frequentes
Ferramentas de Interação	– Chat (bate-papo) – Fórum de discussão – Diários – Wikis (textos colaborativos) – Glossários (colaborativo)
Ferramentas de Avaliação	– Avaliação do curso – Questionários de avaliação – Ensaios corrigidos – Tarefas e exercícios – Enquetes

Fonte: Sabbatini (2007)

A forma como as ferramentas são disponibilizadas pelo professor, bem como elas são utilizadas, possibilita que um mesmo AVA aborde o processo de ensino-aprendizagem frente a diferentes enfoques epistemológicos. Sendo possível desde um modelo de ensino linear, com sequências preestabelecidas e com pouca ou nenhuma possibilidade de o estudante ser o protagonista da sua aprendizagem, a outros que trazem em suas propostas uma abordagem metodológica que suporta a construção cooperativa do conhecimento.

Portanto, o professor atuante na modalidade EAD, que utilize AVAs com a determinação de promover a educação *on-line* poderá escolher dentre as várias teorias de aprendizagem, e dispor o conjunto de ferramentas e conteúdos de forma que ela seja evidenciada.

Existem diversos AVAs disponíveis. Entre os AVAs, software livre[14], mais difundidos estão Amadeus[15], Moodle[16], e-ProInfo[17] e Eureka[18].

Segundo Melo et al. (2011, p. 30), "o ambiente virtual de aprendizagem Amadeus tem origem em um conjunto de pesquisas acadêmicas na área de Interação Humano Computador e Tecnologia Educacional". Para esse autor, no Amadeus, a interação entre os usuários, e destes

[14] Software que estão sob as licenças reconhecidas pela Free Software Fundation

[15] http://amadeus.cin.ufpe.br/

[16] https://moodle.org/

[17] http://e-proinfo.mec.gov.br

[18] https://eureka.pucrp.br

com o conteúdo no ambiente permite a execução de novas estratégias de ensino e de aprendizagem orientadas por teorias construtivistas[19] ou sócio-interacionista[20] do desenvolvimento humano.

De acordo com Ribeiro et al. (2007), Moodle é uma plataforma Software Livre, sob a licença GPL[21], podendo ser instalado, utilizado, modificado e mesmo distribuído gratuitamente . Tem como objetivo o gerenciamento do aprendizado e do trabalho colaborativo no ambiente virtual, permitindo a criação e administração de cursos on-line, grupos de trabalho e comunidades de aprendizagem.

Conforme Gabardo (2010), e-ProInfo é um software público, desenvolvido pela Secretaria de Educação a Distância (SEED), do Ministério da Educação, licenciado por meio da GPL-GNU (Licença Pública Geral). Oferece projetos colaborativos e, no item interatividade, disponibiliza vários recursos tais como: tira-dúvidas, agenda, diário, biblioteca, aviso, correio eletrônico e chat.

Para Gabardo (2010, p. 4), Eureka é um "projeto de pesquisa do Laboratório de Mídias Interativas (LAMI), da Pontifícia Universidade Católica do Paraná (PUCPR). O AVA Eureka tem o objetivo de promover educação e treinamento a distância por meio da internet. Seu principal diferencial em relação às plataformas observadas é a utilização de áudio do texto escrito em todas as telas acessadas"

Mesmo com a farta oferta de ferramentas, conforme visto nos Quadros 8 e 9, a detecção de deficiências de aprendizagem não é trivial, o que mostra que não basta abrir uma sala com conteúdos e ferramentas síncronas e assíncronas; é preciso acompanhar o estudante a fim que ele se mantenha entusiasmado e participando das diversas atividades propostas.

Além disso, segundo Moran (2003), se a manutenção da motivação em ambientes presenciais de ensino-aprendizagem já merece atenção, no virtual essa tarefa é ainda mais crítica. É preciso envolver os estudantes em processos participativos, afetivos, e que inspirem confiança. Os cursos que se limitam à transmissão de informação, de conteúdo, mesmo que brilhantemente produzidos, correm o risco da desmotivação em longo prazo. É preciso, entre outras estratégias, equilibrar conteúdos teóricos e práticos. Em sala de aula, se atentos, os pro-

[19] Construtivismo é uma das correntes empenhadas em explicar como a inteligência humana se desenvolve. Ela parte do principio de que o desenvolvimento da inteligência é determinado pelas ações mútuas entre o individuo e o meio. (BECKER, 2003).

[20] Sócio-interacionismo é uma teoria de aprendizagem cujo foco está na interação. Segundo esta teoria, a aprendizagem dá-se em contextos históricos, sociais e culturais e a formação de conceitos científicos dá-se a partir de conceitos quotidianos. (BECKER, 2003).

[21] GPL – Licença Publica Geral (do Inglês *General Public Licence*), criada por Richard Stallman, fundador da Free Software Fundation. Entre os Softwares Livres, a GPL atualmente é a licença de Software Livre mais difundida. (LICENSE, 2016)

fessores podem mais facilmente obter *feedback* da ocorrência de problemas e negociar novas estratégias pedagógicas. Isso é mais difícil em ambientes virtuais, onde o estudante normalmente só é acessível por e-mail ou pelas ferramentas disponíveis no AVA.

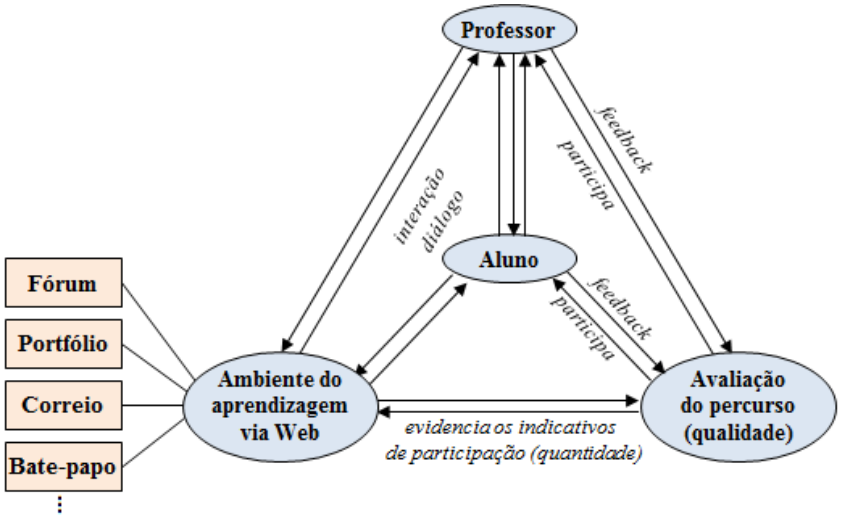
Moran (2003), afirma que os processos convencionais de ensino associados à atual dispersão da atenção da vida urbana dificultam a autonomia e a organização pessoal, indispensáveis para os processos de aprendizagem a distância. O estudante desorganizado em relação à sua agenda terá dificuldade em acompanhar o ritmo de um curso à distância. Isso influencia sua motivação, sua própria aprendizagem e a do grupo, o que cria tensão ou indiferença. Esses estudantes pouco a pouco vão deixando de participar, de produzir, e muitos têm dificuldade em retomar o entusiasmo pelo curso. Entretanto, a pouca quantidade de interações não significa necessariamente um indicador de evasão. Podem ser interessantes informações como dados pessoais (e.g., sexo, idade, curso e polo presencial), os acessos dos estudantes ao AVA como quantidade e frequência, as avaliações realizadas e seus respectivos desempenhos, o acesso aos materiais disponibilizados em áreas de conteúdo ou biblioteca virtual, as participações nos fóruns e as trocas de mensagens e e-mail realizadas.

Tecnologias computacionais tais como os agentes de software[22] e KDD, vêm sendo consideradas para prestar auxílio aos professores da modalidade a distância no acompanhamento dos estudantes, não só em aspectos objetivos de seu desempenho (avaliações, exercícios, etc.), mas também em aspectos subjetivos, como a motivação (LACHI, et al., 2002). Para isso, busca-se explorar melhor os registros das interações dos estudantes em ambientes de EAD e prover suporte ao professor na coleta, identificação, seleção e análise de informações relevantes à avaliação formativa.

Existem muitas possibilidades de interação nos ambientes virtuais de aprendizagem: a interação entre o estudante e o professor, entre o estudante e a turma e entre o estudante e os recursos disponíveis no AVA. A Figura 2, abaixo, apresenta a principais interações entre os diversos estudantes, professores e os recursos envolvidos em uma atividade de ensino-aprendizagem via Internet, todas alimentando as bases de dados, matéria prima para aplicação do KDD.

[22] Agente de Software é um programa que executa em segundo plano e tem como principais requisitos: um ciclo de vida contínuo no tempo, um ambiente de atuação, sensores para recolher informações do ambiente, atuadores que alteram o ambiente e têmautonomia, ou seja, funcionamento independente da interferência do usuário.

Figura 2 – Mapa de interações dos estudantes



Fonte: Souza (2007)

2.2 MINERAÇÃO DE DADOS

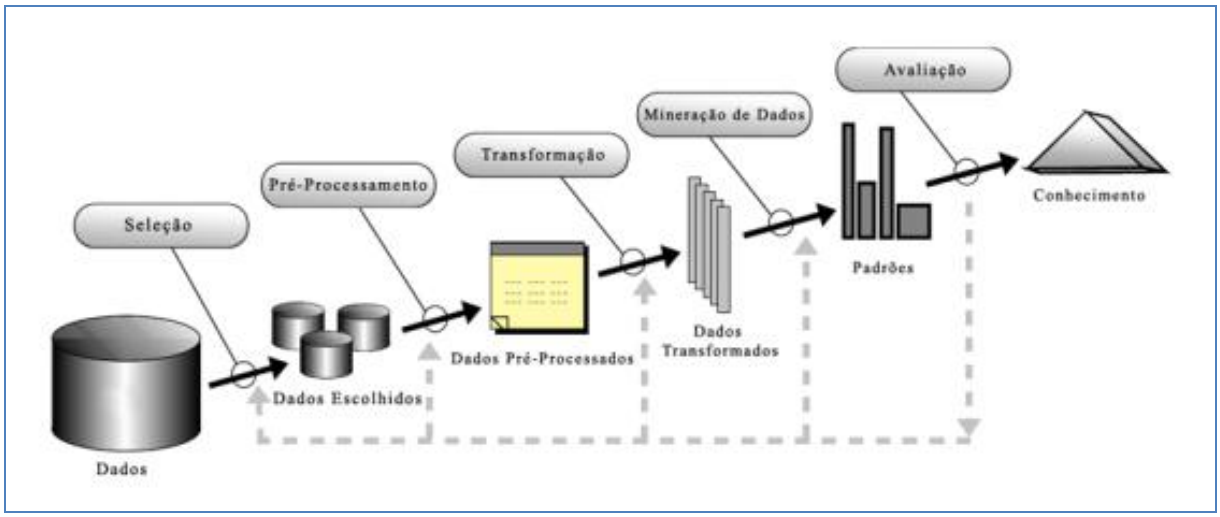
Conforme Camilo e Silva (2009), desde o surgimento dos sistemas computacionais, um dos principais objetivos das organizações tem sido o de armazenar dados. Nas últimas décadas essa tendência ficou ainda mais evidente com a queda nos custos de hardware, tornando possível armazenar quantidades cada vez maiores de dados. Em termos mundiais, o volume de dados armazenado é gigantesco e continua crescendo rapidamente (BARBOSA, 2014). Mas o que fazer com todos esses dados? Como podemos utilizá-los para obtenção/extração de informações úteis? Segundo Camilo e Silva (2009), com a finalidade de responder a estas questões, foi proposta, no final da década de 80 a Descoberta de Conhecimento em Base de Dados (do Inglês *Knowledge Discovery in Databases* - KDD).

Fayyad et al. (1996), define KDD como um "processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados".

> Essas informações são de difícil detecção por métodos tradicionais de análise e devem ser potencialmente úteis para tomada de decisão. Enquanto os métodos tradicionais são capazes de tratar apenas as informações explícitas, a extração de conhecimento é capaz de detectar informações implícitas armazenadas nos bancos de dados. (LIMA; MOURA, 2015, p. 4).

Ainda segundo Fayyad et al. (1996) para que o conhecimento seja obtido através do processo de KDD, é necessário a execução de 5 etapas: seleção, pré-processamento, transformação, mineração de dados e avaliação. Conforme demonstrado na Figura 3.

Figura 3 – Etapas do KDD



Fonte: Fayyad (1996)

Todo processo é iterativo e interativo, pois todas as etapas estão conectadas e contêm tarefas e decisões a serem realizadas pelos usuários. Caso a etapa de avaliação não seja satisfatória, de acordo com o processo, podemos voltar para a etapa de preparação dos dados.

O conhecimento que se consegue adquirir através do KDD tem se mostrado bastante útil nas mais diversas áreas, como educação, medicina, finanças, comércio, marketing, telecomunicações, meteorologia, agropecuária, bioinformáticas, entre outras (GOLDSCHMIDT; PASSOS, 2005).

O KDD possui 5 etapas, sendo que uma delas é a de Mineração de Dados (do Inglês *Data Mining* - DM). Vários autores se referem ao KDD por DM, mas para esse trabalho, vamos utilizar a o conceito KDD como todo o processo de descoberta do conhecimento em base de dados e DM como uma das cinco etapas do KDD, conforme descrito por Silva (2004).

> As ferramentas e técnicas empregadas para análise automática e inteligente dos imensos repositórios de dados de indústrias, governos, corporações e institutos científicos são os objetos tratados pelo campo emergente da Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases*-KDD). Mineração de dados é a etapa em KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão. (SILVA, 2004, p. 1)

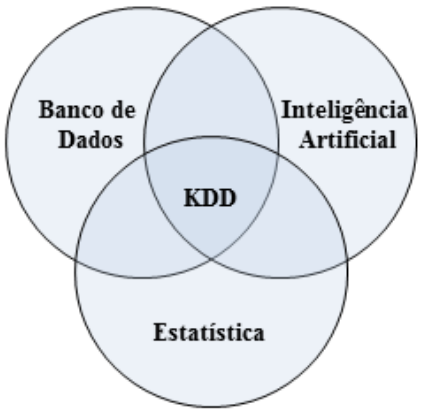
Segundo Cardoso e Machado (2008), nas últimas décadas em que a maioria das operações e atividades das instituições privadas e públicas são registradas computacionalmente e se acumulam em grandes bases de dados, a técnica de KDD é uma das alternativas mais eficazes para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem ajudar as institui-

ções nas tomadas de decisões mais rápidas ou, até mesmo, a atingir um maior grau de confiança.

De acordo com Cardoso e Machado (2008), corroborado por Costa et al. (2014), afirma que a organização que emprega o processo de KDD na análise de seus dados é capaz de: *i)* criar parâmetros para entender o comportamento dos dados, que podem ser referentes a pessoas envolvidas com a organização; *ii*) identificar afinidades entre dados que podem ser, por exemplo, entre pessoas e produtos e/ou serviços; *iii*) prever hábitos ou comportamentos das pessoas e analisar hábitos para se detectar comportamentos fora do padrão, entre outros. Portanto, a investigação dos dados utilizando as diversas técnicas de KDD, pode revelar informações úteis para toda organização, trazendo mais conhecimento para auxilio na tomada de decisões e diferenciais competitivos em relação às organizações correlatas que não utilizam o KDD.

O KDD é, portanto, uma forma de explorar e analisar bancos de dados, buscando identificar regras, padrões ou desvios e tem sido objeto de estudos interdisciplinares, envolvendo, principalmente, as áreas de Estatística, Inteligência Artificial (IA) e Banco de Dados (BD), representado na Figura 4.

Figura 4 – Áreas que envolvem o KDD



Fonte: Lin e Cercone (1997)

2.3 MINERAÇÃO DE DADOS EDUCACIONAIS (EDM)

Para Pinheiro et al. (2009), KDD tem sido utilizada com o intuito de investigar perguntas cientificas na área de educação como por exemplo, quais são os fatores que afetam a aprendizagem? Como desenvolver sistemas educacionais mais eficazes? Ou ainda, a relação da abordagem pedagógica e o aprendizado do estudante. Estas informações podem ser úteis não somente para os educadores, mas também para os próprios estudantes, uma vez que pode

ser orientada para diferentes fins por diferentes participantes no processo. Dentro deste contexto surgiu a Mineração de Dados Educacionais (do inglês, *Educational Data Mining* - EDM) como sinônimo de KDD, para extração do conhecimento especificamente em base de dados educacionais.

Segundo Baker (2009), frente à expansão dos cursos a distância e também daqueles com suporte computacional, muitos pesquisadores da área de Informática na Educação (em particular, Inteligência Artificial Aplicada à Educação) têm mostrado interesse em utilizar mineração de dados dentro deste contexto.

O conceito de EDM é complementado por Baker (2009) e Gottardo et al. (2012) afirmando que o foco do EDM é o desenvolvimento de métodos para realizar descoberta de conhecimento em bases de dados educacionais.

> A mineração de dados educacionais (EDM)[...] tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Atualmente ela vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino. (BAKER, 2011, p 1).

Pinheiro (2014) reforça que o principal foco do EDM é o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Dessa maneira é possível compreender de forma adequada os estudantes: como eles aprendem o papel do contexto no qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem.

Com a recente ampliação dos cursos de educação a distância e daqueles com suporte computacional, surgem perguntas que impulsionam as pesquisas no ramo da EDM, por exemplo, como utilizar os dados gerados por AVAs para aperfeiçoar o ensino a distância? Ou como predizer quais estudantes tem uma maior tendência a evadir? (GOTTARDO et al., 2012).

Com EDM é possível, por exemplo, verificar a relação entre uma abordagem pedagógica e o aprendizado do estudante. Com esta informação, o professor pode compreender se a metodologia de ensino utilizada está realmente ajudando o estudante, possibilitando a consideração de métodos alternativos para um ensino mais eficaz (BAKER, 2009).

Romero e Ventura (2010) descrevem os papéis de interesse do EDM (Quadro 6), dentre os quais destacamos no quadro abaixo a predição do desempenho dos estudantes, que, a partir das informações geradas, os educadores (professores e tutores) podem agir de forma preventiva, caso sejam detectados padrões no comportamento dos estudantes que podem levar a evasão.

Quadro 6 – Papéis de interesse da mineração de dados educacionais (EDM)

Atores	Objetivos para o uso de Data Mining
Estudantes	– Personalizar o e-learning; – Recomendar atividades, recursos e tarefas para que possam melhorar ainda mais a sua aprendizagem; – Sugerir experiências de aprendizagens interessantes para os estudantes; – Sugerir um caminho mais curto ou simplesmente links, para gerar sugestões de adaptação, para recomendar cursos, discussões relevantes, livros, etc.
Professores	– Obter feedback objetivo sobre a instrução; – Analisar a aprendizagem e o comportamento dos estudantes; – Detectar quais os estudantes necessitam de apoio; – **Predizer o desempenho dos estudantes;** – Classificar os estudantes em grupos; – Encontrar padrões regulares dos estudantes, bem como padrões irregulares; – Encontrar os erros mais frequentes; – Determinar as atividades mais eficazes; – Melhorar a adaptação e personalização de cursos, etc.
Desenvolvedores de cursos/pesquisadores educacionais	– Avaliar e manter material didático; – Melhorar a aprendizagem do estudante; – Avaliar a estrutura do conteúdo do curso e sua eficácia no progresso da aprendizagem; – Construir automaticamente modelos de estudantes e modelos de tutor; – Comparar as técnicas de extração de dados a fim de ser capaz de recomendar a mais útil; – Desenvolver ferramentas de mineração de dados específicas para fins educativos; etc.
Instituições de ensino	– Melhorar os processos de decisão em instituições de ensino superior; – Agilizar a eficiência no processo de tomada de decisão; – Alcançar objetivos específicos; – Sugerir certos cursos que podem ser importantes para as classes de estudantes; – Encontrar a melhor forma de melhorar a retenção e notas; – Selecionar os candidatos mais qualificados para a graduação; – Ajudar a admitir estudantes que farão bem em universidade (quando o estudo é realizado em bases de dados do ensino médio), etc.
Gestores escolares	– Desenvolver a melhor maneira de organizar os recursos institucionais e sua oferta educacional; – Utilizar os recursos disponíveis de forma mais eficaz; – Para aprimorar as ofertas de programas educacionais e de determinar a eficácia da abordagem de ensino a distância; – Avaliar professor e currículo; – Definir parâmetros para a melhoria da eficiência do AVA e adaptá-lo aos usuários.

Fonte: ROMERO e VENTURA (2010)

Apesar de EDM ser uma abordagem relativamente nova, já existem vários segmentos de atuação. A taxonomia das principais sub-áreas de EDM que foi adaptada de Baker (2011), e está apresentada no Quadro 7.

Quadro 7 – As principais categorias de EDM.

Atividade/Método	Objetivo do Método	Principal Aplicação
Predição – **Classificação** – Regressão – Estimação	O objetivo é desenvolver modelos que deduzam aspectos específicos dos dados, conhecidos como variáveis preditivas (*predicted variables*), através da análise e fusão dos diversos aspectos encontrados nos dados, chamados de variáveis preditoras (*predictor variables*).	Detectar comportamentos dos estudantes, como por exemplo, comportamentos fora do padrão; possíveis resultados educacionais.
Agrupamento	Encontrar dados que naturalmente se agrupam, dividindo dados completos em conjuntos de categorias.	Descobrir novos padrões de comportamentos e investigar diferenças e semelhanças entre os grupos.
Associação – Mineração de Regras de associação – Mineração de Correlações – Mineração de Padrões Sequenciais – Mineração de Causas	Descobrir relações entre as variáveis.	Descobrir associações curriculares na sequência do curso; descobrir quais estratégias pedagógicas levam a uma aprendizagem mais robusta e eficaz.
Descoberta de Modelos	Um modelo de um fenômeno desenvolvido com predição, aglomeração, ou conhecimento de engenharia, é utilizado como um componente adicional em relação a predição.	Descoberta de relações entre os comportamentos dos estudantes e características dos estudantes ou variáveis contextuais; análise da questão de pesquisa em toda a grande variedade de contextos.
Destilação de dados para facilitar as decisões humanas	Os dados são refinação para permitir a um ser humano identificar ou classificar rapidamente características dos dados.	Identificação humana de padrões na aprendizagem dos estudantes, comportamento, ou de colaboração; dados de rotulagem para uso no desenvolvimento posterior do modelo de previsão.

Fonte: Adaptado de Baker (2011)

Na taxonomia de Baker (2011), os métodos de EDM mais utilizados para realizar atividades preditivas, utilizadas nesta dissertação são:

- **Classificação**: visa identificar à qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de 'aprender' como classificar um novo registro (aprendizado supervisionado). Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os estudantes de um determinado curso: abaixo da média, na média e acima da média. O modelo analisa os registros e então é capaz de dizer em qual categoria um novo estudante se encaixa.
- **Estimação/Regressão:** é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores das notas de diversos tipos estudantes de

acordo com os hábitos de cada um. Após ter analisado os dados, o modelo é capaz de dizer qual será a nota de um novo estudante.

Entretanto, todas as atividades/métodos, conforme mostrado no Quadro 10, podem ser utilizadas para fornecer informações necessárias para auxiliar a tomada de decisão em uma abordagem educacional. Por exemplo, a aplicação da atividade de **associação** sobre os dados armazenados ao longo do tempo num banco de dados pode considerar as atividades planejadas para o estudante e as atividades que o estudante efetivamente cumpriu, indicando o percentual de estudantes aprovados que cumpriram o cronograma de atividades dessa forma o professor pode ou não manter o cronograma de acordo com o índice descoberto.

Apesar das diversas subáreas, o processo de EDM converte dados brutos provenientes das bases de dados de sistemas educacionais em informações úteis, que podem ter um grande impacto na pesquisa e na prática educacional (ROMERO; VENTURA, 2010).

Todo o processo não difere muito das outras áreas de aplicação do KDD como nas áreas de negócios, genética, medicina, etc., pois segue os mesmos passos: pré-processamento, mineração de dados e pós-processamento (ROMERO et al., 2004). No entanto, é importante notar que KDD é um termo usado num sentido mais amplo do que EDM, que usa técnicas típicas de KDD voltadas para área educacional. Dentre os seus objetivos, estão o de fornecer informações úteis para os estudantes, professores, desenvolvedores de curso, gestores escolares e para a própria instituição de ensino. Assim como no KDD, no EDM existem atividades preditivas e descritivas sendo as preditivas abordadas no contexto dessa dissertação para predizer os estudantes propensos a evadir, fornecendo informações relevantes para tomada de decisão no combate ao fenômeno.

2.4 EVASÃO

Inicialmente devemos compreender o que vem a ser o fenômeno evasão. Apesar do foco do presente trabalho estar voltado para educação superior a distância, esse é um fenômeno internacional que atinge todos os níveis de ensino em instituições públicas ou privadas, nas modalidades presenciais e a distância. Suas perdas são consideradas desperdícios sociais, acadêmicos e econômicos. “No setor público, são recursos públicos investidos sem o devido retorno. No setor privado, é uma importante perda de receitas. Em ambos os casos, a evasão é uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico”. (SILVA FILHO et al., 2007).

Dore e Lüscher (2011 apud Santos 2014) dissertam sobre da ausência de uma defini-

ção única para termo evasão, tendo este sido associado a diversas situações, tal como "a retenção e repetência do estudante na escola, a saída do estudante da instituição, a saída do estudante do sistema de ensino, a não conclusão de um determinado nível de ensino, o abandono da escola e posterior retorno", portanto, na literatura, o termo evasão apresenta-se com diversas definições, conforme pode ser observado no Quadro 8.

Quadro 8 – Definições de Evasão

Autor	**Definição**
Gaioso (2005)	Interrupção no ciclo de estudos, em qualquer nível de ensino.
Brasil/MEC (2010)	Saída definitiva do estudante de seu curso de origem, sem concluí-lo
Kira (1998)	Perda ou fuga de estudantes da universidade.
Baggi (2011)	Saída do estudante da instituição antes da conclusão de seu curso.
Tinto (1975 apud SANTOS 2014)	Abandono voluntário (não motivado pelo fracasso escolar) e permanente.
Martins (2007)	Saída do estudante de uma IES ou de um de seus cursos de forma temporária ou definitiva por qualquer motivo, exceto a diplomação.
Laguardia e Portela (2009)	Saída do estudante de um curso ou programa educacional sem tê-lo completado com sucesso.
Favero (2006)	A evasão se caracteriza pela desistência do curso pelos estudantes.
Santos et al (2008)	**A evasão refere-se à desistência definitiva do estudante em qualquer etapa do curso.**

Fonte: do Autor

No estudo desenvolvido pela Comissão Especial de Estudos Sobre Evasão, Brasil/MEC (2010), admitiu-se o conceito de evasão como sendo "Saída definitiva do estudante de seu curso de origem, sem concluí-lo" entretanto, a própria Comissão ressalta que o termo evasão é ambíguo e deve-se questionar qual tipo de evasão está sendo pesquisado, ou seja, de qual evasão está se falando: se evasão de curso, ou evasão da instituição, ou evasão do próprio sistema educacional"

Para o presente trabalho, de acordo com a Brasil/MEC (2010), que indica que na pesquisa deve-se descrever como a evasão será observada, tomaremos com definição a contribuição Santos et al. (2008) que define a evasão como sendo a desistência definitiva do estudante em qualquer etapa do curso. Portanto, consideraremos o abandono + cancelamento no decorrer do curso para registro do índice de evasão. Optou-se por essa definição por ela está de acordo com a definição adotada pela Secretaria Acadêmica da Instituição.

Esse fenômeno vem sendo pesquisado de forma significativa nos últimos anos e é uma das principais preocupações das instituições de educação pois ela afeta diretamente a sustentabilidade, atingindo os diferentes níveis de ensino, sejam quais forem as denominações que eles tenham nesses diversos níveis" (CUNHA; MOROSINI, 2014).

2.4.1 Evasão no ensino superior

Segundo Ribeiro (2005), o tema evasão escolar tem aparecido com frequência nas discussões acerca da universidade, pois se tornou um fenômeno complexo e que está interferindo na gestão universitária por todo Brasil, seja em Instituições de Nível Superior (IES) públicas, seja em IES privadas.

> A evasão/abandono escolar tem se constituído, no âmbito da educação superior, numa temática "nova" e instigante, que tem conduzido diferentes estudiosos a enveredarem na busca de maiores informações e de dados consistentes que possam subsidiar de forma particular ou coletiva (nesse caso, as instituições de ensino) a adotar estratégias que busquem minimizar os efeitos danosos que o fenômeno causa tanto para os estudantes como para as instituições. (CUNHA; MOROSINI, 2014)

Apesar dos índices alarmantes de evasão no ensino superior no Brasil, no estudo desenvolvido por Silva Filho et al. (2007) aponta que as taxas brasileiras ficam um pouco abaixo dos países da América Latina e Central, ficando acima somente em relação a CUBA.

2.4.2 Evasão em EAD

A evasão na EAD tem se mostrado um desafio para gestores de IES, assim como para pesquisadores que buscam identificar as causas e encontrar maneiras de administrar sua contenção, ela é um dos fenômenos que mais atinge, preocupa e desafia essa modalidade de ensino (DAUDT; BEHAR, 2013).

Daudt e Behar (2013) apresentam resultados observados a partir dos dados do Censo EaD.br de 2009 e 2010, investigados a partir dos três primeiros anos de cursos de graduação, ao que chamam de período crítico do curso. É possível observar na Tabela 5 um significativo aumento dos índices de evasão de 2009 para 2010 e um decréscimo no passar dos anos curriculares, indicando uma criticidade maior no primeiro ano do curso.

Tabela 5 – Índice de evasão nos três primeiros anos de curso

Período do curso	Dados 2009	Dados 2010
1º ano	14,4%	17,4%
2º ano	5,4%	13,5%
3º ano	2,7%	8,3%

Fonte: Censo EaD.br 2009 e 2010.

Um dos fatores apontado pelos autores é o desconhecimento por parte dos estudantes sobre o funcionamento e o envolvimento necessário para estudar a distância. Muitos dos que procuram a modalidade o fazem por julgá-la mais fácil do que a presencial, pela comodidade

e pelo acesso facilitado, já que pode-se acessar o curso de qualquer lugar que se tenha conexão, sem sair do nosso espaço profissional ou familiar, contudo, a interação entre os pares (estudante-estudante) empobrece, favorecendo a desmotivação e consequentemente a evasão (MORAN, 2003).

Nesse sentido se faz importante investir em um monitoramento e acompanhamento mais próximo das disciplinas iniciais do curso. De fato, para a Educação a Distância, o monitoramento da evasão é uma tarefa importante para investigação das situações que levam a permanência e evasão do curso (OLIVEIRA et al., 2014).

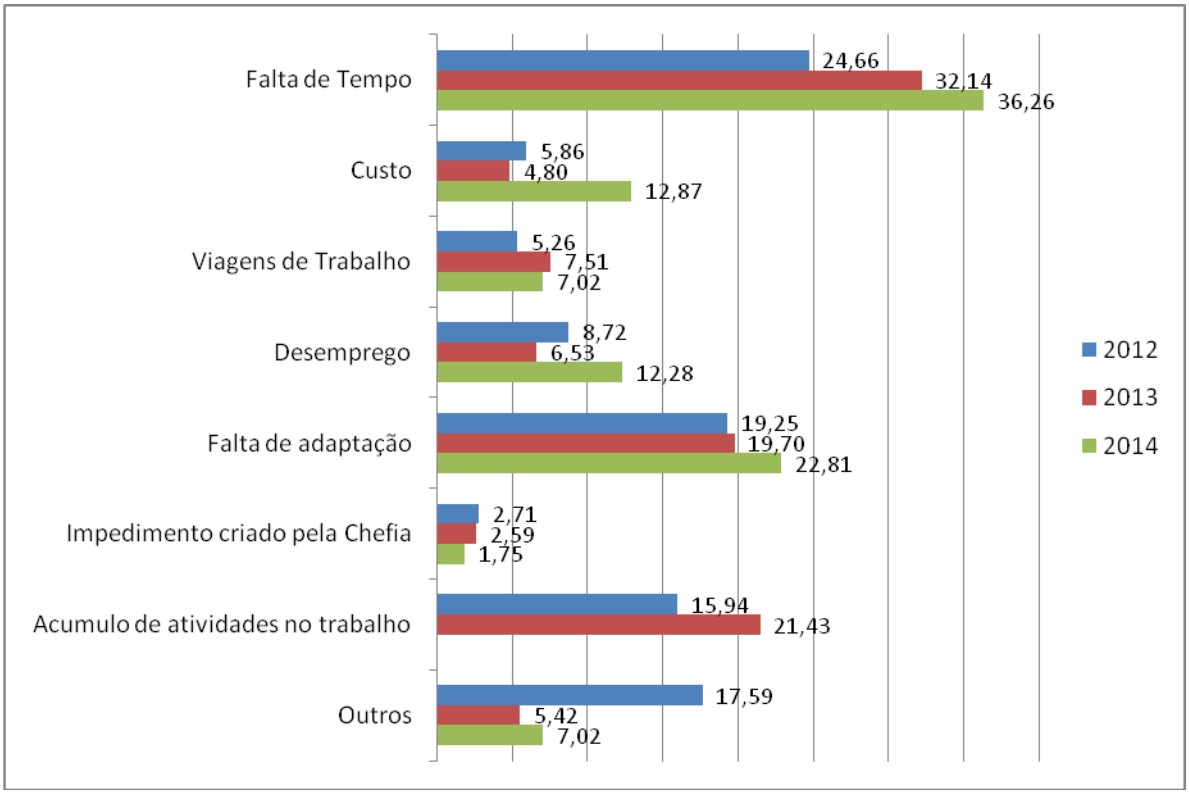
Conforme Fernandes (1996 apud MOORE; KEARSLEY, 2014), um monitoramento eficaz exige uma rede de indicadores que disponibilizem os dados necessários sobre o desempenho e atuação dos estudantes no AVA; isso precisa ser feito frequente e rotineiramente. Os dados precisam chegar aos professores e/ou gestores para que, baseados no monitoramento dos indicadores, possam tomar medidas preventivas de combate à evasão.

2.4.3 Fatores que levam à evasão no ensino superior EAD

Na modalidade EAD, o estudo realizado pelo Censo EaD.br (2014) revela as principais causas apontadas pelos estudantes para as instituições de ensino participantes do Censo EaD.br 2012, 2013 e 2014. Como a evasão constitui um grande obstáculo para o desenvolvimento das ações em EAD, o Censo buscou coletar dados sobre os índices de evasão observados pelas instituições participantes da ABED, mostrados no Gráfico 5.

Foram analisados e confrontados dados do Censo de três anos consecutivos na tentativa de buscar convicções a respeito dos motivos que levam a evasão. Contudo, ao analisar e confrontar os dados, percebeu-se a ausência do motivo “Acumulo de atividades no trabalho” em 2014. Analisando os dados obtidos no Censo EaD.br, destacam-se os motivos: falta de tempo, falta de adaptação, acumulo de atividades no trabalho e curso.

Gráfico 5 – Motivos que levaram a evasão em 2012, 2013 e 2014



Fonte: Censo EaD.br 2012, 2013 e 2014

Em relação à FALTA DE TEMPO, percebe-se um aumento gradativo, chegando a 11,6% de 2012 para 2014. De fato, a EAD possibilita mobilidade e flexibilidade de estudo nos processos formativos e isso está se transformando em referência para mudanças no ensino superior, mas, ironicamente, esses mesmos fatores (mobilidade e flexibilidade) estão também entre os possivelmente relacionados à evasão, uma vez que são frequentemente associados às expectativas geradas pelas falsas ideias de maior "facilidade" e "menor esforço" na EAD. Ajudar o estudante a compreender as dimensões do processo e do desafio com os quais estará envolvido torna-se fundamental para que ele consiga estabelecer um fluxo de estudo e assim realizar as leituras, interações e demais atividades acadêmicas envolvidas na EAD.

Os cursos de educação a distância são, em geral, mais baratos do que os cursos presenciais. Isto proporciona maior acesso para muitas camadas da população. Quanto ao indicador CUSTO, pode-se observar um aumento gradativo de 7,01% do ano de 2012 para 2014, e de acordo com o Censo EAD 2008, um dos itens que definem o perfil do estudante de educação a distância é a faixa salarial, o maior número de matriculados está entre 1 e 3 salários mínimos, chegando ao patamar de 26,7%.

A FALTA DE ADAPTAÇÃO AO MODELO EAD é um indicador alto, chegando a casa de 22,81% em 2014, mostrando uma pequena diferença em relação a 2009 de 3,56%. Segundo

Pretti (2003), na EAD, o estudante é o condutor da sua própria aprendizagem, ele deve provocar em si a motivação necessária para seguir adiante, saber gerenciar seu tempo, ter autodirecionamento, possuir um bom nível de leitura, ter capacidade para resolver problemas, ter disciplina e convicção do que quer aprender, etc. A falta dessas características no estudante EAD pode levá-lo à desistência do curso.

Quanto ao motivo ACÚMULO DE ATIVIDADES, ele aparece em 2012 e 2013 não aparecendo em 2014. Entretanto, podemos considerar que o acúmulo de atividade leva à ausência de tempo, levando a considerar esse motivo como falta de tempo.

Tresman (2002) e Oliveira Neto (2008), categorizam em suas pesquisas os fatores da evasão em extrínsecos e intrínsecos. Dentre as causas extrínsecas os autores destacam a não adequação ao modelo de aprendizagem da EAD, a carga horária de trabalho, dificuldade de priorizar horários para seus estudos e a incerteza da vocação (curso escolhido); e dentre as causas intrínsecas estão a falta de acompanhamento do professor-tutor e de apoio/incentivo institucional, a relevância do conteúdo para seus interesses, a qualidade do material, a usabilidade das tecnologias utilizadas, a metodologia empregada no curso, a forma de avaliação, a interação com professores e colegas e a quantidade, a natureza do feedback recebido com relação às tarefas realizadas e ao progresso no curso.

Correlacionando as pesquisas com o Censo EaD.br (2012, 2013 e 2014), podemos perceber que a falta de tempo e falta de adaptação ao modelo aparecem em primeiro e segundo lugar na pesquisa. No censo não são evidenciados como causa de evasão fatores intrínsecos ao curso. A falta de tempo evidenciada em primeiro lugar no Censo EaD.br (2012, 2013 e 2014), pode ser ainda maior, se somarmos ao seu índice os indicadores: viagens de trabalho e acúmulo de atividades no trabalho, já que tais indicadores incrementam no indicador falta de tempo.

Para analisar e confrontar os motivos que levam a evasão da instituição observada com os motivos da ABED, foi realizada uma pesquisa com os estudantes evadidos. O questionário foi emitido para 635 evadidos, com 51 repostas. Os resultados são apresentados na Tabela 6.

Destaca-se que ao item “Outros” foram atribuídas 14 respostas, entretanto, após a observação literal das respostas eles foram adicionados nas suas respectivas categorias, quais sejam: para o item “Falta de Adaptação para Estudar a Distância”, adicionou-se 4 respostas e para o item “Falta de Tempo”, adicionou-se 1 resposta, restando ao item “Outros” 9 respostas diversificadas.

Os índices obtidos estão próximos entre a instituição pesquisada e os da ABED, com exceção das categorias: Falta de Tempo e Outros.

Tabela 6 – Resultado da pesquisa: motivos da evasão sob o ponto de vista do estudante

Motivo	Percentual após redistribuição dos itens da instituição pesquisada
Falta de tempo	7,84%
Custo	19,61%
Viagens de trabalho	1,96%
Desemprego	9,80%
Falta de adaptação para estudar a distância	25,49%
Impedimentos criados pela chefia	0,00%
Acúmulo de atividades no trabalho	17,65%
Outros	17,65%

Fonte: do autor

Observa-se ainda que para as instituições pesquisadas pela ABED, o fator que mais influência para a evasão é a "Falta de Tempo" e para a instituição pesquisada a "Falta da Adaptação para Estudar a Distância", entretanto, percebe-se que é necessário uma amostragem maior e um questionário mais detalhado para uma observação mais abrangente e conclusiva.

2.5 PREDIÇÃO DE EVASÃO

A cada dia o número de estudantes e cursos cresce na modalidade EAD, com isso, um grande volume de dados é gerado diariamente seja pelo acesso dos estudantes/professores nos AVAs ou pelas diversas possibilidades de interação que a modalidade proporciona. Corriqueiramente, esses dados não são analisados, já que seu formato nativo armazenado nos bancos de dados é de difícil compreensão e acabam sendo deixados de lado. Portanto, é importante utilizar mecanismos capazes de auxiliar na análise dos dados de maneira automatizada, buscando descobrir conhecimento que possa ajudar a compreender e mitigar o fenômeno da evasão.

Com a aplicação do KDD na predição de evasão, os dados armazenados nos bancos de dados podem se transformar em informações úteis e consistentes a respeito dos estudantes propensos a evadir.

Para Fayyad (1996), no contexto do KDD, predição envolve o uso de algumas variáveis ou campos do banco de dados para prever valores desconhecidos ou futuras variáveis de interesse. Trazendo a definição de Fayyad para o contexto educacional, O KDD possibilita, encontrar padrões variados no comportamento dos estudantes e predizer, entre outras coisas, aqueles que estão propensos a evadir, permitindo adotar medidas preventivas.

Quanto mais preciso o padrão for e mais precoce for a sua aplicação para detecção dos estudantes propensos a evadir, mais cedo as ações preventivas podem ser executadas pelos envolvidos no combate a esse fenômeno.

Rigo et al. (2014) diz ser possível utilizar na EAD sistemas para a predição de evasão construídos a partir de diversas fontes de dados, tendo em vista a heterogeneidade de dados disponíveis muitas vezes em sistemas desenvolvidos separadamente, como por exemplo, Sistemas de Processo Seletivo, Sistemas de Gestão Acadêmico, Ambientes Virtuais de Aprendizagem, etc.

De modo geral, algumas etapas são necessárias para que a predição de evasão seja possível. Para tal finalidade, utilizaremos o modelo CRISP-DM para: *i)* compreender o funcionamento da instituição que fomenta a EAD, seu modelo educacional, ferramentas mais utilizadas no AVA; *ii)* compreender como os dados são armazenados no banco de dados; *iii)* preparar os dados, fazendo junções base de dados separadas a fim de aumentar a precisão dos resultados; *iv)* modelar/executar as técnicas de DM, para visualização dos primeiros resultados; *vi)* avaliar os resultados e implantar o modelo que apresentou a melhor acurácia.

Portanto, baseado nas etapas supracitadas, o modelo CRISP-DM será aplicado nos dados consolidados do AVA da instituição a fim de identificar padrões no comportamento dos estudantes que evadem, de forma que os comportamentos identificados possam ser confrontados com comportamento dos estudantes do semestre corrente, identificando os propensos a evadir e assim adotar ações preventivas no combate à evasão.

3. METODOLOGIA

Minayo (1993) considera a pesquisa como “atividade básica das ciências na sua indagação e descoberta da realidade. É uma atitude e uma prática teórica de constante busca que define um processo intrinsecamente inacabado e permanente. É uma atividade de aproximação sucessiva da realidade que nunca se esgota, fazendo uma combinação particular entre teoria e dados”.

Frente às considerações de Minayo, aborda-se nesta dissertação tanto o lado teórico (qualitativo), na verificação das pesquisas produzidas, quanto prático, caracterizado pela descoberta do conhecimento em uma base de dados (quantitativo), ou seja, é um experimento, fundamentado e aplicado.

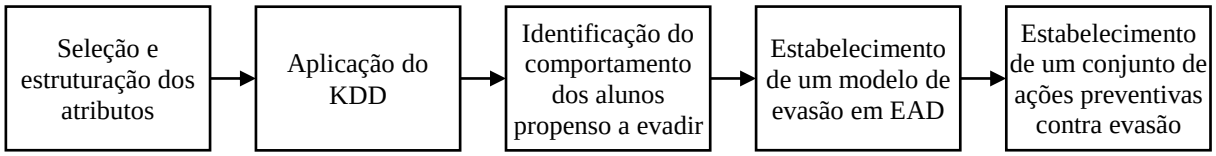
Para Moresi (2003), a pesquisa busca soluções para um problema, para isso, faz uso de um conjunto de ações, tendo como base, procedimentos criativos, racionais e sistemáticos, desenvolvendo métodos científicos para o enfrentamento de um problema para o qual se buscam possíveis respostas, portanto, como o trabalho foca o problema da evasão na instituição observada e não utiliza modelos presentes em outras pesquisas, dado que o contexto institucional e modelo de ensino-aprendizagem diferem dos trabalhos referenciados nessa pesquisa, caracterizamos essa pesquisa como experimental.

As próximas etapas descreverão em detalhes os procedimentos metodológicos utilizados no presente trabalho.

3.1 CLASSIFICAÇÃO DA PESQUISA

Quanto à abordagem esta pesquisa é quantitativa, já que se serve de recursos e técnicas com base em modelos estatísticos. Quanto à sua natureza, ela é aplicada, pois “objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Envolve verdades e interesses locais” (MORESI, 2003). Quanto aos objetivos, é exploratória, uma vez que, no contexto institucional estudado, “não há conhecimento acumulado ou sistematizado” sobre o fenômeno da evasão no contexto da EAD. Finalmente, quanto aos meios de investigação, esta pesquisa é experimental, pois busca controlar “variáveis independentes e observa as variações que tal manipulação e controle produzem em variáveis dependentes”. A Figura 5 representa a estrutura experimental desta pesquisa.

Figura 5 – Estrutura experimental da pesquisa



Fonte: do Autor

3.2 CONDUÇÃO DA PESQUISA

Para conclusão dessa pesquisa, as seguintes etapas e atividades foram necessárias:

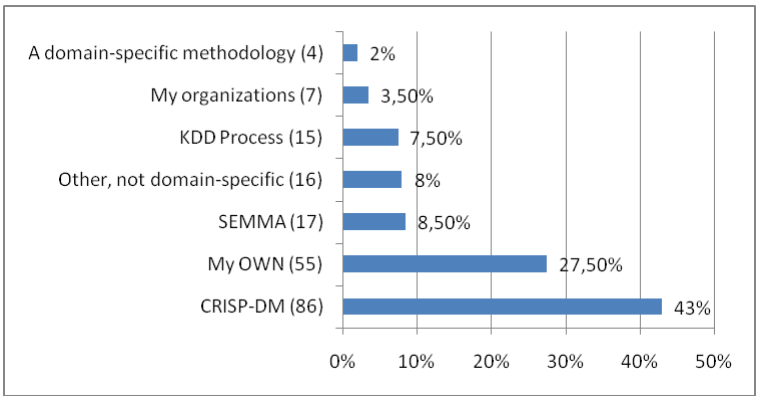
- **Etapa 1 –** As atividades agrupadas na primeira etapa consistem *(i)* na definição da linha de pesquisa, onde se procurou observar um fenômeno não tratado, passível recuperação e de interesse coorporativo. Após essa atividade buscou-se realizar *(ii)* pesquisa bibliográfica, para a fundamentação teórica e metodológica do trabalho em pauta (ALMEIDA, 2012). A pesquisa bibliográfica permeou todas as etapas do trabalho objetivando além do embasamento teórico a sustentação e contraposição de argumentos verificados no contexto da pesquisa. Após a pesquisa bibliográfica, definiu-se o *(iii)* tema da pesquisa, alinhando o interesse do pesquisador com linha de pesquisa e interesse coorporativo. Logo após foi iniciada a construção do *(iv)* Referencial Teórico, que forneceu embasamento sobre todos os assuntos tratados na pesquisa bem como para *(v)* a metodologia utilizada na pesquisa.
- **Etapa 2 –** Esta etapa consiste em selecionar as bases de dados que serão objeto de estudo. A pesquisa pretende observar o padrão de comportamento dos estudantes propensos a evadir. Para tal finalidade, serão selecionados os dados dos estudantes e suas interações do primeiro semestre de 2015, verificando o status dos mesmos estudantes no segundo semestre do mesmo ano. Com esta seleção será obtida a classe evadidos, alvo do presente trabalho e o comportamento dos estudantes poderão ser identificados através do KDD.
- **Etapa 3** – Nessa etapa os resultados dos experimentos realizados serão analisados sendo comparados entre eles. Inicialmente os resultados comparados através da Matriz de Confusão, posteriormente pela precisão e finalmente pela acurácia. Por fim será desenvolvido os Modelos de: (*i*) Predição de Evasão e (*ii*) Tomada de Decisão.

3.3 EXTRAÇÃO DE PADRÕES

A extração do conhecimento das bases de dados não é uma tarefa trivial, e para essa finalidade surgiram algumas metodologias que apoiam todo o processo, dentre elas a CRISP-DM que segundo Kdnuggets (http://www.kdnuggets.com) “descreve um modelo de referência a ser utilizado em projetos de KDD e impõe ao projeto um detalhado planejamento e uma rigorosa avaliação do processo em suas fases com o intuito de facilitar a organização, a compreensão e o controle dos eventos na coordenação do projeto”.

Ainda Kdnuggets (http://www.kdnuggets.com), “a metodologia CRISP-DM se revela como a mais usual e adequada para atender aos objetivos e aos problemas que envolvam DM e, por isso, é o preferido por 43% dos profissionais”, conforme comparativo do Gráfico 6.

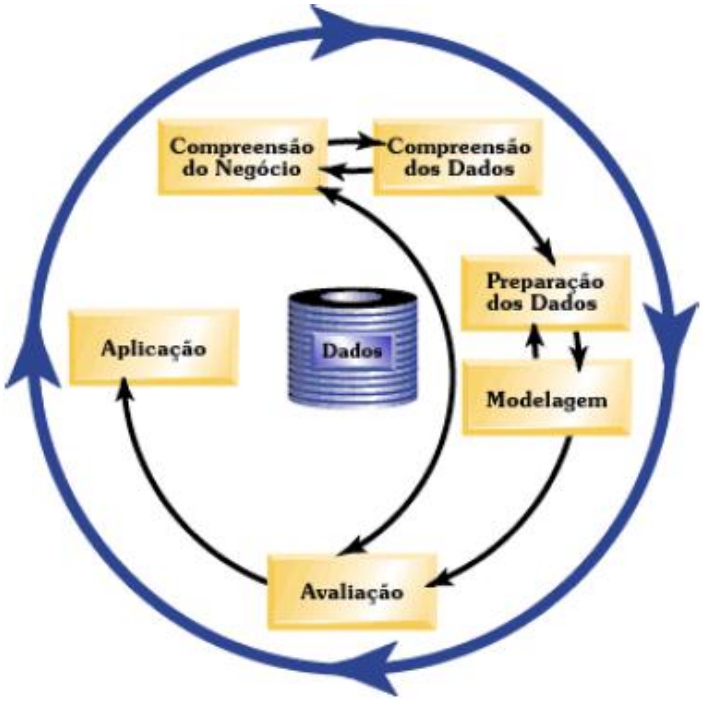
Gráfico 6 – Metodologias para DM mais utilizadas (2014)



Fonte: Kdnuggets (2016)

Para a investigação da base de dados na busca do comportamento do estudante propenso a evadir, foi utilizado esta metodologia, que é um padrão fornecedor de um roteiro para a desenvolvimento de um projeto de KDD. A metodologia CRISP-DM (2008) possui um conjunto de 6 fases que devem ser cumpridas para cobrir todo o ciclo de vida de um projeto de DM as quais são mostradas na Figura 6 e nos tópicos subsequentes. Pode-se observar que o círculo representa o ciclo normal do KDD enquanto as setas indicam a sequência e dependências importantes entre as fases da metodologia.

Figura 6 – Fases da metodologia CRISP-DM



Fonte: CRISP-DM, 2008

De forma breve, as fases do ciclo de vida do projeto podem ser assim descritas:

- **Compreensão do negócio** - Esta fase busca a compreensão dos objetivos do projeto e suas necessidades do ponto de vista dos negócios, ou seja, nessa fase é muito importante identificar fatores que possam influenciar os resultados de KDD. As tarefas realizadas nesta fase são: *(i)* determinação dos objetivos do negócio, *(ii)* avaliação da situação, *(iii)* determinar os objetivos da mineração de dados e *(iv)* produzir o plano de projeto.
- **Compreensão dos dados** – Nessa fase inicia-se a de coleta e investigação inicial dos dados. Também verifica-se aqui, a qualidade dos dados. Nesta fase, é comum a descoberta de padrões interessantes na pesquisa. Suas principais tarefas são: *(i)* coletar os dados iniciais, *(ii)* descrever os dados, *(iii)* explorar os dados e *(iv)* verificar a qualidade dos dados.
- **Preparação de dados –** É um dos aspectos mais importantes e muitas vezes demorados de mineração de dados. De fato, estima-se que a preparação de dados geralmente leva 50-70% de tempo e esforço de um projeto. O bom entendimento anterior do negócio e dos dados minimiza essa sobrecarga. As principais tarefas realizadas nesta fase são: *(i)* seleção dos dados, *(ii)* limpeza dos dados, *(iii)* construção dos dados, *(iv)* integração dos dados e *(v)* formatação dos dados.
- **Modelagem -** Nessa fase é escolhida a técnica de modelagem e os resultados começam

a lançar alguma luz sobre o problema do negócio. A modelagem é geralmente conduzida em múltiplas iterações. Normalmente nessa fase podem ser executados vários modelos usando os parâmetros padrão e, em seguida, afinar os parâmetros ou reverter para a fase de preparação de dados para manipulações exigidas pela escolha do modelo. As principais tarefas realizadas nesta fase são: *(i)* seleção da técnica de modelagem, *(ii)* geração de teste de desempenho, *(iii)* construção do modelo e *(iv)* avaliação do modelo.

- **Avaliação -** Neste ponto completa-se grande parte do projeto de *data mining*, e já se determinou, na fase de modelagem, que os modelos construídos são tecnicamente corretos e eficazes de acordo com os critérios de sucesso de mineração de dados que você definiu anteriormente, então, essa fase consiste basicamente em verificar se os modelos escolhidos atingiram os objetivos selecionados, revisando os passos seguidos no modelo. As tarefas realizadas nesta fase são: *(i)* avaliação dos resultados, *(ii)* revisão do processo e *(iii)* determinação dos próximos passos.
- **Implantação –** Nessa etapa é realizada uma comparação dos resultados com objetivos do projeto. A implantação também é o processo de usar seus novos conhecimentos para fazer melhorias em sua organização. Por exemplo, talvez se tenha descoberto um processo de KDD eficiente para descoberta preditiva de evasão dos cursos, estes resultados podem ser formalmente integrados em seus sistemas de informação para subsidiar as tomadas de decisão dos gestores. As principais tarefas realizadas nesta fase são: *(i)* planejamento do desenvolvimento, *(ii)* planejamento do monitoramento e a manutenção, *(iii)* produção do relatório final e *(iv)* revisão do projeto.

4. ESTUDO DE CASO

O estudo de caso foi orientado e desenvolvido segundo a metodologia CRISP-DM, abordando os aspectos: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação dos resultados e implantação do modelo.

4.1 COMPREENSÃO DO NEGÓCIO

Abordou-se nesse ponto o *background*, para melhor compreensão das necessidades da instituição, os objetivos de negócio e uma descrição dos critérios utilizados para determinar o sucesso para diminuição dos índices de evasão.

4.1.1 Cenário

A instituição observada implantou a EAD em 1996, com o estabelecimento do Núcleo EAD, entretanto, somente em 1997 os dois primeiros cursos foram lançados, ambos de pós-graduação *lato sensu*. Inicialmente os cursos foram oferecidos em material impresso, sendo a tutoria realizada por correspondência.

Em 2000 foi lançado o primeiro curso online, também de pós-graduação, sendo ministrado inteiramente por intermédio de um AVA, utilizando a Internet como meio. Em 2004 o MEC credenciou a instituição para ministrar cursos de graduação a distância e em 2005 houve o primeiro vestibular para os cursos de graduação *on-line*. Da sua criação até a data do desenvolvimento dessa pesquisa, a quantidade de cursos ofertados nas modalidades de extensão, pós-graduação e graduação vem aumentando significativamente, consolidando cada vez mais, na instituição, as três modalidades de EAD como alternativa educacional de qualidade.

Atualmente os cursos da modalidade EAD são ofertados em três níveis de ensino, graduação, pós-graduação e extensão. Dentre os cursos de graduação, são ofertados cursos de bacharelado, licenciatura e tecnológicos. Sendo que a evasão será pesquisada no curso tecnológico na área de Tecnologia da Informação, que tem a duração de cinco semestres.

Os cursos EAD estão vinculados a um modelo pedagógico que prima pela interação, entre os participantes do curso, contudo, apesar desse formato, o processo de identificação de estudantes propensos a evadir não é simples e depende principalmente da experiência e iniciativa dos coordenadores e professores.

Segundo Amaral (2007), a ausência ou pouca atenção do professor, da coordenação ou

dos monitores, aumenta a possibilidade de afastamento dos estudantes, gerando, frequentemente, o abandono dos cursos, observando esse fato, dentre as iniciativas da instituição pesquisada, pode-se citar o acompanhamento dos coordenadores, professores aos acessos dos estudantes feitos por intermédio de vários relatórios, dentre os quais o log de acesso e o de participação.

Os logs de acesso apresentam os registros detalhados de todas as atividades que os estudantes realizam. Ele registra cada clique realizado pelo estudante e tem um sistema de log. Os registros de log podem ser filtrados por curso, por participante, dia e atividade. O professor pode usar esses logs para determinar quem tem sido ativo no curso, o que eles fizeram e quando eles fizeram alguma atividade (ROMERO et al., 2007). Com esses registros pode-se verificar quais recursos disponíveis no curso que o estudante já visualizou. Porém, esse é um relatório extenso, tornando sua interpretação difícil. O relatório e extraído a partir do Moodle, conforme Figura 7.

Figura 7 – Relatório de log de acesso

Fonte: Captura de tela Moodle 2016

O Relatório da participação na disciplina possibilita o professor selecionar uma atividade da sala de aula e verificar se o estudante fez a atividade ou não. Posterior ao relatório, é possível selecionar os estudantes que não fizeram a atividade e enviar uma mensagem dirigida. Figura 8 e 9.

Com o relatório da participação nas atividades da disciplina, pode-se verificar a quantidade de vezes que o estudante participou de cada uma das atividades disponíveis, se ele submeteu às atividades solicitadas bem como se participou com alguma ou várias postagens em um fórum de discussão.

Figura 8 – Relatório da participação nas atividades da disciplina

Fonte: Captura de tela Moodle 2016

Figura 9 – Envio de mensagem para quem não fez a atividade

Fonte: Captura de tela Moodle 2016

Além dos relatórios supracitados, existem outros que podem ser facilmente obtidos que mostram a utilização do AVA pelos estudantes de forma pormenorizada, mostrando as atividades realizadas pelos estudantes de forma individualizada ou coletiva, mas para o acompanhamento e combate da evasão, os relatórios apresentados são os mais utilizados.

Visando colaborar no combate à evasão, a equipe técnica, desenvolveu um relatório gerado a partir da base de dados do AVA, que contempla de forma segmentada por curso/disciplina, informações sobre os estudantes ausentes por mais de X dias (a quantidade de dias é configurável). Esse relatório é enviado aos coordenadores de curso e esses, baseados nos dados constantes nos relatórios, tomam suas decisões de combate à evasão. Dessa forma, para cumprir os propósitos de combate à evasão é imprescindível o comprometimento do coordenador.

Diante da descrição acima, percebe-se a existência de esforço conjunto para combater a evasão, entretanto, esses dependem de ações segmentadas, tornando-se desejável o desenvolvimento e adoção de meios automatizados que realizem a detecção precoce dos estudantes com risco de evasão, e assim facilitando a disseminação das informações e as medidas preventivas que podem ser tomadas.

Ambiente Virtual de Aprendizagem

O AVA utilizado na EAD da instituição observada é o Moodle. Ele apresenta diversas ferramentas de comunicação, interação e avaliação. Armazenando várias informações dos usuários. Tais informações podem ser utilizadas por técnicas de mineração de dados para diversos fins, entre eles o de identificação do comportamento dos estudantes no AVA (BAR-

BOSA, 2014).

O AVA está organizado para que ocorram os processos de ensino-aprendizagem e principalmente a interação entre professor-estudante, estudante-estudante e estudante-recursos disponíveis.

Observa-se que nas salas de aula virtual, nas disciplinas dispostas no AVA, encontra-se diversos recursos dos quais os mais utilizados podem ser observados na Tabela 7. Esse quadro mostra o nome do recurso, a quantidade de vezes que eles foram inseridos no AVA, bem como o percentual em relação aos outros recursos, independente do curso ou disciplina.

Tabela 7 – Recursos mais utilizados no AVA

Recurso	Número de instâncias	Percentual de instâncias
Arquivo	11594	23,71%
Fórum	10746	21,98%
URL	9215	18,84%
Rótulo	7609	15,56%
Questionário	4729	9,67%
Tarefa	3478	7,11%
Chat	576	1,18%
Página	450	0,92%
Pasta	305	0,62%
Wiki	199	0,41%

Fonte: do Autor

Por intermédio dos relatórios disponíveis no AVA e descritos no *Background*, é possível acompanhar diretamente no AVA o acesso/ausência de cada estudante aos recursos apresentados no Quadro 17, entretanto, a interpretação dos relatórios não é elementar e torna-se dispendiosa para todos os envolvidos. O tempo gasto nessa atividade, caso seja automatizada, pode ser destinado a outras atividades necessárias, pedagógica ou administrativa conforme função do envolvido.

4.1.2 Objetivos do negócio

Ofertar com qualidade cursos na modalidade a distância, priorizando a aprendizagem, os processos interativos, a construção da autonomia.

4.1.3 Critérios para o sucesso

São apontadas nesse item algumas questões importantes para obtenção de sucesso do negócio:

- Sustentabilidade com o menor índice de evasão possível;
- Buscar inovações tecnológicas e pedagógicas a fim de obter satisfação dos estudantes e professores;
- Obter melhores notas no ENADE;
- Aumentar a quantidade de cursos com sustentabilidade;
- Criar meios para diminuir o esforço da docência virtual com ganho qualitativo por intermédio de automações tecnológicas.

4.1.4 Avaliação da situação

Inventário de recursos

Os recursos disponíveis para o projeto foram organizados em dois tópicos, quais sejam:

- Pessoal
  - Coordenador de curso para contextualização dos problemas de evasão e avaliação da efetividade do processo e predição de evasão;
  - Um analista para interagir com o software de mineração de dados;
  - Um DBA MySQL para extração dos dados do AVA.
- Hardware
  - Um notebook HP, Core i5 com 8GB de RAM;
  - Um notebook DELL, Core i5 com 4GB de RAM.
- Software
  - MySQL Server 5.5 – Servidor de Banco de Dados;
  - MySQL Workbench 6.3 – Ferramentas para manipulação de Banco de Dados;
  - Excel 2013 – Planilha Eletrônica;
  - Weka 3.7.13 – Ferramenta realização do KDD.

Requisitos e pressupostos

A execução deste projeto auxiliará o curso pesquisado a enfrentar o problema da evasão. Seu enfoque é detectar de forma preditiva os estudantes propensos a evadir, gerando informações para que os gestores possam tomar decisões.

Para atingir essas expectativas a pesquisa foi idealizada e desenvolvida por esse pesquisador, esperando que os resultados do KDD gerem informações suficientemente boas para

o gestor tomar decisões no enfrentamento da evasão.

Limitações

A base de dados analisada estende-se por dois períodos, 2015-1 e 2015-2, sendo o primeiro utilizado para extração dos dados variantes e não variantes no tempo e o segundo para verificar o status dos estudantes pesquisados. Foi utilizado somente um curso na área de Tecnologia da Informação com 110 estudantes.

Riscos e contingências

Riscos: (*i*) obter o comportamento dos estudantes propensos a evadir com baixo índice de precisão; (*iii*) a ferramenta de KDD ser de difícil compreensão não permitindo a extração dos resultados esperados.

Contingências: (*i*) inserir no contexto atributos que possam melhorar os índices obtidos. (*ii*) usar outra ferramenta para realização do KDD.

Custos e benefícios

Não houve necessidade de investimento financeiro, dado que se optou para esta pesquisa utilizar uma ferramenta Software Livre, o WEKA. Contudo, houve investimento em horas para pesquisa e desenvolvimento da parte teórica, aprendizado da base de dados, da ferramenta e da execução do KDD.

Um dos principais benefícios foi o aumento do conhecimento sobre o fenômeno evasão, a geração de um processo que possa identificar os estudantes propensos a evadir de forma preditiva a fim de fornecer informações relevantes para que os gestores possam executar ações para enfrentá-la.

4.1.5 Objetivos do KDD

Neste estudo, o KDD será aplicado com o objetivo de desenvolver um processo que detecte, através dos dados, o comportamento dos estudantes propensos a evadir.

4.1.6 Critérios de sucesso do KDD

O processo de KDD desenvolvido neste trabalho será considerado bem sucedido se as variáveis preditoras de evasão em EAD forem identificadas e o modelo preditivo gerado pos-

sibilitar a definição de ações efetivas para a mitigação da evasão na instituição estudada.

4.1.7 Avaliação inicial das ferramentas

Ferramenta WEKA

A ferramenta utilizada será o *Waikato Environment Knowledge Analysis* (WEKA) que é um conjunto de algoritmos de aprendizado de máquina poderoso e completo para execução das tarefas de mineração de dados. Ele contém ferramentas para pré-processamento dos dados, classificação, regressão, clusterização, regras de associação, e visualização, sendo adequado também para o desenvolvimento de novos sistemas de aprendizagem máquina (WAIKATO, 2015).

A opção pelo WEKA deu-se pelos seguintes motivos: *i*) a maior parte dos trabalhos pesquisados faz referência para esse software, *ii*) pela facilidade de aquisição e disponibilidade para download gratuitamente, devido à sua característica *Open Source*, e *iii*) disponibilidade de recursos estatísticos para comparação de desempenho dos algoritmos disponíveis.

Atividades, tarefas e técnicas de KDD

O KDD pode ser dividido em tarefas preditivas (classificação e regressão) e descritivas (associação, agrupamento e sumarização). Um projeto de KDD pode envolver uma ou mais tarefas. Essas, por sua vez, se servem de várias técnicas advindas da Estatística ou da Inteligência Artificial (aprendizagem de máquina). Os algoritmos relativos a essas técnicas estão implementadas e disponibilizadas por pacotes de análise de dados, como SPSS[23], Weka[24], R[25]. A combinação entre as tarefas, técnicas e algoritmos trarão os resultados do KDD. O Quadro 9, mostra de forma abrangente as principais tarefas, técnicas e algoritmos utilizados no KDD.

[23] http://www.ibm.com/analytics/us/en/technology/spss/

[24] http://www.cs.waikato.ac.nz/ml/weka/

[25] https://www.r-project.org

Quadro 9 – Tarefas, técnicas e algoritmos de KDD

TAREFA	DESCRIÇÃO	EXEMPLOS DE USO	TÉCNICA	DESCRIÇÃO	ALGORITMOS
Classificação	Desenvolve modelos que deduzam aspectos específicos dos dados, possibilitando que eles possam ser classificados.	– Classificar estudantes propensos a evadir; – Classificar estudantes propensos a reprovar; – Classificar pedidos de crédito; – Esclarecer pedidos de seguros fraudulentos; – Identificar a melhor forma de tratamento de um paciente.	Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes".	Algoritmo Genético Simples; Genitor, CHC; Algoritmo de Hillis; GA-Nuggets; GAPVMINER.
			Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.	CART, CHAID, C4.5, C5.0, Quest, ID-3, SLIQ e SPRINT.
			Raciocínio Baseado em Casos	Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança.	BIRCH, CLARANS e CLIQUE
			Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões.	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB.
Regressão	Usada para definir um valor para alguma variável contínua desconhecida.	– Prever a nota final do estudante; – Estimar o número de filhos ou a renda total de uma família; – Estimar o valor em tempo de vida de um cliente; – Prever a demanda de um consumidor para um novo produto.	Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.	CART, CHAID, C4.5, C5.0, Quest, ID-3, SLIQ e SPRINT.

TAREFA	DESCRIÇÃO	EXEMPLOS DE USO	TÉCNICA	DESCRIÇÃO	ALGORITMOS
Associação	Usada para determinar a associação entre itens, sejam eles recursos, produtos, etc.	– Prever quais itens são acionados após o outro; – Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado.	Descoberta de Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados.	Apriori, AprioriTid, AprioriHybrid, AIS, SETM e DHP.
Clusterização	Encontrar dados que se agrupam, dividindo-os em conjuntos.	– Agrupar estudantes com comportamento similar no AVA; – Agrupar estudantes por desempenho; – Agrupar estudantes por região, pais, etc.	Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes".	Algoritmo Genético Simples; Genitor, CHC; Algoritmo de Hillis; GA-Nuggets; GAPVMINER.
			Raciocínio Baseado em Casos	Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança.	BIRCH, CLARANS e CLIQUE
			Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões.	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB.

Fonte: do Autor

4.1.8 Plano do projeto

Nessa etapa, observa-se o plano de projeto conforme Quadro 10. Nele são apresentadas as principais atividades executadas segundo a metodologia CRISP-DM, sendo que os dados advindos do fim de uma etapa servem se subsídios para o início da próxima.

Quadro 10 – Plano do projeto

#	Fase
1	Entendimento do negócio
2	Entendimento dos dados
3	Preparação dos dados
3.1	- Extração
3.2	- Limpeza
3.3	- Transformação
4	Modelagem
5	Avaliação
6	Implantação

Fonte: do Autor

Dentre os trabalhos pesquisados sobre predição de evasão, utilizando técnicas de KDD, os classificadores que mais se destacaram foram listados nos Tabela 8. Nela, são apresentadas as acurácias mais elevadas dos trabalhos pesquisados, e por esse motivo, serão considerados, neste trabalho, esses mesmos classificadores: Bayes Net, SMO, J48, Naive Bayes e MultilayerPerceptron**,** a fim de verificar quais apresentarão os melhores resultados quando submetidos aos conjuntos de dados não variantes no tempo (dataset-1), variantes no tempo (dataset-2) e na fusão desses dois conjuntos.

Tabela 8 – Classificadores utilizados na pesquisa

Referência	Bayes Net	SMO	J48	Naive Bayes	Multilayer Perceptron
Amorim et al. (2008)	89,70%	91,25%	89,65%	–	–
Santana et al. (2015)	–	92,03%	86,46%	85,50%	90,86%

Fonte: Amorin et al. (2008) e Santana et al. (2015)

Os classificadores serão aplicados utilizando o método *Cross Validation*, que garante que cada instância do dado será utilizada pelo menos uma vez para treinamento e testes. Com esse método, os dados são divididos em k grupos e os classificadores são acionados k vezes. Em cada uma dessas vezes, um dos grupos serve para testes e os outros são utilizados para treinamento, garantindo que todos os grupos são utilizados uma vez para testes e k-1 vezes

para treinamento.

Os resultados dos experimentos dessa pesquisa serão avaliados com as métricas de qualidade de classificadores abaixo, detalhados no Anexo I (SILVA, 2016):

- **Acurácia**: percentual de classificações corretas em relação ao total de predições, ou seja, após a submissão dos dados aos classificadores supracitados, será observado o percentual de acertos e erros dos classificadores.
- **Matriz de confusão**: número de classificações corretas versus as classificações preditas para cada classe, que nessa pesquisa se resume em evadido e matriculado. Sendo possível verificar a quantidade de predições corretas e predições erradas com clareza e ainda se a ocorrência de erros foi privilegiada de falsos negativos ou falsos positivos.
- **Precisão**: calculada a partir da matriz de confusão dividindo as previsões positivas corretas pelo total de previsões positivas classificadas (positivos verdadeiros + falsos positivos), tornando essa forma de avaliação a mais relevante para o contexto desta pesquisa, haja vista que indicará os índices de acertos de previsão de evasão.

4.2 COMPREENSÃO DOS DADOS

4.2.1 Descrição dos dados

Buscou-se, nessa etapa, identificar informações relevantes para o estudo, bem com familiarização com os dados a serem utilizados quanto à quantidade, qualidade e utilidade. Com base no referencial teórico, foram analisados, selecionados e agrupados em três dimensões os atributos tidos, empiricamente, como relevantes, como mostrado no Quadro 11. Os dados foram extraídos do Sistema de Gestão Acadêmico - SGA, que compreende os dados do processo seletivo, e também AVA, objetivando obter todos os dados com o qual se trabalhou durante a pesquisa.

Quadro 11 – Dimensões e atributos iniciais do modelo de predição evasão

Pessoal	Interação	Acadêmica
– Sexo – Estado – Polo de EAD – Idade – Estado Civil	– Quantidade de acesso ao curso – Quantidade de acesso aos fóruns – Quantidade de participações nos fóruns – Quantidade de acesso ao plano de ensino – Quantidade de acesso aos conteúdos – Quantidade de acesso à lista de participantes – Quantidade de acesso ao recurso perfil – Quantidade de acesso ao quadro de notas – Quantidade de acessos aos exercícios	– Tipo de ingresso – Tipo de matrícula – Situação no período letivo – Período letivo do estudante – Quantidade de períodos cursados – Situação final

Fonte: do Autor

Processo Seletivo

No decorrer do semestre, abrem-se inscrições para os processos seletivos de graduação. O candidato que desperta interesse pelos cursos ofertados faz sua inscrição e preenche os dados: nome, data de nascimento, estado natal, naturalidade, CPF, sexo, nacionalidade, e-mail, endereço, cidade, estado, país, identidade, órgão emissor, curso desejado e polo. Tais dados posteriormente são migrados para o SGA.

Sistema de Gestão Educacional - SGA

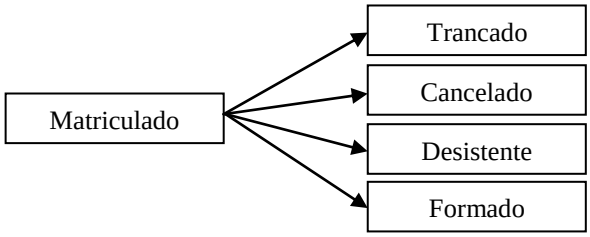
Após a aprovação no vestibular os estudantes são matriculados na instituição. O SGA guarda as informações do cadastro realizado via processo seletivo e os mantém em sua base de dados independente do status do estudante no curso, dentre os status pode-se citar:

- Matriculado: estudante que se encontra regular no curso;
- Trancado: estudante que solicita formalmente trancamento de matrícula;
- Cancelado: estudante que solicita formalmente cancelamento de matrícula;
- Desistente: estudantes que não abrem solicitação formal de trancamento ou cancelamento de curso e simplesmente se afasta do curso não renovando sua matrícula;
- Formado: estudantes que cumpriram todas as disciplinas do curso.

Observa-se que do status Matriculado, o estudante pode assumir qualquer um dos status conforme Figura 10. Conforme definição de evasão utilizada nesse trabalho, vamos consi-

derar os status Cancelado e Desistente para contabilização dos estudantes evadidos.

Figura 10 – Possíveis status dos estudantes



Fonte: do Autor

Ambiente virtual de aprendizagem

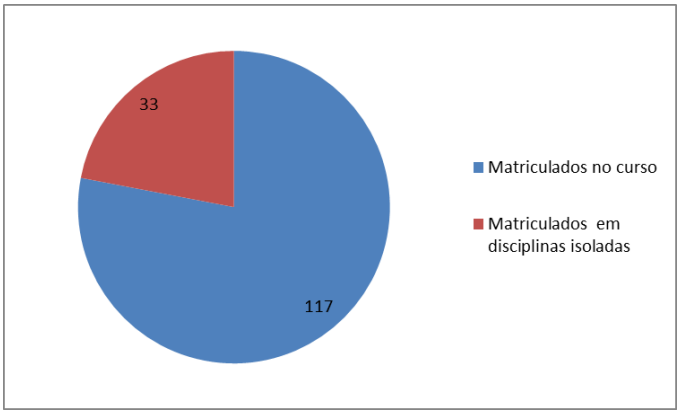
Embora a base de dados do Moodle acumule uma grande quantidade de informações, os registros não estão em um único lugar. Eles estão em tabelas distintas que armazenam dados tais como: identificação pessoal, interações e notas. Os dados são advindos dos acessos dos estudantes e professores aos diversos recursos disponíveis no AVA, bem como das notas alcançadas nas avaliações, exercícios, testes, tarefas as quais os estudantes são submetidos durante sua vida acadêmica.

Atualmente os dados do AVA estão armazenados em uma base de dados MySQL que possui 375 tabelas e para fazer a manipulação dos dados foi utilizado o conjunto de ferramentas da MySQL Workbench, uma ferramenta visual e unificada para Administradores de Banco de Dados e desenvolvedores de Banco de Dados (MYSQL, 2015). Com esse conjunto de ferramentas será possível realizar o pré-processamento e a modelagem dos dados. Dado sua característica de atualização constante, os dados do AVA são considerados variantes no tempo.

4.2.2 Relatório da coleta inicial dos dados

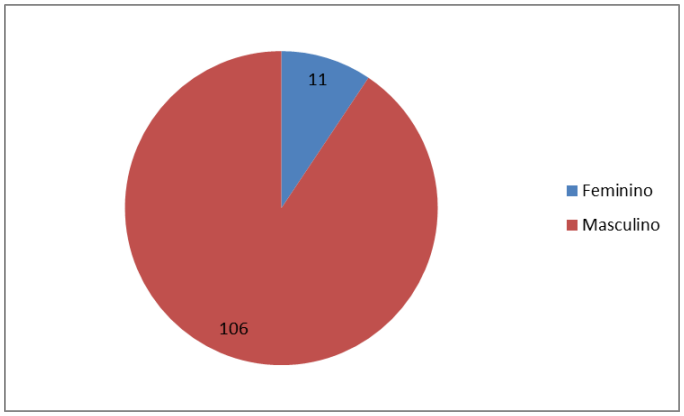
No período compreendido entre o 1º semestre de 2015 e o 2º Semestre de 2015, foram extraídos 150 registros de estudantes ativos/matriculados, dos quais 117 (78%) eram estudantes regulares do curso e 33 (22%) são estudantes do ensino presencial que optaram por alguma disciplina do curso na modalidade EAD. Quanto ao gênero, como mostrado no Gráfico 8, dos 117 estudantes regulares, apenas dez (9%) são do gênero feminino. Essa é uma característica peculiar nos cursos na área de tecnologia da informação independente da modalidade, EAD ou presencial.

Gráfico 7 – Distribuição dos estudantes por tipo (regular do curso EAD ou de disciplina isolada)



Fonte: do Autor

Gráfico 8 – Distribuição dos estudantes por gênero



Fonte: do Autor

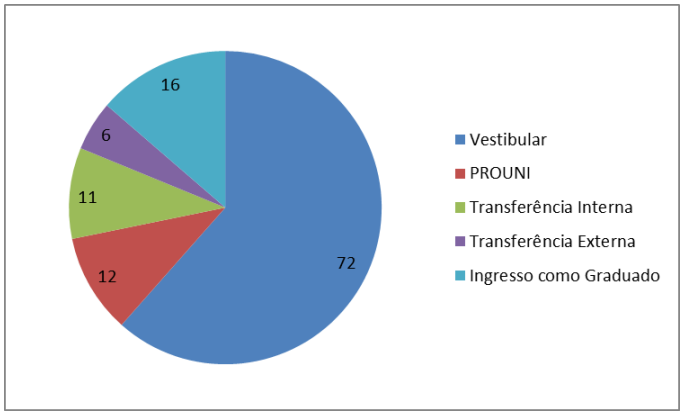
Dos 117 estudantes regulares, 72 (62%) ingressaram no curso via vestibular e 45 (38%) por outras formas: 12 (10%) pelo PROUNI, 11 (9%) por transferência interna, 6 (5%) por transferência externas e 16 (14%) já graduados, conforme mostrado no Gráfico 9. O vestibular continua sendo o grande portal de entrada para os cursos superiores, contudo, chama atenção o número significativo de estudantes já graduados. Nesse trabalho não se pretende aprofundar nessa questão, mas há indícios que a área de TI desperta atenção para pessoas graduadas, carecendo de maior detalhamento para saber as causas.

Quanto à distribuição por polo, entre os alunos regulares, 65 (56%) estão concentrados no Distrito Federal, e os demais estão geograficamente espalhados nos estados mostrados no Gráfico 10. Do restante, 27 (23%) estão nos demais estados e 25 (21%) estão fora do Brasil.

Quanto à distribuição por faixa etária, conforme mostrado no Gráfico 11, a maior concentração está entre 26 e 35 anos, totalizando 67 (57%) estudantes. 17 (15%) dos estudantes

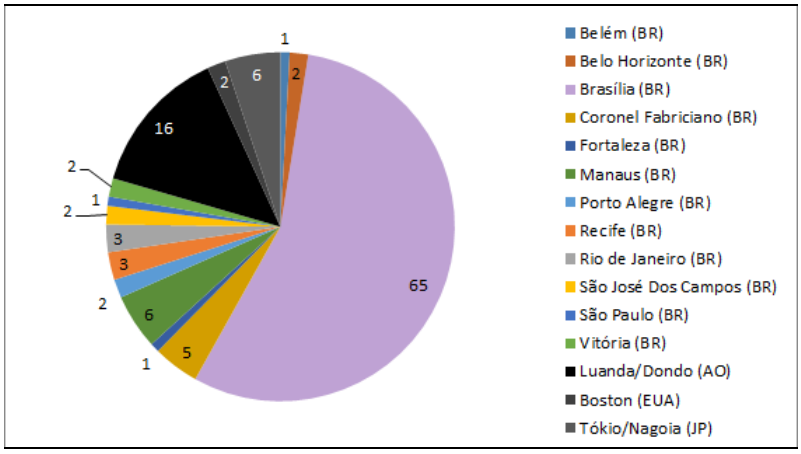
estão entre 20 e 25 anos, mostrando que existe um público jovem aderindo aos estudos nessa modalidade.

Gráfico 9 – Distribuição dos estudantes por tipo de ingresso



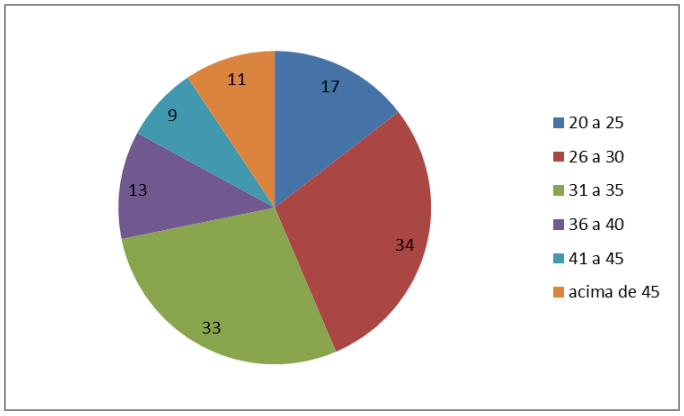
Fonte: do Autor

Gráfico 10 – Distribuição de estudantes por polo de EAD



Fonte: do Autor

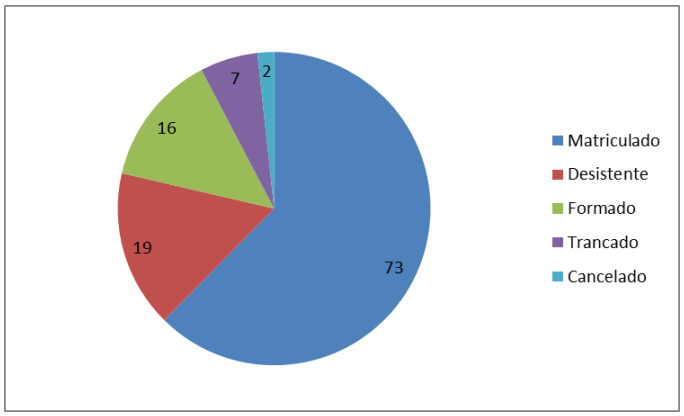
Gráfico 11 – Distribuição de estudantes por faixa etária



Fonte: do Autor

O Gráfico 12 foi construído a partir dos dados do 2º semestre de 2015. Nele percebe-se que dos 117 estudantes matriculados no 1º semestre de 2015, 73 (62%) continuaram matriculados, 19 (16%) desistiram do curso, 16 (14%) se formaram, sete (6%) trancaram e dois cancelaram, totalizando 21 (18%) estudantes evadidos (desistentes + cancelados).

Gráfico 12 – Situação dos estudantes 2015-1 em 2015-2



Fonte: do Autor

Assim, se estabeleceram dois grupos para estudo: (*i*) o primeiro composto por 73 estudantes que estavam matriculados no 1º semestre de 2015 e que continuaram matriculados no segundo semestre de 2015 e (*ii*) o segundo composto de 21 estudantes que estavam matriculados no 1º semestre de 2015 e evadiram no decorrer do semestre. Nesse caso, o estudo observa o comportamento dos dois grupos a fim de verificar no que divergem e predizer quais estudantes são propensos a evadir.

4.2.3 Qualidade dos dados

A baixa qualidade dos dados pode comprometer os resultados da pesquisa. Hand et al. (2001) destacam que a efetividade da mineração de dados depende criticamente da qualidade dos dados.

Com o tempo os sistemas são atualizados e novos campos podem ser criados para melhorar o controle sobre as informações do cadastro dos estudantes, entretanto, a inclusão de novos dados em versões mais recentes nos sistemas, pode gerar a ausência desse dado nos cadastros mais antigos. Em consequência dessa observação, entre outras, a etapa de validação da qualidade dos dados torna-se necessária.

Nos dados do SGA verificou-se: *(i)* a existência de 55 registros sobre o estado civil

não informados; *(ii)* dois estados natal não informados; *(iii)* nove estados natal informados como numeral; *(iv)* um atributo sexo não informado; *(v)* dez estados preenchidos como numeral; *(vi)* um estado não informado. Nos dados do AVA verificou-se: *(i)* as tabelas do AVA não são relacionadas, dificultando o entendimento de como os dados se relacionam; *(ii)* foram retirados do AVA dados da navegação dos estudantes com 30 dias a partir do início do curso[26]; *(iii)* alguns atributos foram retirados como NULL. Na etapa de preparação dos dados para execução do KDD, esses problemas foram tratados visando o não comprometimento dos resultados esperados.

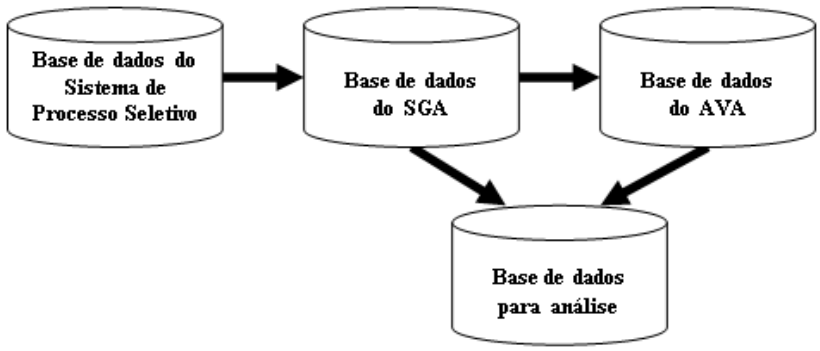
4.3 PREPARAÇÃO DOS DADOS

Nessa etapa foi gerado o conjunto de dados que foram submetidos à ferramenta de KDD, contudo, anteriormente, foi realizada a limpeza de dados inconsistentes e o tratamento dos dados faltantes, verificados e apontados na etapa de validação da qualidade dos dados.

Para melhor compreensão e transformação dos dados, eles foram organizados em uma planilha eletrônica, contendo nas linhas os estudantes e nas colunas os atributos selecionados.

Os dados foram selecionados a partir do SGA e AVA, sendo que a base de dados do SGA contém dados advindos do Sistema de Processo Seletivo, conforme mostrado na Figura 11. Os atributos foram selecionados tendo como base o referencial teórico e posteriormente organizados em uma planilha eletrônica para possibilitar as etapas de: limpeza, transformação e redução dos dados.

Figura 11 – Base de dados utilizada na pesquisa



Fonte: do Autor

[26] Acredita-se que no período de até 30 ainda é possível reverter possíveis evasões sem prejuízos acadêmicos para o modelo adotado pela instituição.

Conforme também indicado no referencial teórico, não existe um procedimento que possa ser aplicado de forma geral para todas as situações. Cada instituição tem uma forma singular de capturar e tratar os dados, vive particularidades e desenvolve modelos educacionais diferenciados. Esta pesquisa capturou os atributos das pesquisas relacionadas ao tema, conforme Quadro 12 e adaptou sua aplicação para a instituição pesquisada. Nesse sentido, foram gerados dois conjuntos de dados: o primeiro originou-se do SGA (dataset-1) e segundo originou-se do AVA (dataset-2). As estruturas desses datasets são apresentadas, respectivamente, nos Quadros 13 e 14.

No Quadro 13, os valores do atributo “idade” foram discretizados a partir do campo DTNASCIMENTO (Data de Nascimento) em faixas de 5 anos.

Após a importação dos dados para uma planilha eletrônica, eles foram submetidos ao WEKA separadamente e, posteriormente, os dados foram unificados. A extração das fontes distintas foi necessária para junção dos dados não variantes no tempo (SGA) e os dados variantes no tempo (AVA). Durante o processo de extração teve-se o cuidado de manter a integridade dos dados.

Não houve problemas significativos durante a extração dos dados, não havendo necessidade de limpeza e mesclagem, sendo que a dificuldade ficou a cargo do entendimento da base de dados, tabelas e seus relacionamentos.

Quadro 12 – Conjunto de atributos utilizados para predição de evasão.

Pasta (2011)	Veloso (2015)	Walter (2006)	Gottardo et al. (2012)
– **Cursos;** – **Sexo;** – **Idade;** – Ocupação; – **Estado Civil;** – Com quem mora; – Renda Mensal Familiar; – Ponto de Vista Financeiro; – Ensino Médio; – Meio de Atualização; – Conclusão do Ensino – Médio; – Razão Escolha Curso; – Razão Escolha IES; – Quem Indicou; – Divulgação e Atenção; – Meio mais informação; – Avaliação Site; – Avaliação Telefone; – Pós Curso.	– Tipo de Resp. Financeiro; – Financiamento; – Pronatec; – Condição do estudante; – Situação ocupacional. – Cod. IBGE da unidade de ensino; – Cod. IBGE do local de trabalho; – Cod. IBGE do estudante; – **Região;** – **Idade;** – Raça; – **Estado Civil;** – Naturalidade; – Nacionalidade; – Motivo da falta de CPF; – **Sexo;** – **UF;** – Grau de Instrução; – Matrícula Articulada; – Origem Escolar; – Carga horária do curso; – **Curso;** – Modalidade; – Área de conhecimento; – **Tipo de Entrada.**	– **Gênero;** – **Estado Civil;** – **Curso;** – **Idade.**	– **Nº de acesso ao AVA;** – **Nº de post no fórum;** – **Nº de respostas nos fóruns para outros participantes;** – **Nº de revisões do fórum;** – Nº participações no chat; – **Nº de questões respondidas;** – **Nº de questões acertadas;** – Frequência média de acesso; – Tempo médio de acesso; – Nº de dias desde o início do curso até o primeiro acesso; – Tempo total de acesso; – Nº de postagens que tiveram resposta de outro estudante; – Nº de respostas no fórum para outros estudantes; – Nº de mensagens recebidas de outros estudantes durante o curso; – Nº de mensagens enviadas para outros estudantes durante o curso; – Nº de postagens de estudantes que tiveram respostas feitas pelos professores; – Nº de postagens de professores que tiveram respostas de estudantes; – Nº de mensagens enviadas ao professor/tutor durante a realização do curso; – Nº de mensagens recebidas do professor/tutor durante o curso; – Nota final obtida pelo estudante.

Fonte: do Autor

Quadro 13 – Estrutura do dataset-1

Nome	Descrição
Tipo de ingresso	Vestibular / PROUNI / Transferência interna / Transferência externa / Ingresso como graduados
Tipo de matrícula	Calouro / Reabertura / Renovação / Retorno
Situação no período letivo	Cancelado / Desistente / Desligado / Formado / Matriculado / Trancado / Transferência interna
Período letivo do estudante	*numeral*
Quantidade de períodos cursados	*numeral*
Polo de educação a distância	***Polos no Brasil*** Belém / Belo Horizonte / Brasília / Coronel Fabriciano / Fortaleza / Manaus / Porto Alegre / Recife / Rio de Janeiro / São José dos Campos / São Paulo / Vitória ***Polos do exterior*** Angola / Estados Unidos / Japão
Sexo	M – Masculino / F – Feminino
Polo	***Brasil*** AM / CE / DF / MG / PA / PE / PI / RJ / RS / SC / SP ***Exterior*** Luanda/Dondo (AO) / Boston (EUA) / Tókio/Nagoia (JP)
Idade	Abaixo de 20 / 20 a 25 / 26 a 30 / 31 a 35 / 36 a 40 / 41 a 45 / Acima de 45
Estado civil	Casado / Solteiro / Outros / Não informado
Situação final	Evadido / Matriculado

Fonte: do Autor

Quadro 14 – Estrutura do dataset-2

Nome	Descrição
Quantidade de acessos à lista de participa	*numeral*
Quantidade de acessos ao curso	*numeral*
Quantidade de acessos ao recurso perfil	*numeral*
Quantidade de acessos ao quadro de notas	*numeral*
Quantidade de participações nos fóruns	*numeral*
Quantidade de acessos aos fóruns	*numeral*
Quantidade de acessos aos exercícios	*numeral*
Quantidade de acessos ao plano de ensino	*numeral*
Quantidade de acessos aos conteúdos	*numeral*
Classe	Evadido / Matriculado

Fonte: do Autor

4.4 MODELAGEM

Os dados foram extraídos formando dois datasets, ambos compostos por estudantes com status matriculados no semestre 2015-1 os quais em 2015-2 continuaram com o mesmo status (matriculado) ou assumiram um dos outros status possíveis: trancado, formado, cancelado ou desistente.

Observa-se que os dados do AVA foram extraídos 30 dias após o início das aulas. Tal

período foi selecionado por dois motivos: (*i*) por ser um período em que é possível fazer reversão de um estudante propenso a evadir sem prejuízos pedagógicos, segundo o modelo de EAD adotado pela instituição; (*ii*) por ser um período em que as navegações no AVA já geraram dados suficientes para execução do KDD.

A fim de verificar possíveis melhorias nos resultados do KDD, resolveu-se:

(i) Verificar os resultados dos classificadores utilizando separadamente os datasets 1 e 2 e posteriormente verificar os resultados com os datasets unificados;

(ii) Excluir de todos os experimentos os estudantes trancados, por não fazerem parte do conjunto de estudantes evadidos;

(iii) Testar os classificadores com e sem os estudantes formados.

Observa-se que os estudantes com o status formado em 2015/2 cursaram normalmente o semestre 2015/1. O intuito de, ora inseri-los, ora não inseri-los na massa de dados deu-se a fim de verificar possíveis melhorias nos resultados das classificações.

Dada as premissas acima, realizaram-se seis experimentos, conforme mostrado no Quadro 15. Nesse sentido, cabem as seguintes observações:

(i) Dada a definição de evasão adotada nesta pesquisa, os estudantes cancelados e desistentes foram transformados em evadidos;

(ii) A fim de verificar possíveis melhorias nos classificadores, os estudantes formados, participantes ativos no semestre, ora participaram do experimento sendo transformados para o status de matriculado, ora ficaram fora do experimento;

(iii) Os classificadores Bayes Net, SMO, Naive Bayes, J48 e Multilayer Perceptron que, segundo o referencial teórico obtiveram os melhores desempenhos para o tratamento de evasão, foram utilizados em todos os experimentos, contudo, foram verificados todos os outros classificadores disponíveis no WEKA.

Quadro 15 – Experimentos da pesquisa

Dataset	**Experimento**	**Status dos Estudantes**				
		Matriculado	**Formado**	**Cancelado**	**Desistente**	**Trancado**
Dataset-1	1	X		X	X	
	2	X	X	X	X	
Dataset-2	3	X		X	X	
	4	X	X	X	X	
Dataset-1 e 2	5	X		X	X	
	6	X	X	X	X	

Fonte: do Autor

Para os experimentos, foi necessário a transformação dos dados extraídos em arquivos no formato ARFF do WEKA. Tal arquivo é composto por “Nome do Relacionamento”, “Relação de Atributos e Valores Possíveis”, “Classe” e “Dados” conforme exemplo abaixo:

```
% Exemplo de arquivo ARFF para carga no WEKA
% Nome do Processamento
@Relation Evasão

%Relação de Atributos e Valores Possíveis
@attribute nota1 real
@attribute nota2 real
@attribute class {evasão, não_evasão}

%Dados
@Data
8.1,7.1,não_evadido
3.1,3.1,evadido
```

Dataset-1 (Dados do SGA)

A modelagem iniciou pela extração dos dados do SGA a fim de formar o dataset-1, nessa etapa, os dados foram organizados nos 5 (cinco) agrupamentos acadêmicos tal como representado no Tabela 9. Para essa organização foram selecionados todos os estudantes matriculados no curso em 2015-1 e posteriormente verificou-se sua situação acadêmica no semestre 2015-2, atribuindo a situação acadêmica aos dados de 2015-1, com isso, o dataset-1 foi criado com os dados: Tipo de Ingresso, Tipo de Matrícula, Situação no Período Letivo, Período Letivo do Estudantes, Quantidade de Períodos Cursados, Polo de Educação a Distância, Sexo, Estado, Idade, Estado Civil, Situação no Curso (dado de 2015-2).

Tabela 9 – Agrupamento dos estudantes por situação acadêmica

Situação	Quantidade
Total de estudantes matriculados em 2015-1	117
Total de estudantes que continuaram matriculados em 2015-2	73
Total de estudantes cancelados de 2015-1 para 2015-2	2
Total de estudantes desistentes de 2015-1 para 2015-2	19
Total de estudantes trancados de 2015-1 para 2015-2	7
Formados de 2015-1 para 2015-2	16

Fonte: do Autor

Experimento 1 (dataset-1)

O experimento deu-se com as condições abaixo:

(i) Status dos estudantes utilizados foram matriculado, cancelado e desistente;

(ii) Foram retirados os estudantes com status de trancado, por não fazerem parte do quanti-

tativo de estudantes evadidos;

(iii) Foram retirados os estudantes formados, para verificar possíveis melhorias nos classificadores;

(iv) Os estudantes com status cancelados e desistentes foram transformados em evadido.

Dados os critérios acima estabeleceu-se a classe com os status de evadido (21 estudantes) e matriculado (73 estudantes), totalizado 94 estudantes.

Aplicando os classificadores, foram obtidos os resultados da Tabela 10. Destacam-se no experimento os classificadores Bayes Net e SMO, que chegaram a acurácia de 79,79%, contudo, a matriz de confusão mostra que os classificadores identificaram corretamente 4 de 21 evadidos.

Tabela 10: Resultados obtidos do experimento 1

Classificador	**Classificação**	**Sumário**		**Matriz de Confusão**		**Precisão do Classificador**
		Quant.	**Percentual**	**Evadido**	**Matriculado**	
Bayes Net	**Correta**	**75**	**79,79%**	**4**	**71**	**0,19%**
	Incorreta	**19**	**20,21%**	**17**	**2**	
SMO	**Correta**	**75**	**79,79%**	**4**	**71**	**0,19%**
	Incorreta	**19**	**20,21%**	**17**	**2**	
Naive Bayes	Correta	72	76,60%	2	70	0,10%
	Incorreta	22	23,40%	19	3	
J48	Correta	72	76,60%	0	72	0%
	Incorreta	22	23,40%	21	1	
Multilayer Perceptron	Correta	72	76,60%	4	68	0,19%
	Incorreta	22	23,40%	17	5	

Fonte: do Autor

Experimento 2 (dataset-1)

O experimento deu-se com nas seguintes condições:

(i) Status dos estudantes utilizados foram: matriculado, formado, cancelado e desistente;

(ii) Foram retirados os estudantes com status de trancado, por não fazerem parte do quantitativo de estudantes evadidos;

(iii) Os estudantes com status cancelados e desistentes foram transformados em evadidos;

(iv) Os estudantes formados foram transformados em matriculados para verificar possíveis melhorias nos classificadores.

Dados os critérios acima, estabeleceu-se a classe com os status de evadido (21 estudantes) e matriculado (89 estudantes), totalizado 110 estudantes. Aplicando os classificadores, foram obtidos os resultados da Tabela 11. A maior acurácia obtida ficou a cargo do classificador Multilayer Perceptron, que chegou a acurácia 83,64%, sendo ele o classificador com mai-

or número de acertos na matriz de confusão quanto aos evadidos, 8 de 21.

Tabela 11: Resultados obtidos do experimento 2

Classificador	Classificação	Sumário		Matriz de Confusão		Precisão do Classificador
		Quant.	Percentual	Evadido	Matriculado	
Bayes Net	Correta	89	80,91%	3	86	0,14%
	Incorreta	21	19,09%	18	3	
SMO	Correta	91	82,73%	4	87	0,19%
	Incorreta	19	17,27%	17	2	
Naive Bayes	Correta	84	76,36%	3	81	0,14%
	Incorreta	26	23,64%	18	8	
J48	Correta	88	80,00%	0	88	0%
	Incorreta	22	20,00%	21	1	
Multilayer Perceptron	**Correta**	**92**	**83,64%**	**8**	**84**	**0,38%**
	Incorreta	**18**	**16,36%**	**13**	**5**	

Fonte: do Autor

Dataset-2 (dados do AVA)

Experimento 3 (dataset-2)

Para esse experimento foram utilizados os mesmos critérios do Experimento 1. Aplicando os classificadores, foram obtidos os resultados da Tabela 12. Realizado o experimento, a acurácia alcançada foi de 58,51%, contudo, a matriz de confusão mostra que o classificador Naive Bayes, classificou corretamente 17 dos 21 estudantes propensos a evadir. Tomando o conjunto dos resultados mostrados (acurácia e matriz de confusão), esse experimento foi o que revelou o melhor resultado para o propósito desta pesquisa.

Tabela 12: Resultados obtidos do experimento 3

Classificador	Classificação	Sumário		Matriz de Confusão		Precisão do Classificador
		Quant.	Percentual	Evadido	Matriculado	
Bayes Net	Correta	70	74,47%	1	69	0,05%
	Incorreta	24	25,53%	20	4	
SMO	Correta	73	77,66%	0	73	0%
	Incorreta	21	22,34%	21	0	
Naive Bayes 3	**Correta**	**55**	**58,51%**	**17**	**38**	**0,81%**
	Incorreta	**39**	**41,49%**	**4**	**35**	
J48	Correta	71	75,53%	0	71	0%
	Incorreta	23	24,47%	21	2	
Multilayer Perceptron	Correta	68	72,34%	0	68	0%
	Incorreta	26	27,66%	21	5	

Fonte: do Autor

Experimento 4 (dataset-2)

Para esse experimento foram utilizados os mesmos critérios do Experimento 2. Apli-

cando os classificadores, foram obtidos os resultados da Tabela 13. Apesar do classificador SMO e Multilayer Perceptron chegarem na acurácia de 80,9091% o classificador que mais se destaca para a finalidade desta pesquisa é o Naive Bayes que, apesar da acurárica chegar somente a 57,27%, ele classificou corretamente 17 dos 21 estudantes propensos a evadir. A quantidade de acertos se assemelha ao do Experimento 3, contudo, a acurácia do experimento 3 foi levemente melhor, chegando a 58,51%.

Tabela 13: Resultados obtidos do experimento 4

Classificador	**Classificação**	**Sumário**		**Matriz de Confusão**		**Precisão do Classificador**
		Quant.	**Percentual**	**Evadido**	**Matriculado**	
Bayes Net	Correta	87	79,09%	0	87	0%
	Incorreta	23	20,91%	21	2	
SMO	Correta	89	80,91%	0	89	0%
	Incorreta	21	19,09%	21	0	
Naive Bayes	**Correta**	**62**	**57,27%**	**17**	**46**	**0,8095%**
	Incorreta	**48**	**42,73%**	**4**	**43**	
J48	Correta	86	78,18%	0	86	0%
	Incorreta	24	21,82%	21	3	
Multilayer Perceptron	Correta	89	80,91%	0	89	0%
	Incorreta	21	19,09%	21	0	

Fonte: do Autor

Dataset-1 e Dataset-2

Para os experimentos 5 e 6, os datasests 1 e 2 foram unificados a fim de verificar a ocorrência de melhoria na acurácia dos classificadores.

Experimento 5 (dataset-1 e 2)

Para esse experimento foram utilizados os mesmos critérios do Experimento 1 e 3. Aplicando os classificadores, foram obtidos os resultados da Tabela 14. Realizado o experimento, as maiores acurácias foram obtidas pelos classificadores SMO e J48 com 77,66%, entretanto, o classificador Naive Bayes se mostrou superior ao demais classificadores, acertando 16 dos 21 estudantes propensos a evadir.

Tabela 14: Resultados obtidos do experimento 5

Classificador	Classificação	Sumário		Matriz de Confusão		Precisão do Classificador
		Quant.	Percentual	Evadido	Matriculado	
Bayes Net	Correta	68	72,34%	5	63	0,31%
	Incorreta	26	27,66	16	10	
SMO	Correta	73	77,66%	4	69	0,19%
	Incorreta	21	22,34%	17	4	
Naive Bayes	**Correta**	**57**	**60,64%**	**16**	**41**	**0,76%**
	Incorreta	**37**	**39,36%**	**5**	**32**	
J48	Correta	73	77,66%	0	73	0%
	Incorreta	21	22,34%	21	0	
Multilayer Perceptron	Correta	75	79,79%	8	67	0,38%
	Incorreta	19	20,21%	13	6	

Fonte: do Autor

Experimento 6 (dataset-1 e 2)

Para esse experimento, foram utilizados os mesmos critérios do Experimento 2 e 4. Aplicando os classificadores, foram obtidos os resultados apresentados na Tabela 15. A maior acurácia foi obtida pelo classificador SMO, com 82,73%. Entretanto, o classificador Naive Bayes foi o que mais se destacou, reconhecendo 15 dos 21 estudantes propensos a evadir.

Tabela 15: Resultados obtidos do experimento 6

Classificador	Classificação	Sumário		Matriz de Confusão		Precisão do Classificador
		Quant.	Percentual	Evadido	Matriculado	
Bayes Net	Correta	85	77,27%	3	82	0,14%
	Incorreta	25	22,73%	18	7	
SMO	Correta	91	82,73%	4	87	0,19%
	Incorreta	19	17,27%	17	2	
Naive Bayes	**Correta**	**68**	**61,82%**	**15**	**53**	**0,71%**
	Incorreta	**42**	**38,18%**	**6**	**36**	
J48	Correta	86	78,18%	0	87	0%
	Incorreta	24	21,82%	21	2	
Multilayer Perceptron	Correta	90	81,82%	10	80	0,48%
	Incorreta	20	18,18%	11	9	

Fonte: do Autor

4.5 DISCUSSÃO DOS RESULTADOS

Como já dito, os experimentos foram realizados com dados de duas fontes distintas, Sistema de Gestão Acadêmico (SGA), para utilização dos dados “não variantes no tempo”; e Ambiente Virtual de Aprendizagem (AVA), para utilização de “dados variantes no tempo”.

Do SGA, foram extraídos inicialmente 10 atributos: sexo, estado, polo de ead, idade, estado civil, tipo de ingresso, tipo de matrícula, situação no período letivo, período letivo do

estudante, quantidade de períodos cursados; contudo, após submissão dos atributos à função do WEKA que avalia a relevância ou não dos atributos, os mesmos foram reduzidos para 3 atributos: tipo de matrícula, período letivo do estudante e estado.

Do AVA foram extraídos inicialmente nove atributos: quantidade de acesso ao curso, quantidade de acesso aos fóruns, quantidade de participações nos fóruns, quantidade de acesso ao plano de ensino, quantidade de acesso aos conteúdos, quantidade de acesso à lista de participantes, quantidade de acesso ao recurso perfil, quantidade de acesso ao quadro de notas, quantidade de acesso aos Exercícios; entretanto, após a submissão da relevância dos atributos via WEKA, foram utilizados cinco atributos: quantidade de acessos ao curso, quantidade de acesso aos fóruns, quantidade de acesso ao plano de ensino, quantidade de acesso aos conteúdos, quantidade de acesso à lista de participantes.

Ao todo foram realizados seis experimentos com cinco classificadores em cada experimento, totalizando 30 testes, sendo que os melhores resultados de cada experimento são mostrados na Tabela 16.

Tabela 16: Melhores resultados obtidos nos seis experimentos

Experimento	**01**	**02**	**03**	**04**	**05**	**06**
Classificador	BayesNet / SMO	Multilayer Perceptron	Naive Bayes	Naive Bayes	Naive Bayes	Naive Bayes
Estudantes formados	Não	Sim	Não	Sim	Não	Sim
Dados	Não variantes no tempo	Não variantes no tempo	Variantes no Tempo	Variantes no Tempo	Não variantes e variantes	Não variantes e variantes
Acurácia	79,7872%	83,6364%	58,5106%	57,2727%	60,6383%	61,8182%
Classificação correta dos evadidos	4 de 21	8 de 21	17 de 21	17 de 21	16 de 21	15 de 21
Precisão	0,1904%	0,3809%	0,8095%	0,8095%	0,7619%	0,7142%

Fonte: do Autor

Nos experimentos, foram consideradas três métricas de qualidade: matriz de confusão, precisão e acurácia[27]. A primeira observação que se faz é em relação ao tipo de classificadores. Percebe-se nos melhores resultados a predominância de classificadores Bayesianos: Bayes Net e Naive Bayes, totalizando cinco entre os seis classificadores que apresentaram os melhores resultados, sendo que o Naive Bayes aparece quatro vezes e o Bayes Net, uma vez.

Outra observação é em relação à inclusão dos estudantes formados no experimento 2,

[27] Para definições das métricas, conferir página 69.

usando os dados “não variantes no tempo”. Tal inclusão apresentou 50% de melhora na quantidade e acertos dos estudantes propensos a evadir, saindo de 4 acertos de 21 prováveis no primeiro experimento, para 8 acertos de 21 prováveis no segundo experimento. Apesar da melhora, os experimentos mostram a dificuldade da predição de evasão utilizando somente os dados não variantes no tempo.

Em relação aos experimentos utilizando dados “variantes no tempo”, os resultados apresentados foram superiores aos experimentos com dados “não variantes no tempo”, sendo que a inclusão dos estudantes formados nos experimentos não apresentou ganho significativo na quantidade de acertos dos estudantes propensos a evadir.

Em relação a acurácia dos resultados, a maior conseguida, não representou o melhor resultado para a pesquisa. A maior acurácia de todos os experimentos foi obtida no experimento 2, chegando a 83,64%, entretanto, a quantidade de acertos dos estudantes propensos a evadir foi de apenas 8 dos 21, já nos experimentos 3 e 4, apesar de uma menor acurácia, 58,51% e 57,27% respectivamente, ambos experimentos classificaram corretamente 17 dos 21 estudantes propensos a evadir.

Em relação à quantidade de acertos dos estudantes propensos a evadir, dois experimentos se destacaram: os experimentos 3 e 4. Em comum, ambos sugiram a partir dos dados “variantes no tempo”, ou seja, extraídos do AVA e do classificador Naive Bayes. Os experimentos classificaram corretamente 17 dos 21 estudantes propensos a evadir com a “precisão”, métrica de qualidade utilizada para aferir os resultados obtidos, de 0,81%.

Com a união dos dados “não variantes tempo” e “variantes no tempo” esperava-se que os experimentos 5 e 6, apresentassem melhora tanto na classificação dos estudantes propensos a evadir quanto na precisão, contudo, houve uma discreta piora em relação aos experimentos que utilizaram somente os dados “variantes no tempo”. Com a união dos dados, foi classificado corretamente 16 estudantes dos 21 propensos a evadir, chegando a precisão de 0,76%.

Com os resultados obtidos, observa-se que foi possível predizer os estudantes propensos a evadir, com cinco atributos variantes no tempo, extraídos do AVA, com 30 dias de início das aulas, obtendo uma precisão de 0,81%.

4.6 APLICAÇÃO

Utilizar a metodologia proposta no estudo de caso, no ambiente de produção, após 30

dias do início do semestre, permitirá a viabilidade e a utilidade prática da metodologia desenvolvida para predizer os estudantes propensos a evadir e a tomada de decisões e ações no combate a evasão. Nesse ponto, surgem dois modelos, sendo que o primeiro fornece informações para o segundo. São eles: (*i*) modelo de predição de evasão baseado em dados do AVA, (*ii*) modelo para tomada de decisão e ações de combate a evasão.

O resultado da utilização dos modelos tende a dirimir o fenômeno da evasão na instituição, tornando-os, dessa forma, recursos importantes à disposição dos gestores.

4.6.1 Modelo de predição de evasão baseados em dados do AVA (Modelo 1)

O modelo apresentado no Quadro 16 foi desenvolvido a partir do Experimento 3. Ele utiliza cinco variáveis extraídas do AVA, identificadas como as mais relevantes após análise. No experimento 3, o classificador *Naive Bayes* conseguiu classificar 17 estudantes propensos a evadir do total de 21, chegando na precisão de 0,81%. Com a aplicação do modelo, os resultados apresentados são passados para que os tomadores de decisão possam iniciar o Modelo de Tomada de Decisão e Ações de Combate a Evasão.

Quadro 16: Modelo de predição de evasão em EAD

Passos	**Ação**	**Detalhamento**	**Objetivo**
1	Extração dos dados para mineração de dados.	**Atributos**: – quantidade de acessos ao curso; – quantidade de acesso aos fóruns; – quantidade de acesso ao plano de ensino; – quantidade de acesso aos conteúdos; – quantidade de acesso à lista de participantes.	Extrair do banco de dados do AVA os dados necessários para mineração de dados.
2	Formatação dos dados.	– Formatar os dados no formato ARFF.	Deixar os dados no formato adequado para mineração de dados.
3	Aplicação da Mineração de Dados.	– Classificador Naive Bayes.	Obter informações para combate da evasão.

Fonte: do Autor

4.6.2 Modelo para tomada de decisão (Modelo 2)

Com a aplicação do modelo apresentado no Quadro 16, geraram-se informações para

que os tomadores de decisão, providos das informações, possam tomar decisões e iniciar ações de combate à evasão. Para o acompanhamento dessas ações durante o semestre, sugere-se o modelo do Quadro 17. As ações desse podem conter elementos tais como: (*i*) colocar-se à disposição dos estudantes com alguma dificuldade, (*ii*) solicitar maior empenho dos estudantes que não acessam com frequência o AVA, (*iii*) lembrar dos prazos das atividades e encontros presenciais avaliativos, (*iv*) estimular a participação dos estudantes nas atividades acadêmicas, e /ou (*vii*) envio de informações incentivadoras estimulando a motivação e vínculo do estudante, (*viii*) provocar os professores para que estimulem os estudantes a participaram das atividades propostas nas disciplinas, (*ix*) premiar professores com menores índices de evasão.

Quadro 17: Modelo para tomada de decisão e ações de combate da evasão

Passos	**Ação**	**Objetivo**
1	Acompanhamento das matrículas no curso.	Obter e registrar a quantidade de estudantes matriculados no semestre.
2	Recebimento das informações geradas no Modelo 1	Receber informações para facilitar o processo decisório e inicio das ações.
3	Análise das informações.	Avaliação das informações geradas a partir do Modelo 1.
4	Ação 1 - E-mail convite para retorno às atividades acadêmicas.	Obter retorno dos estudantes nas atividades acadêmicas.
5	Ação 2 – E-mail convite para que os professores estimulem o acesso dos estudantes.	Obter maior engajamento do professor para combater a evasão.
6	Após 7 dias, solicitar nova execução do Modelo 1.	Gerar novas informações sobre os estudantes propensos a evadir.
7	Análise das informações.	Avaliação das informações geradas a partir do Modelo 1.
8	Cruzar as informações para obter os casos persistentes, novos casos e taxa de retorno.	Obter dados refinados.
9	Ação 3 - Contato telefônico realizado pelo coordenador do curso.	Obter retorno dos estudantes nas atividades acadêmicas.
10	Após 7 dias, solicitar nova execução do Modelo 1.	Gerar novas informações sobre os estudantes propensos a evadir.
11	Análise das informações.	Avaliação das informações geradas a partir do Modelo 1.
12	Cruzar as informações para obter os casos persistentes, novos casos e taxa de retorno.	Obter dados refinados.
13	Ação 4 - Contato telefônico realizado pelo coordenador do curso, a fim de verificar o motivo da evasão eminente e tentar reversão.	Obter retorno dos estudantes nas atividades acadêmicas e motivos de levam os estudantes a evadir, analisar e solicitar correção.
14	Durante o semestre melhorar no que couber, os pontos detectados.	Diminuir focos que levam a evasão.

Fonte: do Autor

5. CONCLUSÃO

A pesquisa norteou-se na obtenção de um modelo de predição de evasão e de um modelo para tomada de decisões, baseados nas informações geradas após a aplicação do KDD para mitigar o problema da evasão em cursos EAD.

Uma das finalidades desta pesquisa foi a construção de um processo para predição de evasão dos estudantes EAD, ou seja, descobrir se é possível predizer estudantes propensos a evadir baseados nos dados do SGA e AVA, verificando a precisão da predição com os dados "variantes no tempo" quanto com dados "não variantes no tempo", descobrindo quais são os atributos e classificadores mais significativos para a predição.

Percebeu-se no estudo que é possível considerar o uso KDD nos dados provenientes do tanto do SGA quando do AVA para predizer os estudantes propensos a evadir. Por um lado, os dados fornecidos pelo SGA (não variantes no tempo), possibilitam que o modelo seja aplicado assim que os estudantes finalizem suas matrículas, contudo, a maior precisão obtida foi de 0,3809%, predizendo 8 dos 21 casos possíveis. Por outro lado, os dados fornecidos pelo AVA (variantes no tempo), possibilitam que o modelo seja aplicado decorrido 30 dias do início das aulas, chegando a predizer 17 dos 21 casos possíveis coma precisão de 0,8095%. Gerando informações suficientes para que os gestores de curso possam tomar decisões a fim de reverter à evasão.

Os experimentos possibilitaram gerar informações para tomada de decisão de modo que não era possível fazer antes da aplicação do KDD na busca da predição da evasão. Dessa forma, acredita-se que a contribuição dessa pesquisa como sendo a demonstração da utilidade da aplicação do KDD no estudo da evasão na EAD, servindo de plataforma para o desenvolvimento de outros estudos.

Com a predição de evasão, as informações geradas permitem que gestores de curso, apoiado ou não por outros setores, possam articular ações envolvendo os estudantes pertencentes ao grupo de risco de evasão, intervindo, na tentativa de mitigar seus índices.

Frente aos resultados obtidos, acredita-se que o processo pode ser estendido aos demais cursos da instituição e até mesmo de outras instituições onde se observa o fenômeno da evasão, ajustando-o, para as especificidades dos diferentes públicos.

Dessa forma, a pesquisa propõe um modelo que prediz os estudantes propensos a evadir, e um modelo para suporte à tomada de decisões e ações. O primeiro modelo fornece in-

sumos para que as decisões e ações mitigadoras do segundo modelo possam ser iniciadas.

O desenvolvimento desta pesquisa para a detecção preditiva dos estudantes propensos a evadir, com a aplicação de técnicas de KDD mostrou-se eficaz, principalmente se aplicada com os dados advindos do AVA, obtendo boa precisão e assertividade conforme os resultados apresentados e discutidos anteriormente. Contudo, este é somente um ponto de partida para que outras pesquisas possam ser desenvolvidas tais como as sugeridas a seguir:

- Implantar e validar os modelos: de predição e tomada de decisões em outros cursos e verificar os resultados alcançados;
- Realizar comparativo entre os estudantes alvo, advindos dos dados “não variantes no tempo” com os estudantes alvo, advindos dos dados “variantes no tempo” a fim de descobrir possíveis padronizações entre eles.
- Realizar outros experimentos com dados “variantes no tempo” com antecedência menor que 30 dias de forma a verificar o mínimo de dias necessários para detectar os estudantes propensos a evadir;
- Aplicar e testar, validando o modelo proposto em outras instituições e/ou cursos de diferentes especificidades que apresentem índices de evasão consideráveis;
- Propor uma interface amigável para o modelo de predição de evasão para que os gestores possam acompanhar de forma sistemática os estudantes propensos a evadir e, assim, gerenciar o fenômeno da evasão.

REFERÊNCIAS

ALMEIDA, L. J. de. **A evasão escolar no Programa Senac de Gratuidade (PSG)**: um estudo de caso no Distrito Federal. 2012. 132 f. Dissertação (Mestrado em Educação) - Universidade Católica de Brasília, Brasília, 2012.

ALVES, J. R. M. Educação à Distância e as Novas Tecnologias de Informação e Aprendizagem. **Novas Tecnologias na Educação**. 2015. Disponível em http://www.engenheiro2001.org.br/programas/980201a1.htm. Acesso em: 12 mar 2016.

AMORIM, M. J. V. et al. Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. **Anais do Simpósio Brasileiro de Informática na Educação**. 2008. p. 666-674.

ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA – ABED. Censo EAD.br**. Relatório analítico da aprendizagem a distância no Brasil**. São Paulo: ABDR Education do Brasil, 2009, 2010, 2011, 2012, 2013 e 2014

BAGGI, S. et al. **Evasão e avaliação institucional no ensino superior**: uma discussão bibliográfica. 2011.

BAKER, R. S. J. D.; YACEF, K. The state of educational data mining in 2009: A review and future visions. **JEDM-Journal of Educational Data Mining**, v. 1, n. 1, p. 3-17, 2009.

BAKER, R. S. J. D. et al. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, vol. 19, no. 2, p. 2-13, 2011.

BARBOSA, R. L. L. **Formação de educadores: desafios e perspectivas**. UNESP, 2003.

BARBOSA, W. D. M. et al. Uma Proposta para Identificação de Causas da Evasão na Educação a Distância através de Mineração de Dados. **Anais da ERBASE-Escola de Computação Bahia-Alagoas-Sergipe**, 2014.

BRASIL. Decreto n. 5.622, de 19 de setembro de 2005. Dispõe sobre regulamenta o art. 80 da Lei nº 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional. **Diário Oficial da República Federativa do Brasil, Brasília**, DF: 2006. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2005/Decreto/D5622.htm>. Acesso em:15 mar 2015.

BECKER, F. Vygotski versus Piaget-ou sociointeracionismo e educação.

BRASIL, Censo Escolar da Educação Básica 2012-Resumo Técnico. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**, 2013.

CAMILO, C. O; SILVA, J. C. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC)**,** p. 1-29, 2009.

CARDOSO, O. N. P; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de administração pública**, v. 42,

n. 3, p. 495-528, 2008.

COSTA, S. S. da; CAZELLA, S.; RIGO, S. J. Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS. **RENOTE**, v. 12, n. 2, 2012.

CRISP-DM. **CRISP-DM 1.0**: Step-by-step data mining guide. Disponível em <http://www.crisp-dm.org>. Acesso em: 11 setembro. 2015

CUNHA, E. R.; MOROSINI, M. C. Evasão na educação superior: uma temática em discussão. **Revista Cocar**, v. 7, n. 14, p. 82-89, 2014.

DAUDT, S. I. D; BEHAR, P. A. A gestão de cursos de graduação a distância e o fenômeno da evasão. **Educação**, v. 36, n. 3, p. 412-421, 2013.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. MIT press, 2001.

IAS, P. Da e-moderação à mediação colaborativa nas comunidades de aprendizagem. **Educação, Formação & Tecnologias**, p. 4-10, 2008.

FAVERO, R. V. M; FRANCO, S. R. K. Um estudo sobre a permanência e a evasão na Educação a Distância. **RENOTE**, v. 4, n. 2, 2006.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

FERNANDES, J. et al. Identificação de Fatores que Influenciam na Evasão em um Curso Superior de Ensino a Distância. **Perspectivas OnLine 2007-2010**, v. 4, n. 16, 2014.

GABARDO, P; QUEVEDO, S. R. P de; ULBRICHT, V. R. Estudo comparativo das plataformas de ensino-aprendizagem. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, n. 2. sem., p. 65-84, 2010.

GAIOSO, N. P. L. **Evasão discente na educação superior**: a perspectiva dos dirigentes e dos estudantes. 2005. Dissertação (Mestrado em Educação) – Universidade Católica de Brasília, Brasília, 2005.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining**: Um Guia Prático-Conceitos, Técnicas, Ferramentas, Orientações e Aplicações. Rio de Janeiro: Campus, v. 1, 2005.

GONZALEZ, M. **Fundamentos da tutoria em educação a distância**. Avercamp, 2005.

GOTTARDO, E. et al. Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais. In: **Anais do Simpósio Brasileiro de Informática na Educação.** 2012.

GOTTARDO, E. et al**.** Avaliação de Desempenho de Estudantes em Cursos de Educação a Distância Utilizando Mineração de Dados. In: **Anais do Workshop de Desafios da Computação Aplicada à Educação**. 2012. p. 30-39.

INSTITUTO MONITOR (Org.). **Anuário Brasileiro Estatístico de Educação Aberta e a Distância: ABRAED-2005**. ABED, 2005.

JORGE, B. G. et al. Evasão na Educação a distância: um estudo sobre a evasão em uma instituição de ensino superior. In: **Ciaed Congresso Internacional Abed de Educação a Distância**. 2010.

KAMPFF, A. J. C. Identificação de Perfis de Evasão e Mau Desempenho para Geração de Alertas num Contexto de Educação a Distância. **Revista Latino americana de Tecnología Educativa-RELATEC**, v. 13, n. 2, p. 61-76, 2014.

KIRA, L. F. **A evasão no ensino superior: o caso do curso de pedagogia da Universidade Estadual de Maringá (1992-1996).** 1998. 106 f. 1998. Tese de Doutorado. Dissertação (Mestrado em Educação)–Programa de Pós-graduação em Educação da Universidade Metodista de Piracicaba, Piracicaba.

KDNUGGETS. **CRISP-DM, still the top methodology for analytics, data mining, or data science projects (Oct 28, 2014).** Disponível em http://www.kdnuggets.com. Acesso em: 15 de out. 2015.

LACHI, R. L. et al. Uso de agentes de interface para auxiliar a avaliação formativa no ambiente TelEduc. **XIII Simpósio Brasileiro de Informática na Educação**. São Leopoldo-RS, Novembro, p. 2-9, 2002.

LAGUARDIA, J.; PORTELA, M. Evasão na educação a distância Dropout in distance education. **ETD-Educação Temática Digital**, v. 11, n. 1, p. 349-379, 2009.

LANDIM, C. M. M. P. F. **Educação a distância: algumas considerações**. Rio de Janeiro, 1997.

LICENSE, GNU General Public. License used by the Free Software Foundation for the GNU Project**. Free software foundation**. Disponível em: https://www.gnu.org/licenses/gpl-3.0.html. <acesso em 19/06/2016>. Acesso em: 19/02/2016.

LIMA, J. R. C.; MOURA, K. H. S. Mineração De Dados Em Redes Sociais Usando o Nodexl**. Revista Cogitatem**, v. 1, n. 1, p. 1-19, 2015.

LOBO, R. L et al. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641-659, 2007.

MAIA, M. de C. et al. Análise dos índices de evasão nos cursos superiores a distância do Brasil. In: **Anais do XI Congresso Internacional de Educação à Distância**. Salvador; Bahia. 2004.

MARTINS, C. B. N. **Evasão de alunos nos cursos de graduação em uma instituição de ensino superior**. Montes Claros, 2007.

MORAN, J. M. **Contribuições para uma pedagogia da educação on-line. Educação on-line: teorias, práticas, legislação, formação corporativa**. São Paulo: Loyola, 2003. p. 39-50.

MELO F. et al. Percepção social em EAD: Identificando necessidades para o LMS Amadeus. **Revista Brasileira de Informática na Educação**, v. 19, n. 3, 2011

MINAYO, M. C. de S. O desafio do conhecimento científico: pesquisa qualitativa em saúde. **São Paulo/Rio de Janeiro: Hucitec-Abrasco**, p. 01-10, 1993.

MORESI, E. et al. Metodologia da pesquisa. **Brasília: Universidade Católica de Brasília**, v. 108, 2003.

NUNES, I. B. A história da EaD no mundo. In: LITTO, Fredric; FORMIGA, Marcos. *Educação a distância.* **O estado da arte. São Paulo: Pearson Education do Brasil**, 2009, p. 2-8.

OLIVEIRA, A. E. F. et al. **Avaliação da Produtividade no Processamento dos Dados em um Curso de Pós-graduação Lato Sensu da UNASUS/UFMA na Modalidade a Distância utilizando o Sistema de Monitoramento (SIM).** Disponível em: http://www.telessaude.uerj.br/resource/goldbook /pdf/27.pdf: Acesso em 19/02/16.

PASTA, A. **Aplicação da técnica de data mining na base de dados do ambiente de gestão educacional**: um estudo de caso de uma instituição de ensino superior de Blumenau – SC. 2015. 153 f. Dissertação (Mestrado em Computação Aplicada) – Universidade do Vale do Itajaí, 2011.

PINHEIRO, M. F. et al. Identificação de Grupos de Estudantes em Ambiente Virtual de Aprendizagem: Uma Estratégia de Análise de Log Baseada em Clusterização. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2014.

PRETI, O. O.; OLIVEIRA, G. Estado da Arte sobre "Tutoria": modelos e teorias em construção. **PRETI, O.; OLIVEIRA, GMS O sistema de Orientação Acadêmica no curso de Pedagogia a distância da Universidade Federal de Mato Grosso. Relatório de Pesquisa. Programa CAERENAD-Téléuniversité du Québec, Canadá**, 2003.

PRETI, O.; OLIVEIRA, GMS O sistema de Orientação Acadêmica no curso de Pedagogia a distância da Universidade Federal de Mato Grosso**. Relatório de Pesquisa. Programa CAERENAD-Téléuniversité du Québec**, Canadá, 2003

RIGO, S. J. et al. Educação em Engenharia e Mineração de Dados Educacionais: oportunidades para o tratamento da evasão. **EaD & Tecnologias Digitais na Educação**, v. 2, n. 3, p. 30-40, 2014

RIBEIRO, E. N. et al. A importância dos ambientes virtuais de aprendizagem na busca de novos domínios da EAD. In: **Congresso da Associação Brasileira de Educação a Distância, Goiás.** Disponível em: http://www. abed. org. br/congresso2007/tc/4162007104526am. pdf. 2007

RIBEIRO, M. A. O projeto profissional familiar como determinante da evasão universitária: um estudo preliminar. **Revista Brasileira de Orientação Profissional**, v. 6, n. 2, p. 55-70, 2005.

ROMERO, C. et al. Knowledge discovery with genetic programming for providing feedback

to courseware authors. **User Modeling and User-Adapted Interaction**, v. 14, n. 5, p. 425-464, 2004.

ROMERO, C. et al. Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: **Applications and Reviews, IEEE Transactions on**, v. 40, n. 6, p. 601-618, 2010.

SABBATINI, R. M. E. **Ambiente de Ensino e Aprendizagem via internet: a plataforma moodle**. Campinas: Instituto Edumed, 2007.

SANTANA, M. A. et al. A predictive model for identifying students with dropout profiles in online courses. **CEUR**, v.1446: Workshops Proceedings of EDM 2015 - 8th International Conference on Educational Data Mining, 2015. Disponível em: http://ceur-ws.org/Vol-1446/smlir_submission3.pdf,

SANTOS, A. P**. A predição da evasão de estudantes de graduação como recurso de apoio fornecido por um assistente inteligente**. 2014. 64f. Dissertação (Mestrado em Gestão do Conhecimento e Tecnologia da Informação) – Universidade Católica de Brasília, Brasília, 2014.

SANTOS, E. M. et al. **Evasão na Educação a Distância: identificando causas e propondo estratégias de prevenção**. Paidéia, 2008.

SANTOS, R. N. et al. **Uma Abordagem Genérica de Identificação Precoce de Estudantes com Risco de Evasão em um AVA utilizando Técnicas de Mineração de Dados**. 2014.

SILVA, D. C. F. **Variáveis preditoras de evasão em cursos a distância**. 2016. 94 f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Católica de Brasília, Brasília, 2016.

SILVA FILHO, R. L. L. et al. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641-659, 2007.

SILVA, M. P. S. Mineração de dados: Conceitos, aplicações e experimentos com weka. **Livro da Escola Regional de Informática Rio de Janeiro-Espírito Santo. Porto Alegre: Sociedade Brasileira de Computação**, v. 1, p. 1-20, 2004

SOUZA, E. P. de. Avaliação Formativa em Educação a Distância via Web. **13º Congresso Internacional de Educação à Distância. Curitiba –PR**. 2007.

VELOSO, L. A. **A predição da Evasão Escolar dos Cursos Técnicos de Nível Médio**: um estudo de caso no SENAI. 2015. 94 f. Dissertação (Mestrado em Gestão do Conhecimento e Tecnologia da Informação) – Universidade Católica de Brasília, Brasília, 2015.

WALTER, A. M. **Variáveis preditoras de evasão em cursos a distância**. 2006. Tese de Doutorado. Dissertação de mestrado, Universidade de Brasília, Brasília.

ANEXO A - MÉTRICAS DE QUALIDADE DE CLASSIFICADORES PARA PREDIÇÃO DE EVASÃO EM EDUCAÇÃO À DISTÂNCIA (SILVA, 2016)

Após a obtenção de um classificador, ele deve ter a sua qualidade verificada por algum método confiável. O método usualmente utilizado é a medição da "precisão da classificação" (*Classification Accuracy*), que é o percentual de predições corretas em relação ao total de predições. Há diversas formas de analisar a precisão da classificação, como o *Split Test* e *Cross Validation*.

No *Split Test*, os dados são separados aleatoriamente em duas partes. Uma parte é analisada (a parte de treino) para estabelecer as regras da classificação e a outra serve para testes, onde os resultados, para casos independentes, são previstos pelas regras estabelecidas. Se os testes forem feitos em dados que o classificador já viu antes, ele pode simplesmente informar o resultado que ele já conhece, invalidando o seu poder preditivo. Os dados de teste são novos para o classificador e tem uma classificação conhecida, servindo para calcular a porcentagem de predições corretas. Uma das variações do *Split Test* é o *Multiple Split Test*, em que a classificação é feita algumas vezes, cada uma com as bases de treinamento e teste definidas aleatoriamente, e sua acurácia é a média das acurácias encontradas em cada classificação. Isso visa atenuar o problema da aleatoriedade da classificação. No entanto, mesmo fazendo a classificação dessa forma, não há garantia de que todos os dados serão utilizados para treinamento e testes. A Figura I.1 representa a diferença entre esses *Split Test* é o *Multiple Split Test***.**

O método *Cross Validation* garante que cada instância do dado será utilizada pelo menos uma vez para treinamento e testes. Os dados são divididos em *k* grupos e o classificador é acionado *k* vezes (Figura I.2). Em cada uma dessas vezes, um dos grupos serve para testes e os outros são utilizados para treinamento, garantindo que todos os grupos serão utilizados uma vez para testes e *k*-1 vezes para treinamento.

Figura I.1 – O modelo de análise de precisão da classificação *Split Test* e sua variação *Multiple Split Test*

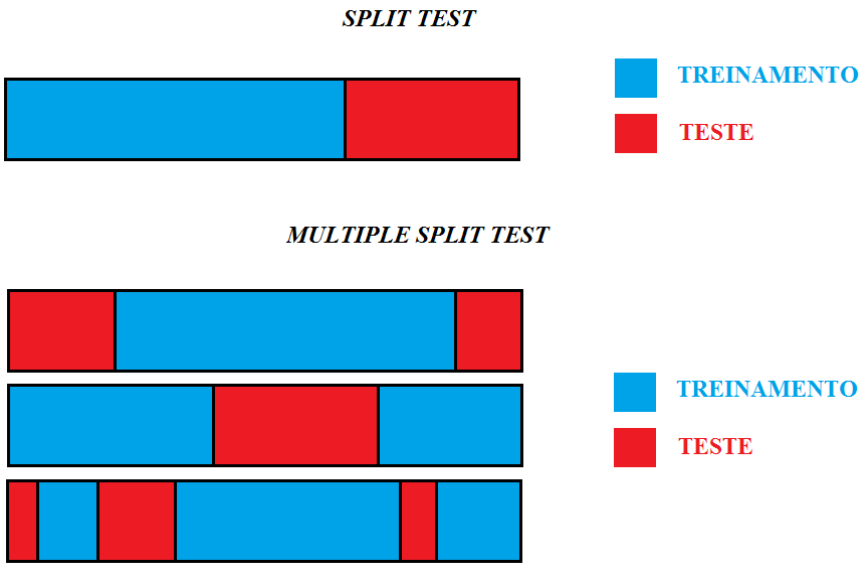
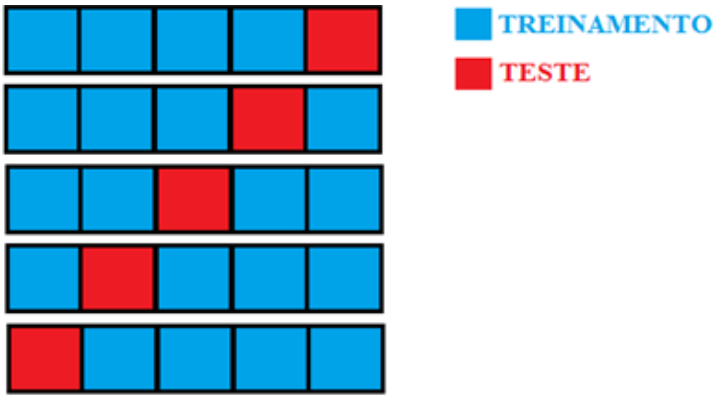


Figura I.2 – O modelo de análise de precisão da classificação *Cross Validation*



Nem sempre a acurácia e a qualidade de um modelo são equivalentes. Parece óbvio que um modelo com a acurácia elevada terá um desempenho melhor, e mais significância para o negócio, mas não é difícil ver situações onde modelos com uma acurácia alta são triviais e inúteis. Uma simples análise de acurácia não leva em conta itens importantes para o negócio, expressos nos falsos positivos e falsos negativos, e apenas considera a acurácia de modo grosseira. Ocorre que, comumente, modelos com uma menor acurácia são melhores para o negócio se o número de falsos negativos ou falsos positivos também for menor, resultando em um custo menor. Também há o problema de classes desbalanceadas, onde um modelo trivial, que classifica todos os casos de acordo com a classe mais representada, teria uma acurácia alta, mas não produz informação útil para o negócio. A acurácia não é irrelevante ao negócio, mas,

ela não deve ser utilizada como critério único e absoluto de qualidade do modelo, cabendo a utilização de métricas alternativas de qualidade., principalmente em casos onde as classes estão desbalanceadas, ou o custo para os diferentes tipos de erro do classificador são diferentes[28].

Há algumas maneiras para evitar esses erros, como utilizar outras medidas de desempenho, que podem ser utilizadas como complemento da acurácia da classificação, como *Kappa, F-Measure* e *ROC Curves*.

Para ilustrar melhor a representação dos resultados de uma predição, é utilizada uma matriz de confusão, como mostrado na Figura I.3. No caso de um problema de classificação binário, como o objeto desse estudo, ela será uma matriz com duas linhas e duas colunas. As colunas são os resultados reais, e as linhas os resultados previstos. Desse modo, é possível visualizar a quantidade de previsões corretas e erradas claramente, e se os erros foram de falsos negativos ou falsos positivos.

Figura I.3 - Matriz de confusão

Resultado / **Predição**	**Positivo**	**Negativo**
Positivo	Positivo verdadeiro	Falso positivo
Negativo	Falso negativo	Negativo verdadeiro

Um modelo de predição perfeito teria essa matriz preenchida apenas na diagonal principal, que começa na célula superior esquerda. As predições incorretas são definidas nas outras células, e separadas pelo tipo de erro delas, seja predizendo uma ocorrência onde não haverá nenhuma, ou não identificando uma ocorrência real.

Utilizando essa matriz é mais fácil ilustrar o paradoxo da acurácia. No exemplo mostrado anteriormente, um classificador trivial, para o qual todos os 100 estudantes são classificados como não evadidos, sendo que, na realidade só 80 deles não evadiram, cuja matriz de confusão seria representado conforme mostrado na Figura I.4, ao responder a pergunta “quais estudantes vão evadir o curso?”, teria uma acurácia de 80%.

[28] CHAWLA, N et al (2002). SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357. Disponível em: http://jair.org/media/953/live-953-2037-jair.pdf.

Figura I.4 - Matriz de confusão para classificador trivial

Resultado / Predição	Evasão	Conclusão
Evasão	0	0
Conclusão	20	80

Nesse caso, perdeu-se a informação mais importante: a quantidade de estudantes que irão evadir o curso. Sem essa informação, não é possível a adoção de medidas preventivas para a evasão. Esse tipo de problema é relativamente comum em casos onde há um desbalanceamento de classes, tendendo a classificação à classe melhor representada. O resultado para a instituição são os custos advindos por não se ter tomado medidas preventivas em relação a possíveis evasões.

Supondo que um segundo classificador chegasse à matriz de confusão apresentada na Figura I.5, ele é um pouco pior quanto à sua acurácia, de 76% (14 + 62 resultados corretos, de um total de 100), mas, diferentemente do primeiro classificador, mesmo sendo uma menor acurácia, apresenta um número de falsos negativos significativamente menor (6), ou seja, menos estudantes que vão evadir foram classificados como estudantes que vão concluir. Isso significa que para pelo menos 14 estudantes, medidas preventivas poderiam ser tomadas corretamente, podendo diminuir o impacto dessas evasões.

Figura I.5 - Matriz de confusão para classificador hipotético

Resultado / Predição	Evasão	Conclusão
Evasão	14	18
Conclusão	6	62

Nesse sentido, medidas de desempenho podem ser agregadas à acurácia no sentido de avaliar os classificadores: a **precisão e regressão**.

A **precisão** responde a pergunta: “de todos os casos que o classificador previu um resultado positivo, qual a frequência que ele estava certo?”. Ela é calculada dividindo as previsões positivas corretas pelo total de previsões positivas classificadas (positivos verdadeiros + falsos positivos). No exemplo, seria o total de evasões corretas que o classificador identificou em relação ao total de evasões que ele previu.

No primeiro caso (classificador trivial), não haveria precisão, já que não foi prevista nenhuma evasão. No segundo caso (classificador hipotético), a precisão seria de 43% (14/ (14+18) = 0,43).

A regressão responde a pergunta: “de todos os casos que (tal fenômeno) poderia ocorrer, quantas o classificador conseguiu identificar?”. Ela é calculada dividindo o número de previsões positivas corretas pelo número total de casos positivos (positivos verdadeiros + falsos negativos). No exemplo acima, seria o total de evasões corretamente previstas divididas pelo total de evasões que ocorreram.

No primeiro classificador, também não haveria regressão, já que ele não foi prevista nenhuma evasão. No segundo, seria de 70% (14/ (14+6) = 0,70).

Essas duas medidas dão uma visão melhor sobre a qualidade real do classificador, mas elas ainda não são o suficiente. Cada uma delas tem uma visão diferente: a primeira identifica a precisão da classificação do algoritmo, mas olhando apenas para as previsões dele, sem considerar o número de falsos negativos. Muito comumente os falsos negativos tem um custo maior do que os falsos positivos. A regressão, por outro lado, busca a completude das predições, ou seja, ela olha para as previsões feitas e o total de ocorrências, sem considerar a quantidade de falsos positivos para o cálculo. Um terceiro classificador que classificasse que todos os estudantes iriam evadir o curso teria 100% de regressão, mas também não seria uma solução boa.

Há o modelo de medição, *F-Measure*, que leva em consideração a precisão e a regressão para avaliar a qualidade do modelo, fazendo um balanceamento entre as duas, oferecendo uma visão melhor sobre a verdadeira qualidade do classificador.

Outro método que chama atenção é o *Kappa*. Uma predição pode estar correta apenas por sorte. *Kappa* é um método estatístico para determinar se as observações que foram feitas estão corretas apenas por chance, ou se elas são de fato melhores do que se poderia esperar de uma classificação baseada em chance.

Utilizando esse método, é possível encontrar um valor numérico que diz a relação entre o quanto o resultado encontrado se compara com um resultado de chance. Esse valor é padronizado para variar entre -1 e 1. Um valor de 1 indica que nenhuma observação estava correta por chance, um sistema de predição perfeito. Um valor de 0 indica que o sistema está prevendo do mesmo modo que um sistema de predição aleatório criaria as predições, nada melhor do que uma classificação por chance. Um valor negativo é interessante, e indica que o

sistema se sairia melhor fazendo previsões aleatórias, o que pode indicar um problema na formulação da lógica, na pergunta, ou uma anomalia.

Para encontrar o coeficiente *kappa*, é utilizada a fórmula:

$$kappa = \frac{Acurácia\ Total - Acurácia\ Aleatória}{1 - Acurácia\ Aleatória}$$

Onde a "acurácia total" já foi apresentada: O número de predições corretas dividido pelo número total de observações. A "acurácia aleatória" é:

$$Acurácia\ Aleatória = \frac{Predição\ negativa \times Resultado\ negativo + Predição\ positiva \times Resultado\ positivo}{Total\ de\ predições^2}$$

O *kappa* dá a possibilidade de ver a relação dos resultados obtidos, o que permite que ele seja utilizado em conjunto com outros métodos, e prover informações importantes, como, por exemplo, uma análise de acurácia alta, mesmo utilizando o *F-measure*, pode ser observado junto ao coeficiente *kappa*, que, caso seja baixo, vai indicar que o resultado foi produzido pela chance. Por exemplo, uma análise de acurácia de 82%, *F-measure* de 79%, e o *kappa* de 0,19, indica que a predição, mesmo tendo uma boa acurácia, como indicado no *F-measure*, está chegando a esse resultado em boa parte por sorte.

Não há uma tabela fixa para os valores do *kappa*, mas normalmente, é aceito um valor acima de 0,8 como significantemente melhor que um resultado por chance. Valores entre 0,4 e 0,8 são moderadamente melhores, e abaixo de 0,4 tem menos significância estatística.

Outra forma de medir a qualidade real de um modelo, preferencialmente de classificação binaria, é a curva *ROC*. Essa técnica, assim como a *F-Measure*, utiliza dois conceitos, a sensibilidade e a especificidade.

A sensibilidade equivale à regressão, ou seja, a capacidade de um classificador identificar corretamente um caso "positivo" dentro de todos que podem ocorrer. Ela é medida dividindo a quantidade de casos positivos encontrados pelo total de casos positivos que ocorreram (positivos reais mais os falsos negativos).

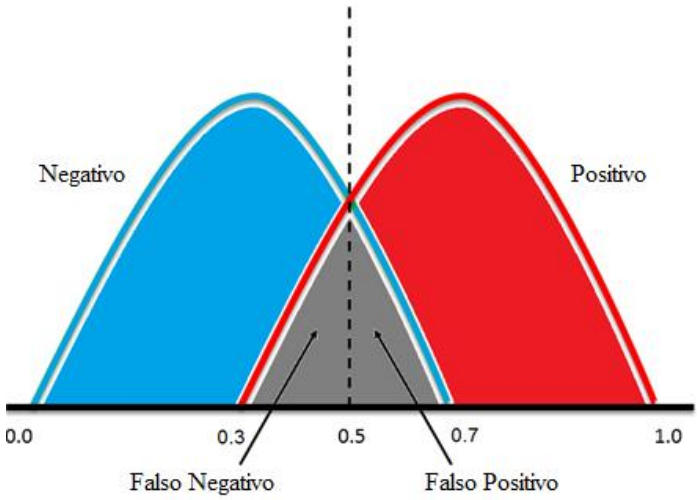
A especificidade é o oposto, a habilidade do classificador de corretamente rejeitar os casos negativos. Ela é medida dividindo os casos negativos corretamente classificados pelo algoritmo pelo total de casos negativos (negativos reais mais falsos positivos).

Os classificadores normalmente não dividem suas predições em grupos definidos. Eles, na verdade, definem uma probabilidade para cada observação de estar junto a determi-

nada classificação, e, a partir dessas probabilidades definem a classificação. O normal em uma classificação binaria é definir qualquer observação com uma probabilidade acima de 50% como positiva, e as outras como negativas. A curva *ROC* só pode ser usada em algoritmos que gerem essa pontuação de probabilidade, ao invés de apenas decisões binarias.

A Figura I.6 mostra um modelo hipotético que poderia ser criado em um classificador.

Figura I.6 - Gráfico de demonstração de predições



O eixo vertical é a quantidade de observações, o horizontal a probabilidade que o classificador definiu para cada caso. A área azul são os casos negativos, a vermelha os positivos, a cinza é a intersecção, tendo tantos casos negativos quanto positivos.

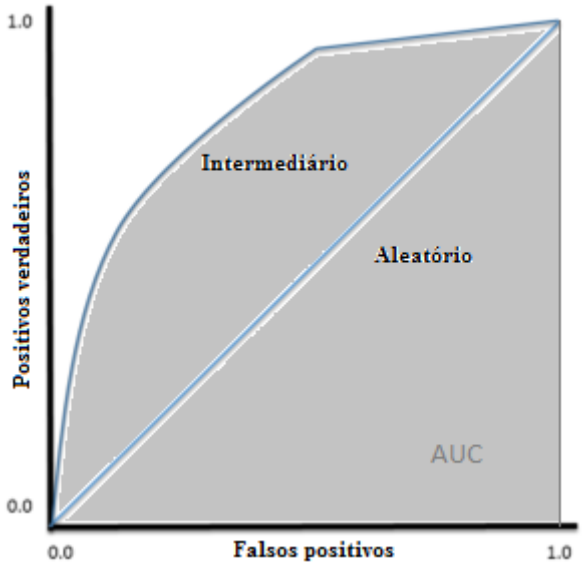
Visualizando esse modelo fica claro que, não importa qual seja o limite definido pelo classificador, ele não vai ser perfeito, sempre gerando falsos positivos ou falsos negativos. Por exemplo, se o limite fosse definido em 70% (0,7), ele teria uma especificidade de 100%, porque todos os casos negativos seriam corretamente classificados, mas iria gerar um número alto de falsos negativos, e teria uma sensibilidade baixa. Um modelo com o limite em 30% (0,3) teria uma sensibilidade de 100%, mas apresentaria um número alto de falsos positivos.

A curva *ROC*, diferente dos outros modelos, olha para todos os limites que podem ser definidos por um modelo. Para montá-la, todos os valores da sensibilidade e o complemento da especificidade (1 – especificidade, a porcentagem de falsos positivos) são calculados, para todos os limites possíveis. (0,01; 0,02,...). Esses pontos são colocados em um gráfico, onde a sensibilidade (positivos verdadeiros) está no eixo y e os falsos positivos no eixo x, formando a curva *ROC*.

Um exemplo de uma curva *ROC* é apresentado na Figura I.7. No gráfico, é utilizada

uma curva que o corta na diagonal, que serve como medida de uma classificação aleatória, que estaria certa em 50% dos casos (teoricamente). Quanto mais perto da esquerda a curva estiver, melhor. Quanto mais próxima da linha diagonal, mais o modelo se parece com um modelo aleatório, sem inteligência.

Figura I.7 - Curva ROC



Essa curva não nos dá uma análise direta da acurácia, mas tem algumas vantagens sobre os outros métodos de análises, sendo o mais evidente a possibilidade de utilizá-la para selecionar o limite que mais atende à nossa necessidade, que pode ser minimizando os falsos positivos, falsos negativos, ou alguma combinação.

Para realizar a análise da qualidade utilizando a curva *ROC*, a área embaixo da curva (*Area Under ROC Curve* - AUC) é calculada. Essa área fornece medida para comparar desempenho de classificadores. Quanto maior a área AUC melhor o desempenho global do classificador.