UNIFACS UNIVERSIDADE SALVADOR
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO
MESTRADO EM SISTEMAS E COMPUTAÇÃO

PÉRICLES NOGUEIRA MAGALHÃES JÚNIOR

UM MODELO DE DADOS PARA APOIAR A MINERAÇÃO DE DADOS EDUCACIONAIS NA INVESTIGAÇÃO DE EVASÃO DE ESTUDANTES

Salvador
2013

PÉRICLES NOGUEIRA MAGALHÃES JÚNIOR

UM MODELO DE DADOS PARA APOIAR A MINERAÇÃO DE DADOS EDUCACIONAIS NA INVESTIGAÇÃO DE EVASÃO DE ESTUDANTES

Dissertação apresentada ao Programa de Pós-graduação – em Sistemas e Computação da Universidade Salvador, como requisito parcial para obtenção do grau de Mestre.

Orientador: Prof. Dr. Rodrigo Oliveira Spínola.

Salvador
2013

FICHA CATALOGRÁFICA
Elaborada pelo Sistema de Bibliotecas da UNIFACS Universidade Salvador, Laureate Internacional Universities

Magalhães Júnior, Péricles Nogueira

Um modelo de dados para apoiar a mineração de dados educacionais na investigação de evasão de estudantes./ Péricles Nogueira Magalhães Júnior. – Salvador, 2013.

134 p.: il.

Dissertação apresentada ao Curso de Mestrado em Sistemas e Computação, UNIFACS Universidade Salvador, como requisito parcial para obtenção do grau de Mestre.

Orientador: Prof. Dr. Rodrigo Oliveira Spínola.

1. Mineração de dados educacionais. 2. Evasão escolar. I. Spínola, Rodrigo Oliveira , orient. II. Título.

CDD.004.22

TERMO DE APROVAÇÃO

PÉRICLES NOGUEIRA MAGALHÃES JÚNIOR

UM MODELO DE DADOS PARA APOIAR A MINERAÇÃO DE DADOS EDUCACIONAIS NA INVESTIGAÇÃO DE EVASÃO DE ESTUDANTES

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Sistemas e Computação, UNIFACS Universidade Salvador, pela seguinte banca examinadora:

Rodrigo Oliveira Spínola – Orientador _____
Doutor em Engenharia de Sistemas e Computação pela COPPE/UFRJ
UNIFACS Universidade Salvador

Expedito Carlos Lopes _____
Doutor em Ciência da Computação pela Universidade Federal de Campina Grande (UFCG)
UNIFACS Universidade Salvador

Marcos Kalinowski _____
Doutor em Engenharia de Sistemas e Computação pela COPPE/UFRJ
Universidade Federal de Juiz de Fora (UFJF)

Salvador, de setembro de 2013.

Dedico esse trabalho a meu pai (*in memorian*), eterna referência em minhas ações.

Dedico, também, a minha família, Veruza, Alice e Cecília, pela paciência e compreensão pelas ausências.

AGRADECIMENTOS

A Edinaldina, minha mãe, e a Poliana e Patrícia, minhas irmãs, companheiras ativas em minha jornada pela vida.

Aos amigos, Profa. Dra. Eulália Azevedo, Profa. Dra. Tereza Cristina Fagundes, Profa. Mônica Massa e Prof. Dr. Guna Alexander, pelas várias conversas que, em momentos importantes, me mantiveram firme no propósito.

Ao orientador Prof. Dr. Rodrigo Oliveira Spínola, por sua dedicação e empenho no acompanhamento do projeto.

Ao Prof. Dr. Sidney Viana, pelo direcionamento e orientações iniciais.

Aos colegas, Prof. Luís Michelon, Jailson Oliveira, Jaizo Araujo e Rafael Guimarães pela importante contribuição na pesquisa.

Aos professores e colegas do programa de programa de Mestrado em Engenharia de Sistemas e Computação da Universidade Salvador.

A todos que direta ou indiretamente contribuíram para a realização deste trabalho.

RESUMO

O advento da inclusão da Educação a Distância – EAD na Lei de Diretrizes e Bases da Educação – Lei 9.394, de 20 de dezembro de 1996, alçou a modalidade como uma das principais estratégias de inclusão educacional no Brasil, provocando um efervescente crescimento da oferta de cursos, com uma variada gama de metodologias e formatos educacionais. O uso intenso de soluções de *software* na operação e gestão desses cursos, tais como sistemas de gestão de aprendizagem, produz um grande volume de dados inerentes ao comportamento de seus estudantes, matéria prima pouco aproveitada nos processos decisórios dessas instituições. As técnicas e algoritmos de mineração de dados, por sua capacidade de processar grandes montantes de dados na identificação de padrões comuns, têm sido utilizadas por pesquisadores e gestores de cursos da modalidade como apoio nos seus processos decisórios. Esses trabalhos, porém, guardam pouca ou nenhuma relação entre si no que se refere a abordagens e terminologias. A contribuição dessa dissertação de mestrado baseia-se na construção e proposição de um modelo de dados que reúna indicadores aplicáveis em diversas situações educacionais e que possa ser utilizado como referência em futuras pesquisas, proporcionando uma maior homogeneidade de terminologias e, consequentemente, permitindo análises comparativas em diferentes trabalhos. Para a seleção das entidades e atributos que compõem o modelo proposto, foi realizado um levantamento bibliográfico acerca das pesquisas realizadas na área da mineração de dados educacionais e dos principais modelos conceituais de análise comportamental de estudantes. Dessa forma, foi proposto um modelo para aplicações de mineração de dados educacionais e, tomando como foco o fenômeno da evasão, um estudo de avaliação foi realizado, mostrando que o modelo é aplicável, permitindo a identificação de indícios de evasão de estudantes, além de reduzir os esforços necessários para a seleção de atributos e subsequente preparação dos dados para a mineração de dados. A possibilidade de utilização do modelo em futuras pesquisas permitirá a convergência de termos e conceitos, propiciando uma maior troca de experiências entre pesquisadores.

Palavras chaves: Mineração de Dados Educacionais. Evasão de Estudantes. Modelo de Dados.

ABSTRACT

The insertion of Distance Education in the Brazilian Law of Guidelines and Bases of Education (Law 9,394 of December 20, 1996), lifted the modality as one of the main strategies for inclusive education in the country, causing a growth of courses offer, with a diverse range of methodologies and educational formats. Extensive use of software solutions in the operation and management of these courses, such as learning management systems, produces a large volume of data related to the behavior of their students, unexplored raw material in decision making processes of these institutions. The data mining techniques and algorithms, for their ability to process large amounts of data to identify common patterns, have been used by researchers and managers of online courses supporting their decision processes. These papers, however, have little or no relation with each other regarding to their approaches and terminologies. The contribution of this dissertation is based on the construction and proposal of a data model that can join applicable indicators in various educational situations and can be used as a reference in future studies, providing greater uniformity of terminology and thus allowing comparative analyzes in different works. For the selection of entities and attributes that made up the proposed model, we conducted a literature review regarding the research conducted in the Educational Data Mining area and about the major conceptual models of student's behavior analysis, focusing on the attrition phenomenon. Thus, we propose a reference model for Educational Data Mining applications and, focusing the attrition phenomenon, an evaluation study was conducted, showing that the model is applicable, allowing the identification of evidence of evasion of students, as well as to reduce the effort required for attributes selection and subsequent preparation of the data for data mining. The use of a reference model for future researches will enable the convergence of terms and concepts, providing greater exchange of experiences between researchers.

Keywords: Educational Data Mining. Students Attrition. Data Model.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A definição e execução de políticas educacionais para a qualificação de profissionais é imprescindível para o desenvolvimento de um país. No Brasil, iniciativas públicas e privadas de educação corporativa e educação continuada têm sido fundamentais para a aceleração da lacuna existente na qualificação profissional da população, principalmente com o crescimento da oferta de cursos da modalidade a distância. A partir do desenvolvimento de uma legislação que regule o funcionamento desses cursos e com a isonomia do reconhecimento de certificações de cursos de presenciais e a distância, a atratividade desses cursos perante o mercado tem aumentado progressivamente gerando, consequentemente, uma maior preocupação com a sua melhoria de qualidade (ABED, 2011).

Nesse sentido, gestores e professores da modalidade necessitam cada vez mais de informações e conhecimentos que sustentem suas análises e apoiem suas decisões, conforme apontado no censo (ABED, 2011):

> Há uma valorização da busca pelo conhecimento. No ensino, há um esforço constante para a organização de informações relevantes do ponto de vista pedagógico e tecnológico, traçando-se caminhos múltiplos para que a aprendizagem seja facilitada. Conceitos do paradigma presencial, como “semestre”, “sala de aula” e os limites geográficos, deixam de fazer sentido. Essas mudanças são grandes desafios para pesquisadores tanto do meio acadêmico como do mercado. As informações quantitativas, análises qualitativas, projeções e avaliações detalhadas podem indicar onde estávamos, onde estamos e para onde estamos indo e auxiliar na tomada de decisões. Os dados e informações oferecem maior segurança na tomada de decisões, seja para instituições, seja para empresas ou mesmo para o governo e órgãos reguladores no estabelecimento de diretrizes para a política educacional.

No cenário apresentado, identifica-se uma forte convergência entre os dados gerados a partir da operação de cursos na modalidade a distância, pela sua natureza e pelo seu volume, e a utilização de recursos tecnológicos que suportem a sua transformação em informações úteis aos processos decisórios (ROMERO *et al*, 2006). Identificar o conhecimento oculto em grandes bases de dados é, justamente, o objetivo da mineração de dados, principal etapa da KDD, sigla em inglês para *Knowledge Discovery in Databases*, ou Descoberta de Conhecimento em Bases de Dados, área da tecnologia da informação voltada para a inteligência em processos gerenciais.

1.1 O PROBLEMA ANALISADO

A aplicação de técnicas e tarefas da mineração de dados na área educacional tem sido objeto de pesquisa de diversos trabalhos em organizações e instituições de ensino e pesquisa,

definindo, inclusive, um termo específico para a Mineração de Dados Educacionais (MDE) (BAKER *et al*, 2011). As iniciativas de MDE encontradas abordam, em sua maioria, situações específicas de suas instituições de ensino ou de um interesse em particular dos seus autores, gerando uma grande heterogeneidade no material produzido, o que dificulta sobremaneira a sua utilização em análises transversais ou na sua própria continuidade evolutiva.

Em uma das poucas iniciativas com caráter homogeneizador, a Associação Brasileira de Educação a Distância (ABED), realiza o seu censo anual buscando proporcionar informações compiladas e estratificadas sobre o mercado brasileiro. Na versão do ano de 2011, a ABED já aponta que:

> Dados e informações são indispensáveis para que análises científicas sejam possíveis e, no Brasil, nem sempre as pessoas e instituições dispõem-se a compilá-los e disponibilizá-los a partir de esforços sistemáticos. (ABED, 2011).

O trabalho atual identifica uma lacuna gerada pela falta de padronização nas diversas produções acadêmicas sobre mineração de dados educacionais e propõe um modelo de dados que possa ser aplicado em futuros estudos, acadêmicos ou comerciais, independentemente do problema abordado ou da instituição de ensino analisada. A utilização de um único modelo de dados em aplicações de MDE, além de promover uma maior homogeneidade de conceitos, possibilita análises comparativas entre seus resultados.

O Censo ABED (2011) aponta, também, que o maior problema das instituições de ensino superior com ofertas na modalidade EAD encontra-se nas altas taxas de evasão de estudantes. Dessa forma, o trabalho atual toma como cenário para a construção e aplicação do modelo proposto, o estudo do fenômeno da evasão de estudantes na modalidade.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral dessa dissertação é propor um modelo de dados para utilização em trabalhos de mineração de dados educacionais, aplicável em distintas situações problema, independentemente da sua instituição de ensino, estrutura tecnológica de software e bancos de dados.

1.2.2 Objetivos Específicos

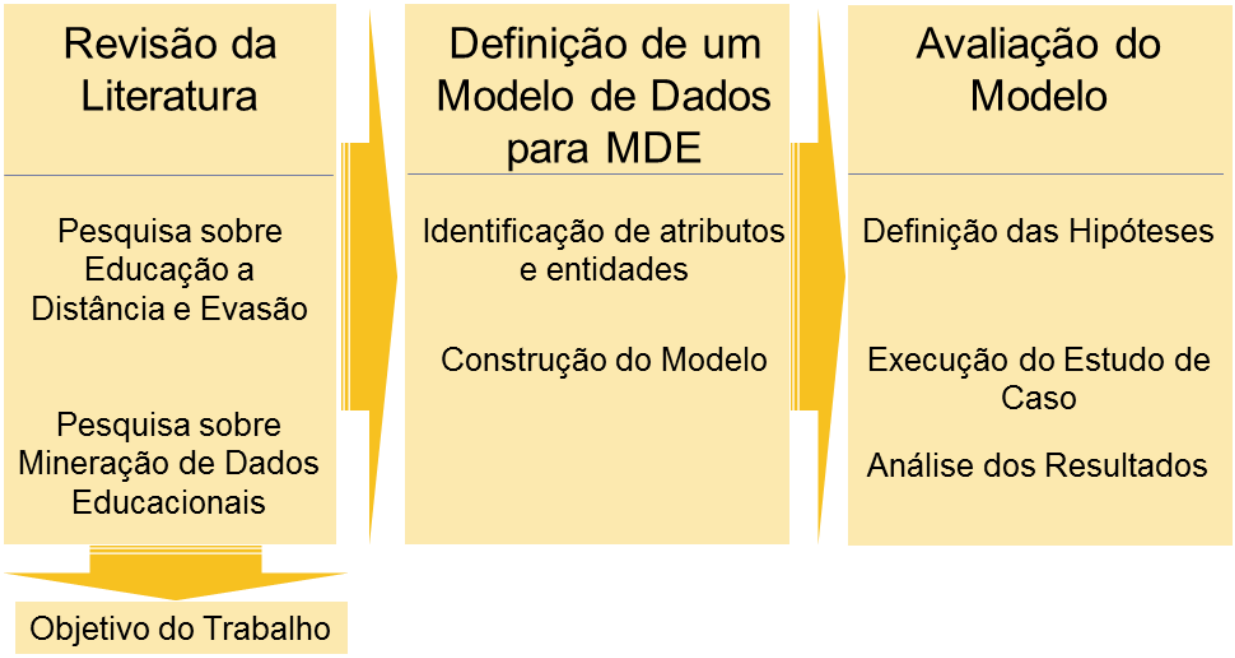
Para a concretização do objetivo geral, são traçados os seguintes objetivos específicos:

a) A identificação de um conjunto mínimo de entidades e atributos relacionados a situações problema comuns em cursos EAD, alinhados com os principais modelos conceituais de análise pedagógica e acadêmica da relação dos estudantes com seus cursos e instituições de ensino.
b) Definir um modelo de dados a partir do conjunto de atributos e entidades identificados.
c) Avaliar a aplicabilidade do modelo proposto com dados reais e com a utilização de algoritmos de mineração de dados na identificação de indícios de evasão de estudantes.
d) Avaliar os benefícios da utilização do modelo de dados proposto em aplicações de mineração de dados educacionais.

1.3 METODOLOGIA UTILIZADA

O processo de construção do trabalho atual partiu da revisão da literatura existente sobre aplicações na Educação a Distância que abordassem o fenômeno da evasão de estudantes. Como resultado dessa pesquisa, identificou-se modelos conceituais sobre o comportamento de estudantes em seus cursos e os fatores que podem leva-los a evadir. Também foram encontrados diversos trabalhos sobre a convergência entre esses estudos e a mineração de dados, levando à descoberta da MDE.

Figura 1 – Metodologia Utilizada



Uma nova revisão se fez necessária, desta vez abordando especificamente trabalhos com mineração de dados educacionais, buscando identificar o estado da arte sobre o tema. Os trabalhos encontrados nessa pesquisa apresentaram uma heterogeneidade tanto nos conceitos educacionais utilizados quanto na própria aplicabilidade das técnicas em suas situações problema. A dificuldade em convergir os conceitos dos trabalhos encontrados e em realizar análises comparativas entre os mesmos, apontou a necessidade de uma estrutura de dados que propusesse conceitos comuns à área e facilitasse o uso em técnicas de mineração de dados, determinando, assim, o objetivo do trabalho atual.

Uma vez definido o objetivo, a análise dos modelos conceituais encontrados na literatura, aliado com a experiência prática na gestão de cursos da modalidade, lastreou a identificação e definição das entidades e atributos mais pertinentes ao tema e a composição do modelo proposto.

Conforme apresentado na Figura 1, a construção do modelo proposto foi sucedida à etapa de sua avaliação, através de um estudo de caso. O planejamento do estudo iniciou-se com a definição de uma hipótese nula e três hipóteses alternativas. Cada hipótese serviu-se de um tratamento específico, sempre analisando os resultados obtidos com a utilização do modelo proposto, em comparação com a sua “não-utilização”. Os resultados do estudo de caso foram analisados, apontando indícios de validação do modelo proposto.

1.4 ESTRUTURA DA DISSERTAÇÃO

A presente dissertação é composta por seis capítulos tendo, no capítulo atual (1. Introdução) a contextualização do tema do trabalho, assim como a apresentação de seus objetivos.

Os capítulos 2 e 3 trazem toda a fundamentação teórica que norteou o trabalho, onde são abordados os principais conceitos pertinentes ao trabalho, tais como definições e conceitos da educação a distância (capítulo 2. Educação a Distância) e definições sobre mineração de dados e a área da mineração de dados educacionais (capítulo 3. Mineração de Dados).

O quarto capítulo apresenta a proposta do autor de um modelo de dados para a utilização em trabalhos de mineração de dados educacionais. Inicialmente é proposto um esquema de dados, com grupos de atributos inerentes a processos educacionais de uma maneira geral e uma análise da qualidade de dados necessária para esses atributos. A partir dessa análise o modelo de dados proposto é apresentado nas formas gráfica e descritiva.

No capítulo 5, o modelo apresentado é posto à prova através da realização de um estudo de caso onde são verificadas a sua aplicabilidade no estudo da evasão de estudantes e principais vantagens de sua utilização. O referido estudo parte de uma base de dados real buscando identificar, através de técnicas de mineração de dados, características que possam ser classificadas como indícios de evasão de estudantes.

As considerações finais do trabalho são apresentadas no capítulo 6, que aponta as suas principais limitações e possibilidades de evolução em futuros trabalhos.

2 EDUCAÇÃO A DISTÂNCIA

O mercado da educação tem se tornado cada vez mais competitivo e global, procurando apoio em tecnologias e processos para atender às pressões populacionais com necessidades de uma educação continuada. No Brasil, com a incorporação da modalidade da educação a distância na Lei de Diretrizes e Bases (Lei 9.394, de 20 de dezembro de 1996), transformando-a definitivamente em política pública de inclusão educacional, a oferta de cursos EAD em instituições de ensino superior públicas e privadas, regional e nacionalmente tem crescido acima dos indicadores de crescimento econômico do país (ABED, 2011).

Esse crescimento tem proporcionado, também, um aumento na preocupação com a qualidade do aprendizado dos seus estudantes e, consequentemente, pesquisas sistemáticas sobre teorias pedagógicas e tecnológicas que a suportem. No afã de organizar informações relevantes sobre o tema, a reflexão e a discussão de conceitos oriundos de um pensamento tradicionalmente presencial se fazem necessárias para dar espaço a novos conceitos e metodologias (ABED, 2011).

Este capítulo busca trazer parte dessas reflexões com a apresentação das principais definições e conceitos da modalidade. Inicialmente, são apresentadas as principais discussões conceituais sobre o que vem a ser a educação a distância para, em seguida, serem discutidas as principais abordagens metodológicas sobre o acompanhamento do estudante na modalidade e os fatores que determinam o seu sucesso ou fracasso e, acima de tudo, sua permanência ou desistência nos cursos que seguem.

2.1 AS DISTÂNCIAS NA EDUCAÇÃO

Educação a Distância pode ser definida como uma forma sistematicamente organizada de autodidatismo na qual o estudante se instrui a partir do material de estudo que lhe é apresentado com o acompanhamento e supervisão de tutores e professores (MUYINDA, 2012). De maneira simplificada, alguns autores definem a Educação a Distancia como um formato de educação onde o professor está geograficamente distante dos seus estudantes. Entretanto, o próprio conceito de distância apresenta diversas nuances na literatura de maneira que a definição dessa modalidade de ensino passa, necessariamente, pelas definições e conceitos do que pode vir a ser distância na educação e os desafios das equipes de cursos à distância podem ser resumidas em como essas distâncias podem ser reduzidas (TORI, 2002).

Para ajudar na compreensão das diferentes formas de Educação a Distância, Tori (2002) propõe uma classificação do conceito de distância a partir das diferentes perspectivas dos envolvidos no processo ensino/aprendizagem.

A distância espacial, física ou geográfica está relacionada à ocupação do espaço físico entre o aluno e os elementos envolvidos: professor, material de estudo e demais colegas. Quando os elementos compartilham do mesmo espaço físico, diz-se que a atividade é local. Em contraponto, quando existe uma separação geográfica entre eles, chamamos o processo de remoto ou à distância. Dessa maneira, a educação tradicional é, notoriamente, local, enquanto que telecursos e cursos por correspondencia são exemplos de educação remota. O autor distingue, também, o termo presencial do local, uma vez que é possível a presença simultânea de estudantes e professores, mesmo separados fisicamente. Em videoconferencias, por exemplo, os estudantes participam presencialmente em relação ao material didático e aos colegas, mesmo que remotos em relação ao professor.

Chama-se de distância temporal o nível de simultaneidade das atividades realizadas. Quando as partes participam simultaneamente, o processo é chamado de síncrono e, quando há uma diferença significativa entre os instantes de ação e reação, o processo é considerado assíncrono. Normalmente, a educação tradicional é local e síncrona, enquanto que na educação remota são possíveis tanto atividades síncronas como assíncronas. A Tabela 1 ilustra alguns exemplos de distância na educação, classificando-os temporalmente, como síncronos e assíncronos, e espacialmente, como locais e remotos.

Tabela 1 - Exemplos de distância na educação

		Distância Temporal	
		Síncrono	**Assíncrono**
Distância Espacial	**Local**	Aula com professor na sala	Videoaula exibida em sala de aula.
	Remoto	Aula transmitida ao vivo por satélite ou internet. Discussão em sala de chat.	Aula pré-gravada, assistida pela internet. Telecurso pela TV. Discussão por e-mail, fórum.

Tori (2002) traz ainda o conceito de distância interativa ou operacional, referindo-se ao nível de participação do estudante no processo. Menor é a distancia operacional quanto mais o estudante interage. Dessa maneira, aulas expositivas apresentam maior distância operacional entre o aluno e o professor do que aulas participativas, assim como trabalhos individuais apresentam maior distância operacional entre alunos quando comparados com trabalhos em grupo, cooperativos.

De forma complementar, alguns autores apresentam o conceito de distância transacional, resultado da combinação entre: a autonomia do aluno em buscar informações e estudar; a estrutura oferecida pelo curso, através de seu material e metodologia; e a possibilidade de diálogo para esclarecer dúvidas, orientar e desenvolver uma empatia com o curso (PASTA, 2011).

2.2 EVASÃO NA EDUCAÇÃO A DISTÂNCIA

A retenção de estudantes nos seus cursos representa, atualmente, o maior problema de instituições de ensino superior de acordo com o Censo 2011 da Associação Brasileira de Educação a Distância (ABED). Conforme apresentado na Tabela 2, Vargas (2007, apud ABED, 2011) apresenta os diferentes conceitos de evasão encontrados na literatura onde, de um modo geral, caracteriza-se por evasão a não continuidade de um estudante em um curso no qual havia se inscrito anteriormente, ou seja, a evasão é um ato de desistência. Uma sutil divergência é apresentada na abrangência das definições, no que se refere à inclusão ou não de estudantes que não chegaram a iniciar os estudos nos seus cursos.

Tabela 2 - Diferentes definições de evasão

Autor	Definição	Abrangência da Definição
Utiyama e Borba (2003)	Evasão é entendida como a saída definitiva do aluno de seu curso de origem, sem concluí-lo.	Não foi estabelecido nenhum critério de tempo no curso para a saída do aluno.
Maia e Meireles (2005)	Evasão consiste em alunos que não completam cursos ou programas de estudo, podendo ser considerados como evadidos alunos que se matriculam e desistem antes mesmo de iniciar o curso.	Especifica que, mesmo alunos que nunca começaram o curso devem ser considerados evadidos.
Abbad, Carvalho e Zerbini (2005)	Evasão refere-se à desistência definitiva do aluno em qualquer etapa do curso.	Não deixa claro se evasão se aplicaria apenas aos alunos que chegaram a iniciar o curso ou se abrangeria tambem os que apenas se matricularam e nunca iniciaram o curso.

Fonte: Vargas 2007 *apud* (ABED, 2011).

Do ponto de vista de sua comunicação à instituição de ensino, a evasão pode ser informada ou, numa ação de abandono, silenciosa, não informada. Quanto à sua perenidade, a desistência pode ser temporária ou definitiva. A Tabela 3 apresenta as principais situações de evasão em cursos, classificando-as quanto à sua comunicação e à sua perenidade.

Tabela 3 - Evasão: Perenidade X Comunicação

	Temporária	**Definitiva**
Informada	Trancamento	Cancelamento, Transferência
Não Informada	Abandono	Abandono

Apesar de todos os tipos de evasão serem igualmente danosos, a evasão silenciosa, não informada, merece especial atenção dos gestores de cursos por não permitirem uma identificação imediata da opção do estudante em não concluir seu curso. Para mapear as razões de um abandono, é necessário provocar um contato com o estudante evadido e interrogá-lo diretamente.

Rovai (2002 *apud* PYLE, 1999) revela que as taxas de evasão informada na educação superior são de aproximadamente 23%, enquanto que a evasão geral tem atingido 50%, dependendo da natureza do curso e da definição utilizada para evasão.

No Brasil, segundo o Censo 2011 da ABED, a média de evasão nos cursos da modalidade ofertada está na ordem de 18%. As Tabelas 4 a 6 a seguir apresentam a distribuição dos dados coletados pelo censo segmentados por natureza jurídica da instituição (pública ou privada), por porte da instituição, por região na qual estão sediadas e por tipo de curso (livre ou autorizado).

Do ponto de vista da natureza jurídica da instituição (Tabela 4), a evasão é maior em instituições públicas do que em instituições privadas, com uma predominância, em ambos os casos, em cursos livres sobre os autorizados.

Tabela 4 - Evasão no Brasil por Natureza Jurídica da Instituição,

Natureza Jurídica da Instituição de Ensino	**Tipos de Cursos**	
	Autorizados (%)	**Livres (%)**
Instituições Públicas	22,16	30,9
Instituições Privadas	15,8	20

Fonte: Censo ABED (2011).

Também de acordo com o censo, as maiores concentrações de evasão encontram-se em regiões com maior prevalência de cursos: nas regiões Sudeste, Centro-Oeste e Sul do país, em detrimento das regiões Norte e Nordeste (Tabela 5). Chama a atenção o fato de que nessas regiões os indicadores de evasão em cursos autorizados são aproximadamente os mesmos dos de cursos livres, situação atípica nas outras distribuições apresentadas pela pesquisa.

Tabela 5 - Evasão no Brasil por Região do Brasil

Região do Brasil	Tipos de Cursos	
	Autorizados (%)	Livres (%)
Norte	0	18,7
Nordeste	14,8	33,8
Centro-Oeste	18	19,8
Sudeste	21,74	21,1
Sul	16,4	16,4

Fonte: Censo ABED (2011).

Outro fato relevante apresentado pelo censo da ABED reside na proporcionalidade entre o índice de evasão e o porte das instituições – quanto maior a instituição de ensino, maior é a incidência do fenômeno da evasão, tanto em cursos livres quanto em autorizados (Tabela 6).

Tabela 6 - Evasão no Brasil por Porte da Instituição de Ensino

Porte da Instituição de Ensino	Tipos de Cursos	
	Autorizados (%)	Livres (%)
Micro	5,5	18,3
Pequena	12,7	18,6
Média	13,2	20
Grande	18,5	23,3

Fonte: Censo ABED (2011).

Ao analisar o fenômeno da evasão, as pesquisas realizadas convergem ao determinar a importância de mapear as justificativas para a escolha do estudante em evadir. Pyle (1999) comenta que:

> [...] dados os custos, financeiros e outros, associados à evasão, é óbvio que a maioria das faculdades e universidades gostaria de entender melhor as suas forças motrizes. Se essas instituições pudessem entender melhor por que os indivíduos abandonam o ensino superior poderiam tentar mudar as suas políticas de seleção, ou a forma com que lidam com os seus alunos, tendo em vista a reduzir as suas taxas de evasão.

Ao analisar as causas de evasão, o Censo ABED (2011) aponta que :

> As pesquisas têm indicado que a evasão está relacionada à dificuldade e à longa duração dos cursos; à pouca adequação dos cursos às necessidades dos alunos; à complexidade e à quantidade de atividades escritas exigidas; à falta de tempo do aluno; aos problemas financeiros; às más condições de estudo no trabalho e em casa e à falta de acompanhamento dos alunos pela tutoria.Um curso a distância concorre com a vida do aluno. Como é possível realizar o curso a qualquer tempo e em qualquer lugar, muitas vezes as atividades de estudo são adiadas várias vezes e, quando o aluno decide retomar o curso, nem sempre tem condições de acompanhá-lo

> e acaba desistindo. É preciso empenho e disciplina para conseguir completar um curso a distância. E sempre o término de um curso está associado ao sucesso e não conseguir terminar, ao fracasso sentido pelo aluno.

De acordo com os respondentes do censo, as razões mais comuns apontadas para justificar a evasão estão relacionadas ao tempo de estudo (falta de tempo para estudar, 30,8%; acúmulo de atividades no trabalho, 24,5%; e viagens a trabalho com 9,1%), o que constitui, justamente, o argumento comercial principal utilizado pelas instituições para captar estudantes para os cursos da modalidade.

O Censo aponta que 73,30% das instituições participantes, que responderam ao item sobre evasão, consideram o controle da evasão como importante na gestão de cursos de Educação a Distância, conforme apresentado na Tabela 7.

Tabela 7 - Importância do controle da evasão

	Importante	Pouco Importante	Não importante	Não respondido	Total
Cursos Autorizados	89	3	44	136	**272**
Cursos Livres	46	4	2	30	**82**
Cursos Corporativos	16	2	0	17	**35**
TOTAL	**151**	**9**	**46**	**183**	**389**

Fonte: Censo ABED (2011).

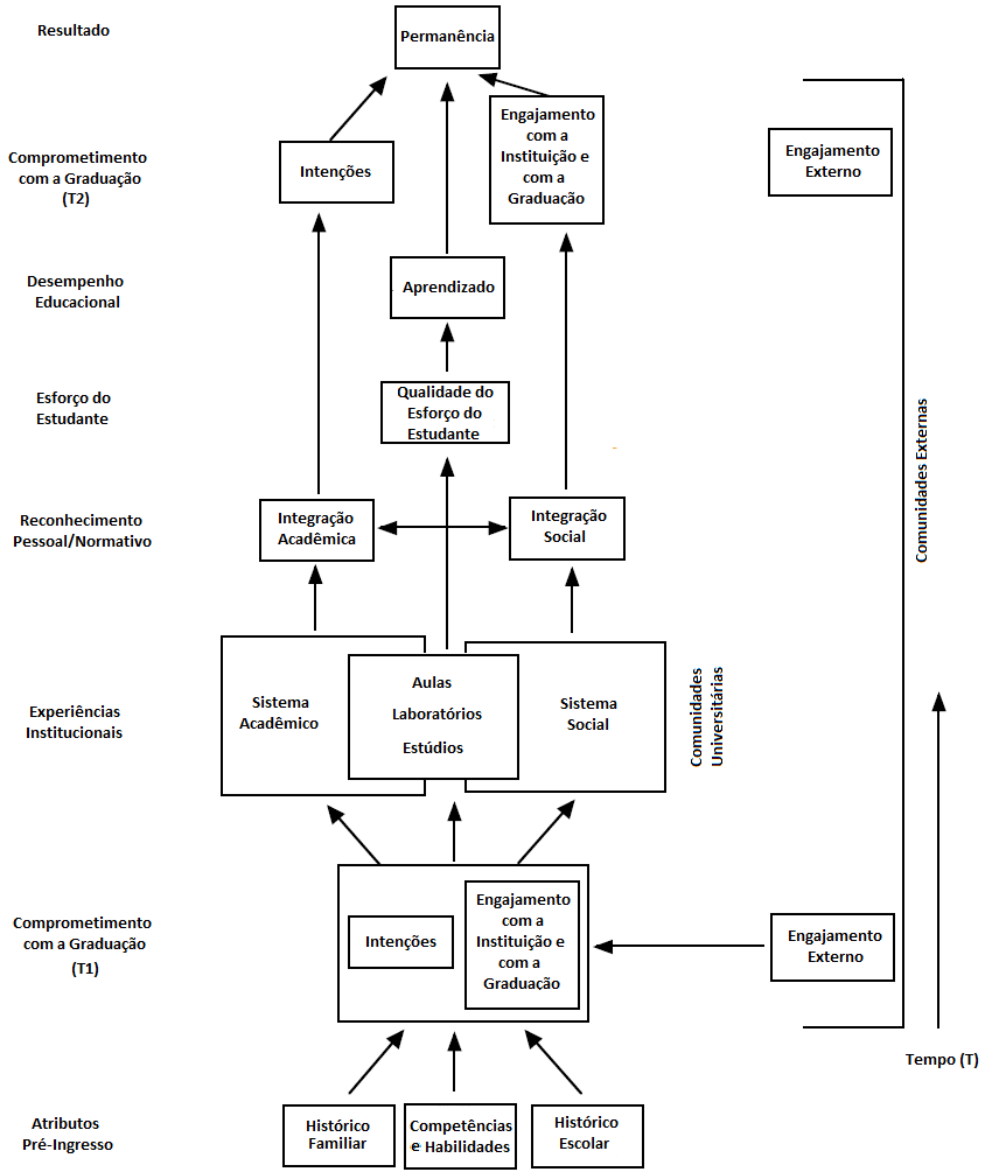
Retenção de estudantes e taxas de evasão na Educação a Distância tem sido um assunto intensamente investigado e discutido nas últimas sete décadas (TYLER-SMITH, 2006). Neste mesmo trabalho, o autor apresenta, complementarmente, que a importância da investigação do fenômeno da evasão na modalidade deve ser levada em consideração por dois motivos principais: primeiro, para diagnosticar a efetividade do custo-benefício do ensino online, quando comparado com o ensino tradicional; e, segundo, para determinar a melhor abordagem na manutenção do estudante nos seus cursos e, consequentemente, no seu aprendizado. Tyler-Smith (2006) indica, também, uma concordância nos trabalhos, em afirmar que as motivações para evadir são diversas e complexas, não havendo soluções simples para o seu combate.

Nas próximas subseções, são apresentadas algumas análises realizadas sobre o fenômeno da evasão e as principais características e situações que podem determinar a decisão de um estudante em permanecer no seu curso. Essas análises são representadas por modelos conceituais que resumem suas principais constatações.

2.2.1 O Modelo de Integração de Tinto

Tinto (1975 *apud* MCCUBBIN, 2003) oferece um modelo conceitual longitudinal com a intenção de explicar os aspectos e processos que podem influenciar uma decisão individual de abandonar um curso, além de apresentar de que maneira esses processos interagem para produzir a evasão. O Modelo de Integração Estudantil – SIM (do inglês, *Student Integration Model*) apresentado originalmente em 1975 foi extensamente discutido, conforme McCubbin (2003), o que gerou diversas revisões, culminando no modelo apresentado pelo próprio Tinto em 1997 e exibido na Figura 2.

Figura 2 - Adaptação do modelo de evasão de estudantes revisado por Tinto em 1997



Fonte: Mccubbin (2003).

O núcleo do modelo de Tinto (1997) está no nível de integração do estudante com os aspectos sociais e acadêmicos da universidade, além do seu comprometimento com a instituição e com o seu objetivo em se graduar.

Pelo modelo, várias características individuais afetam o nível de comprometimento do aluno, ainda enquanto interessado em se matricular, tanto em relação a sua meta de obtenção do grau como com a instituição na qual está se inscrevendo. Tinto (1997) destaca algumas características principais que influenciam o comprometimento do aluno com a sua graduação e com a instituição de ensino - experiências pré-universitárias (experiências sociais e acadêmicas, tais como sua média global no histórico escolar e outras realizações acadêmicas e sociais), antecedentes familiares (status social, valores e expectativas), além de características como raça, sexo e capacidade acadêmica.

Tinto (1997) afirma que há uma relação direta das expectativas educacionais de um indivíduo com sua probabilidade de evasão. Alguns alunos veem a instituição na qual ingressam como fundamental para sua carreira no futuro, outros alunos, porém, poderiam estar tão satisfeitos em outra instituição quanto na que frequentam. Alunos que colocam uma grande importância na faculdade que frequentam são significativamente mais propensos a persistir em detrimento de eventuais problemas acadêmicos ou sociais.

A percepção do estudante de sua própria educação é muito importante em sua decisão de evadir. Tinto (1997) indica que os alunos avaliam a sua própria experiência de educação em termos de custo-benefício e, se eles sentem que poderiam obter maior benefício a um custo menor ou igual fora da instituição, eles são suscetíveis a abandonar.

A evasão também pode ser influenciada por aspectos da personalidade do aluno. Desistentes tendem a apresentar determinados traços de personalidade como maior impulsividade, menor compromisso emocional com a educação, são incapazes de aproveitar o máximo de experiências passadas, são mais instáveis, mais ansiosos e são excessivamente inquietos. De acordo com Tinto (1997), o desempenho acadêmico age como uma espécie de recompensa extrínseca, enquanto o desenvolvimento intelectual é mais uma recompensa intrínseca. Além disso, o sexo do aluno também influencia a importância que eles dão ao desempenho nas disciplinas, na medida em que desenvolvimento intelectual aparenta ser mais importante na determinação da persistência para mulheres do que para os homens.

Tinto (1997) também indica que níveis muito elevados de integração social podem levar a déficits no desempenho acadêmico, mas não necessariamente levam a evasão como, por exemplo, quando a integração acontece com um grupo que tenha "fortes orientações

acadêmicas". O modelo de integração de estudantes também ilustra a afirmação de Tinto (1997) de que integração acadêmica e social, junto com comprometimento com a graduação e com a instituição de ensino, não estão separados e distintos, mas possuem uma relação influente um sobre o outro. Também os graus de comprometimento com a graduação e com a instituição não são, por si, suficientes para levar alguém a evadir. De acordo com o autor, quando um estudante, por exemplo, tem suficiente comprometimento com a sua graduação, ele pode permanecer em uma instituição com a qual tenha pouco comprometimento.

No aspecto cronológico, Tinto (1997) aponta que, fundamentalmente, a integração social se apresenta em curva ascendente pelo primeiro ano, quando é muito importante para os alunos desenvolverem relações sociais que sirvam como uma rede de apoio. Se o aluno não se sente suficientemente integrado no primeiro ano, é provável que venha a abandonar. Por outro lado, se o aluno não se sente suficientemente integrado socialmente no final do seu curso de graduação, é pouco provável que ele venha a evadir uma vez que, neste estágio, está mais comprometido com a conclusão de seu curso. No último ano de uma graduação, integração acadêmica aparenta ser mais importante para a maioria dos estudantes do que a integração social.

2.2.2 O Modelo de Kember

Kember (1989 *apud* TYLER-SMITH, 2006) propôs um modelo conceitual de evasão para educação a distância, alegando que o Modelo de Tinto (1975) foi realizado com alunos presenciais menores de 25 anos. Tinto sugeriu que interação social e integração bem sucedida dos estudantes na cultura acadêmica da instituição são características da experiência do aluno que contribuem significativamente para a sua provável persistência em seus estudos e, estar no campus, torna isso muito mais provável de acontecer do que em cursos a distância.

Já o modelo de Kember aponta que alunos de cursos a distância são, majoritariamente, adultos maduros com família, o que introduz outros fatores, como a capacidade do aluno para gerenciar as demandas de estudo com compromissos de trabalho, familiares e sociais.

De acordo com Kember, situações familiares como o número e idade dos dependentes, condições de habitação e as pressões das responsabilidades, tais como sustentar a família, podem todas ter um impacto significativo na decisão de um aluno a distância interromper seu curso. Kember também identifica os níveis de renda, sexo e distância geográfica da instituição de contribuir para a evasão.

2.2.3 O Modelo de Boyles

Tyler-Smith (2006) traz também o modelo de Boyles (2000, *apud* DAGGER; WADE, 2004) que evoluiu a partir do modelo de Kember (1989) abordando, especificamente, retenção em EaD. Ele identificou três conjuntos de variáveis relacionadas à evasão:

- Variáveis relacionadas ao histórico do aluno: maturidade, circunstâncias pessoais e experiências anteriores;
- Variáveis ambientais: família, compromissos sociais e de trabalho;
- Variáveis acadêmicas: histórico escolar e a aderência entre o aluno e o assunto que está sendo estudado.

A estes conjuntos de variáveis se somam outras variáveis individuais, tais como autoconfiança acadêmica, resultados acadêmicos, facilidade de integração com a instituição, tamanho da instituição, habilidades de integração social e composição psicológica do aluno.

2.2.4 O Modelo de Berge e Huang

O modelo de Berge e Huang (2004) aperfeiçoa os anteriores organizando as variáveis em três outros grupos principais:

- Variáveis pessoais: idade, etnia, gênero, nível de renda, experiência acadêmica anterior e outras características pessoais, como a autodidatismo, organização e motivação pessoal;
- Variáveis institucionais: atitude institucional, valores e crenças, características acadêmicas como sistemas estruturais e processos, apoio ao aluno e grau de congruência entre as necessidades individuais dos alunos e a postura filosófica da instituição;
- Variáveis circunstanciais: natureza e a qualidade da interação da instituição com o aluno; interações acadêmicas, projeto do curso e facilitação, bem como interações específicas da vida do aluno nos seus círculos de trabalho e família, além de sua responsabilidade e satisfação.

Embora todos esses estudos tenham trazido contribuições importantes à compreensão do fenômeno da evasão, entender como aplicá-los para aumentar os índices de retenção ainda é um desafio para pesquisadores e gestores de cursos e instituições de ensino.

2.3 CONSIDERAÇÕES SOBRE OS MODELOS ANALISADOS

O modelo conceitual de Tinto (1997), que serviu de base para estudos posteriores (KEMBER, 1989 *apud* TYLER-SMITH, 2006; BOYLES, 2000 *apud* DAGGER; WADE, 2004; BERGE; HUANG, 2004), introduz a importância de fatores como experiência (histórico familiar, escolar, institucional), integração (social, acadêmica) e comprometimento (com a instituição e com a formação) para a permanência dos estudantes nos cursos. O seu modelo apresenta agrupamentos de características, suas inter-relações e sua disposição cronológica, culminando na permanência do estudante no curso no qual está matriculado. Apesar de apontar diversos atributos como importantes para a permanência de estudantes nos cursos, em nenhum instante, o autor determina o grau de importância de cada um deles.

Tinto (1997) aponta que características relacionadas com a experiência do estudante estão diretamente associadas com o comprometimento e o engajamento do estudante com a sua graduação e com a instituição de ensino que, por sua vez, estão relacionadas com a decisão do estudante em permanecer. Sua classificação de atributos evoluiu com as contribuições de outros pesquisadores, como Kember (1989 *apud* TYLER-SMITH, 2006).

Alguns atributos e classificações sucederam o modelo de Tinto (1997). A Tabela 8 apresenta algumas dessas classificações.

Tabela 8 - Agrupamentos de atributos segundo modelos conceituais

Autores	Agrupamento	Detalhamento
Boyles (2000, *apud* Dagger & Wade, 2004)	Individuais	Maturidade, circunstâncias pessoais, experiências anteriores.
Boyles (2000, *apud* Dagger & Wade, 2004)	Ambientais	Família, compromissos sociais e de trabalho.
Boyles (2000, *apud* DAGGER & WADE, 2004)	Acadêmicas	Histórico escolar, aderência do estudante ao curso.
Berge e Huang (2004)	Pessoais	Etnia, gênero, nível de renda, experiência acadêmica anterior, autodidatismo, organização, motivação pessoal.
Berge e Huang (2004)	Institucionais	Atitude institucional, valores e crenças da instituição, sistemas e processos, apoio ao aluno, congruência entre as expectativas dos estudantes e a postura da instituição.
Berge e Huang (2004)	Circunstanciais	Natureza e qualidade da interação do aluno com a instituição. Interações específicas da vida do aluno nos seus círculos profissionais e familiares.

2.4 CONCLUSÃO

Amplamente discutida nos últimos anos, a Educação a Distância ainda se apresenta em processo de amadurecimento com modelos conceituais distintos e algumas vezes divergentes. Da mesma forma, a análise de fenômenos já conhecidos no modelo presencial sugere que as motivações de estudantes na modalidade à distância são necessariamente distintas daquelas que orientam estudantes de cursos presenciais.

A realização de pesquisas que discutam modelos e paradigmas atuais à luz das novas práticas e suas necessidades de mudança torna-se fundamental para a evolução da modalidade. Nesse sentido, analisar os dados gerados pelas instituições e mecanismos tecnológicos utilizados torna-se tarefa primordial no sucesso dessas pesquisas. De acordo com o Censo ABED (2011):

> As informações quantitativas, análises qualitativas, projeções e avaliações detalhadas podem indicar onde estávamos, onde estamos e para onde estamos indo e auxiliar na tomada de decisões. Os dados e informações oferecem maior segurança na tomada de decisões, seja para instituições, seja para empresas ou mesmo para o governo e órgãos reguladores no estabelecimento de diretrizes para a política educacional

A análise de dados e informações produzidos na gestão e operação de cursos a distância, no nível de detalhe e tempo necessários para o apoio a processos decisórios, demanda a utilização intensiva de recursos computacionais.

Dentre os mecanismos da tecnologia da informação, a Mineração de Dados, através de suas tarefas e técnicas, se destaca pela sua capacidade em analisar e descobrir informações estratégicas, de maneira rápida e eficiente, ocultas em grandes volumes de dados (GOEBEL et al, 1999).

3 MINERAÇÃO DE DADOS

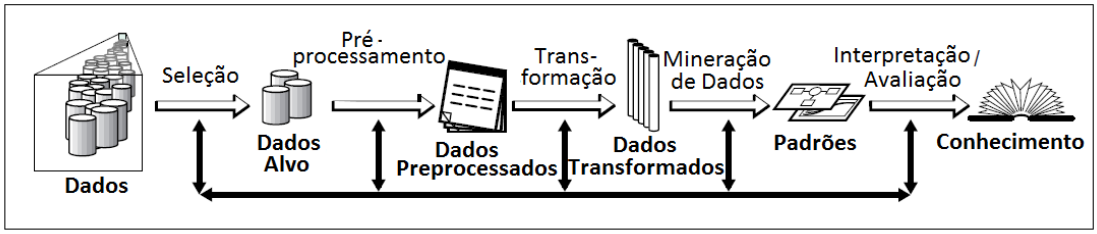
A mineração de dados, com suas tarefas e técnicas, representa a fase principal do KDD, sigla em inglês para *Knowledge Discovery in Databases*, ou Descoberta de Conhecimento em Bases de Dados. O KDD concentra os conceitos e processos para a utilização de bases de dados em processos de tomada de decisão transformando, através de processamentos sucessivos, dados brutos em informações relevantes e conhecimento útil (GOMES, 2002).

Este capítulo apresenta as definições do KDD, concentrando-se nas fases de pré-processamento e mineração de dados, com suas principais tarefas e técnicas.

3.1 KDD – DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

O KDD, segundo Fayyad *et al* (1996), é formado por uma sequência de etapas que, uma vez executadas, resultará na geração do conhecimento útil. Este processo é composto, conforme apresentado na Figura 3, pelas seguintes atividades: seleção dos dados utilizados; sua preparação para a utilização através de um tratamento prévio (pré-processamento); sua subsequente transformação para um formato adequado; o processamento do conjunto de dados por algoritmos especialistas (mineração de dados) e, finalmente, a análise dos resultados obtidos para a sua aplicação no processo decisório (interpretação/avaliação).

Figura 3 - As fases do KDD



Fonte: Fayyad (1996).

A etapa de seleção diz respeito à análise da disponibilidade e relevância dos dados existentes nas suas diversas fontes. Uma escolha errada dos dados pode levar à geração de informações errôneas, prejudicando a tomada de decisão (BATISTA, 2003).

Uma vez identificado e extraído o conjunto de dados relevantes, faz-se necessário prepará-los para a aplicação das técnicas de mineração. Esse tratamento consiste em seu processamento, sob diferentes aspectos, tornando-os qualificados para a mineração.

Preliminarmente, devem ser estabelecidas normas de representação dos dados utilizados, que geralmente são provenientes de origens distintas, com diferentes formatos de armazenamento, determinando a sua padronização na base a ser gerada. De acordo com Pyle (1999), a definição das normas de representação de uma base para mineração sofre influência direta da técnica de mineração desejada e, em alguns casos, da ferramenta de mineração utilizada. Após a definição das normas de representação, os dados identificados devem ser extraídos de suas fontes e integrados em um único repositório.

A fase de pré-processamento é constituída por diversas atividades de tratamento dos dados selecionados, tais como verificação semântica, enriquecimento, deduplicação, unificação e discretização, detalhadas a seguir.

A verificação dos dados armazenados quanto à sua consistência semântica determinará se, ao integrá-los de fontes diferentes, foi produzida alguma inconsistência. Problemas com compatibilidade dos elementos formadores de endereços - CEP, bairro, cidade, Estado, país, são exemplos comuns de inconsistências identificadas nesse processo.

O enriquecimento dos dados consiste no preenchimento de lacunas existentes em registros incompletos. Esse processo somente será possível com a disponibilidade de fontes de consulta externa aos dados ou através da dedução dos valores das lacunas a partir de dados disponíveis, como a identificação de um logradouro não preenchido a partir do CEP preenchido, por exemplo, ou do cálculo da idade de alguém (não preenchida) a partir da data de nascimento disponível.

O processo de deduplicação dos dados é definido pela identificação e supressão de registros duplicados, que dizem respeito a uma mesma entidade. Ao analisar, por exemplo, uma base de funcionários de uma organização, cada funcionário deve estar representado por um único registro. Caso seja identificada uma duplicação, o registro excedente deve ser eliminado da base.

Eventualmente, ao integrar os dados para a mineração, faz-se necessário representar de forma única registros múltiplos de uma mesma entidade. A esse processo chamamos de unificação dos dados. Como exemplo, pode-se citar a representação da média mensal de consumo de usuários de determinada loja onde, apesar de cada usuário realizar ao longo do tempo diversas compras, do ponto de vista da análise do usuário, no seu registro (único) na base de dados analisada, os dados dessas compras só podem ser armazenados de maneira consolidada, através de somatórios, valores médios, mínimos máximos.

Dependendo da técnica de mineração utilizada, pode ser necessária a discretização de dados, ou seja, a representação de valores numéricos contínuos através de valores textuais

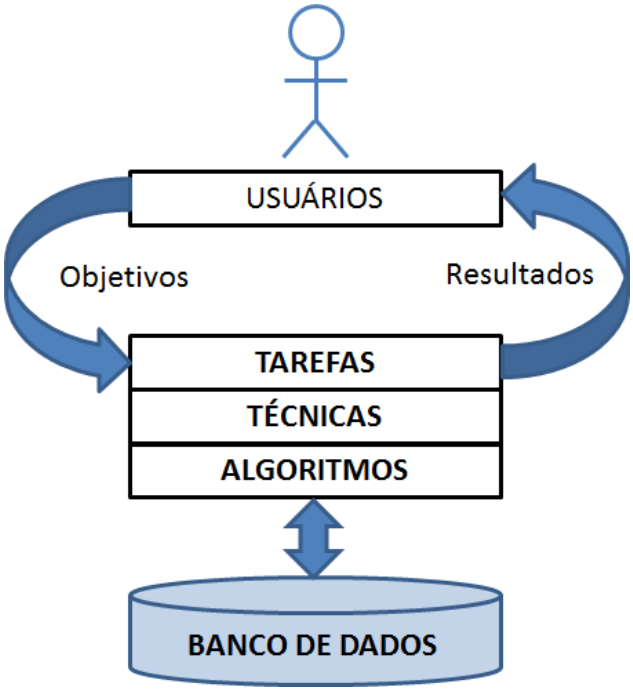
discretos, representando faixas de valores ou conjuntos de valores. Um exemplo comum desse processo é a representação de valores de renda de cidadãos através de faixas salariais – até cinco salários mínimos, de seis a dez salários mínimos, acima de 10 salários mínimos.

As ferramentas de mineração de dados existentes utilizam diferentes formatos de arquivos de entrada. Geralmente os dados são alimentados através de arquivos texto, em formato CSV (*comma separated values*) ou ARFF (*Attibute-Relation File Format*). A fase de organização dos dados no formato exigido pela ferramenta de mineração é denominada de transformação e conclui a preparação dos dados para a mineração.

Uma vez pré-processados e transformados, os dados são, enfim, submetidos aos algoritmos de mineração de dados e, seus resultados, na forma de padrões de comportamento dos dados analisados, tornam-se disponíveis para a interpretação e análise à luz do processo decisório do negócio investigado.

Segundo Pasta (2011), o modo como os dados são minerados denomina-se de tarefas e, conforme o objetivo pretendido, podem ser utilizadas uma ou mais tarefas com diversas abordagens ou técnicas na sua utilização. A aplicação dessas técnicas pressupõe a utilização de um algoritmo especializado para a sua implementação. A partir de seus objetivos no processo decisório, o usuário, gestor, para atingir os resultados esperados, utiliza tarefas de mineração de dados através da aplicação de suas técnicas e algoritmos. Os algoritmos são *software* codificados com a função específica de analisar os dados disponíveis e executar as técnicas de mineração (Figura 4).

Figura 4 - A relação entre tarefas, técnicas e algoritmos



A mineração de dados, com suas técnicas e algoritmos, reúne conhecimento científico interdisciplinar proveniente das áreas da estatística, bancos de dados e da inteligência artificial para extrair novos conhecimentos úteis, relevantes e não triviais que estejam ocultos em grandes massas de dados (TINTO, 2006).

Autores dessas três "áreas matrizes" da mineração de dados apresentam diferentes definições para a disciplina. Hand *et al* (2001) definem, do ponto de vista da estatística, que mineração de dados é "a análise de (frequentemente grandes) conjuntos de dados para encontrar relacionamentos inesperados e para resumir os dados em novas formas que sejam compreensíveis e úteis para o proprietário dos dados". Analisando sob a perspectiva do estudo de bancos de dados, a mineração de dados, segundo Han *et al* (2006), representa "o processo de descoberta de conhecimento interessante a partir de grandes quantidades de dados armazenados em bases de dados, *data warehouses* ou outros repositórios". Para Witten et al (2005), sob a ótica da inteligência artificial, a mineração de dados define a "extração de informação implícita, previamente desconhecida e potencialmente útil a partir de dados".

Nas próximas seções, são apresentadas as principais tarefas e técnicas de mineração de dados, suas distinções e objetivos.

3.2 TAREFAS DA MINERAÇÃO DE DADOS

As principais tarefas da mineração de dados, na literatura, possuem classificações de acordo com o ponto de vista analisado. Segundo CIOS (1998 *apud* KAMPFF, 2009), na perspectiva da descoberta do conhecimento, as tarefas podem trazer descoberta direta ou indireta do conhecimento. Na descoberta direta do conhecimento um atributo focal é definido e a mineração deve explicar como os demais atributos se relacionam a ele. Através da descoberta direta, por exemplo, pode-se analisar como dados cadastrais de um indivíduo (sexo, idade, estado civil, nível de escolaridade) influenciam sua renda. Por outro lado, na descoberta indireta objetiva-se descobrir relações e padrões implícitos através do cruzamento de todos os dados disponíveis. Pasta (2011) nomeia como preditivas as tarefas com descoberta de conhecimento direta e descritivas aquelas tarefas com descoberta indireta.

Dentre as tarefas de mineração de dados com descoberta de conhecimento direta, destacam-se a classificação e a estimativa, Berry e Linoff (1997 *apud* KAMPFF, 2009).

3.2.1 Classificação

A tarefa deve associar os registros minerados a classes pré-estabelecidas, de acordo com as suas semelhanças, buscando associar cada registro da base a um único rótulo categórico chamado de classe, Rabelo (2007 *apud* PASTA, 2011). Segundo Goldschmidt e Passos (2005 *apud* PASTA, 2011), classificar significa examinar as características de um objeto e atribuir-lhe uma classe pré-definida. Uma vez conhecida a relação dos dados com as classes, torna-se possível prever a classificação de um novo dado inserido no grupo.

Os modelos de classificação trabalham com dois tipos de atributos: o atributo objetivo, correspondendo à variável que define as categorias ou classes pré-definidas e, os atributos preditivos, usados para estabelecer as relações e inferir a classe de novos dados.

3.2.2 Estimativa ou Regressão

A partir da análise de dados numéricos contínuos dos elementos disponíveis, procura-se estimar um valor para um novo elemento. Utilizando métodos estatísticos e redes neurais, a tarefa procura estabelecer uma função, linear ou não, a partir da relação entre os atributos preditivos com os diversos valores de atributos objetivos, criando uma fórmula para o cálculo de valores em novos registros.

Um bom exemplo de aplicação de regressão é o cálculo do valor de um imóvel a partir dos valores de outros imóveis e suas características. Ao relacionar valores de imóveis com características como área útil, área construída, benfeitorias, localização, entre outros, a tarefa constrói uma função que calcula o impacto de cada um desses itens no valor final dos imóveis. Dessa forma, ao aplicar a função em um novo imóvel e suas características, obtemos uma estimativa de seu valor.

Enquanto as tarefas de classificação e a estimativa são definidas como descoberta de conhecimento direta, as tarefas mais comuns com descoberta de conhecimento indireta são a associação e a clusterização (PASTA, 2011).

3.2.3 Associação

Essa tarefa objetiva identificar tendências que facilitem a compreensão de padrões em grandes bases de dados ou, nas palavras de Barioni (2001 *apud* PASTA, 2011), “envolve a descoberta de regras de associação que indiquem correlações interessantes entre os objetos de um dado banco de dados”.

Sua maior aplicabilidade está no setor comercial, buscando identificar padrões de consumo que possam orientar as iniciativas de vendas. O caso mais conhecido de mineração de dados é um exemplo de regra de associação que, apesar de comprovadamente fictício, segundo artigo do periódico londrino, Financial Times, em sua edição de sete de fevereiro de 1996 (WITTEN *et al*, 2005), continua sendo reiteradamente referenciado na literatura científica pela sua natureza didática – determinada rede de supermercados, ao analisar os dados de compras realizadas pelos seus clientes, identificou a associação inusitada entre as vendas de fraldas e cerveja. Supostamente, ao comprar fraldas para seus bebês, pais aproveitam e compram cerveja para si. Identificada a relação entre os produtos, as equipes de marketing trabalharam para aumentar suas vendas dos produtos.

3.2.4 Clusterização

Também conhecida como segmentação ou agrupamento (PASTA, 2011), a clusterização consiste na identificação de grupos de registros que apresentam características similares. Os conjuntos encontrados, chamados de clusters, podem determinar padrões de comportamento dos dados. Segundo Martinhago (2005 *apud* PASTA, 2011):

> Um cluster pode ser definido como um conjunto de objetos agrupados pela similaridade ou proximidade e, a segmentação pode ser definida como a tarefa de segmentar uma população heterogênea em um número de subgrupos (ou *clusters*) mais homogêneos possíveis, de acordo com alguma medida.

Na clusterização as classes são resultantes dos dados minerados, ao contrário da tarefa de classificação, onde as classes são pré-determinadas pelo pesquisador. Dessa forma, ao invés de atribuir um rótulo a novos clientes a partir de uma classificação pré-estabelecida, a clusterização permite, por exemplo, agrupar clientes de uma loja que possuem comportamentos de compra similares.

O maior desafio em utilizar a clusterização é que nem sempre os agrupamentos gerados são compreensíveis do ponto de vista da análise (BERRY ; LINOFF, 1997 *apud* KAMPFF, 2009) e, na medida em que a quantidade de atributos utilizados e o volume de dados crescem, essa análise aumenta consideravelmente sua complexidade.

3.3 TÉCNICAS DE MINERAÇÃO DE DADOS

Para a execução das diversas tarefas da mineração de dados, é necessário fazer uso de métodos ou técnicas. Basicamente, as técnicas são suportadas por algoritmos especialistas e aplicadas através de ferramentas de software desenvolvidas para esse fim. De maneira geral, não existe um método universal para a mineração de dados, pois cada técnica aborda melhor alguns problemas do que outros, cabendo ao “minerador” identificar empiricamente qual o algoritmo utilizado e especificar que tipo de informação o algoritmo deve buscar no seu conjunto de dados Martinhago (2005 *apud* PASTA, 2011).

As técnicas existentes são extensões de métodos analíticos conhecidos, amplificados pelo aumento do poder computacional, da capacidade de armazenamento de dados e à redução dos custos de processamento (PASTA, 2011). Diversas técnicas devem ser testadas e combinadas para permitir uma comparação de resultados, viabilizando a identificação do conjunto de técnicas mais adequado ao problema (CAMILO *et al*, 2009).

De acordo com Chen (1996), uma vez que a mineração de dados levanta muitas questões de pesquisa desafiadoras, as aplicações diretas de métodos e técnicas desenvolvidas em estudos voltados a aprendizagem de máquina, estatísticas e sistema de banco de dados não são suficientes para resolver estes problemas, sendo necessário realizar pesquisas dedicadas a inventar novos métodos de mineração de dados ou desenvolver técnicas integradas para a melhoria da eficiência e eficácia da mineração de dados.

A seguir, são apresentadas algumas das principais técnicas de mineração de dados.

3.3.1 Regras de Associação

A técnica e os algoritmos para descoberta de conhecimento por regras de associação buscam identificar relações frequentes entre os dados de uma base. A ideia geral é buscar por padrões de associações fortes entre os registros (utilizando-se do conceito de frequência) e as categorias. Basicamente, consiste em dois passos: primeiro, os dados de treinamento são analisados para que se obtenham os itens mais frequentes. Em seguida, estes itens são usados para a geração das regras (CAMILO *et al*, 2009).

Uma regra de associação é uma expressão na forma A=> B ou “se A, então B” permitindo inferências associativas entre os itens de uma transação. Dessa forma, uma organização comercial, por exemplo, pode identificar que 90% das vezes em que uma compra qualquer acima de R$1.000,00 é efetuada, o pagamento é realizado com cartão de crédito. Dessa forma, dizemos que {Compra > R$ 1.000} => {Pagamento com cartão}. Formalmente, Martinhago (2005 *apud* PASTA, 2011) define uma regra de associação como X => Y (X implica Y) em que X e Y são conjuntos de itens da base de dados e $X \cap Y = \{\}$. X é o antecedente da regra (lado esquerdo) e Y é o consequente da regra (lado direito), podendo envolver qualquer número de itens em cada lado da regra.

Kampff (2009) apresenta um exemplo de aplicação de regra de associação a partir de uma base de 13 estudantes contendo nome, quantidade de atividades entregues, desempenho (baixo, médio, alto), sexo e resultado (aprovado, reprovado). Com a aplicação da técnica, foram geradas 10 regras de associação independentes indicando, entre outras, que 100% das vezes em que o estudante com desempenho baixo entregou somente três atividades, o resultado foi de reprovação. A Figura 5, a seguir, ilustra a representação das regras de associação resultantes dessa aplicação.

Os algoritmos de associação utilizam os parâmetros de confiança e suporte para orientar a sua aplicação. O parâmetro de confiança indica a relação entre o número de vezes em que dois itens A e B aparecem na mesma transação com o número de vezes em que o item antecedente A aparece no conjunto de transações. Ao indicar um nível de confiança aceitável, estabelece-se o grau de exigência esperado das regras a serem identificadas. O parâmetro de suporte indica o número de ocorrências mínimo de casos em que as regras se aplicam no conjunto de dados.

Para Kampff (2009), a boa utilização desses parâmetros interfere diretamente na quantidade e na qualidade das regras geradas, exemplificando que, ao minerar um grande conjunto de dados, muitas regras surgirão e um ajuste de 100% de confiança com apenas um caso de suporte indicará uma situação tão peculiar (e talvez única) que dificilmente poderá ser generalizada.

Figura 5 - Exemplo de regras de associação em formato textual

1) *If* **Atividades_Entregues** *is* **3,00** and **Resultado** *is* **Reprovado**
Then **Desempenho_Medio** *is* **Baixo**
Rule's probability: **1,000**
The rule exists in **2** *records.*

2) *If* **Desempenho_Medio** *is* **Alto** and **Resultado** *is* **Reprovado**
Then **Atividades_Entregues** *is* **1,00**
Rule's probability: **1,000**
The rule exists in **2** *records.*

3) *If* **Atividades_Entregues** *is* **3,00** and **Desempenho_Medio** *is* **Alto**
Then **Resultado** *is* **Aprovado**
Rule's probability: **1,000**
The rule exists in **2** *records.*

4) *If* **Atividades_Entregues** *is* **3,00** and **Desempenho_Medio** *is* **Medio**
Then **Resultado** *is* **Aprovado**
Rule's probability: **1,000**
The rule exists in **2** *records.*

5) *If* **Atividades_Entregues** *is* **1,00**
Then **Resultado** *is* **Reprovado**
Rule's probability: **1,000**
The rule exists in **3** *records.*

3.3.2 Regras de Classificação

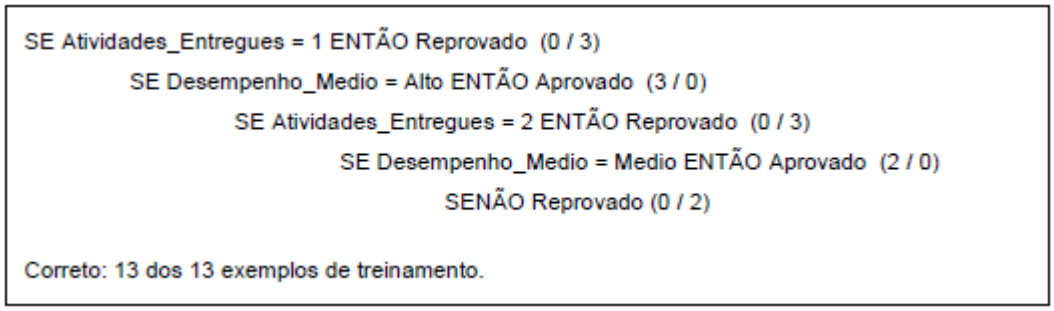
Algoritmos de classificação visam encontrar algum relacionamento entre os atributos e classes pré-estabelecidas gerando uma regra. Posteriormente, com dados de um novo registro, será possível predizer a classe desse novo registro. A qualidade da regra leva em consideração a sua precisão, sua cobertura e seu comprimento (HAN ; KAMBER, 2006 *apud* KAMPFF, 2009). A precisão representa a relação entre o número de casos classificados corretamente e o número total de casos em que as condições apresentadas no antecedente da regra são verdadeiras. A cobertura diz respeito à relação entre o número total de casos classificados

corretamente e o total de casos analisados. O comprimento indica a quantidade de atributos testados no antecedente da regra (KAMPFF, 2009).

O resultado gerado é representado por uma série de estruturas condicionais sucessivas e aninhadas relacionando os atributos e seus valores com as classes. Ao lado de cada regra de classificação, são apresentados entre parênteses, separados por uma barra, os seus indicadores de ocorrência, sendo o primeiro valor indicando quantos dos casos do conjunto de referência são enquadrados na regra estabelecida e satisfazem à classe estabelecida; enquanto que o segundo valor indica o número de casos que não satisfazem ao valor aferido na classe.

De forma complementar ao exemplo utilizado para regras de associação, Kampff (2009) apresenta uma aplicação de algoritmo de classificação sob a mesma base de dados utilizada de estudantes e suas avaliações. A Figura 6 apresenta as regras de classificação resultantes da aplicação desse algoritmo.

Figura 6 - Exemplo de regras de classificação em formato textual



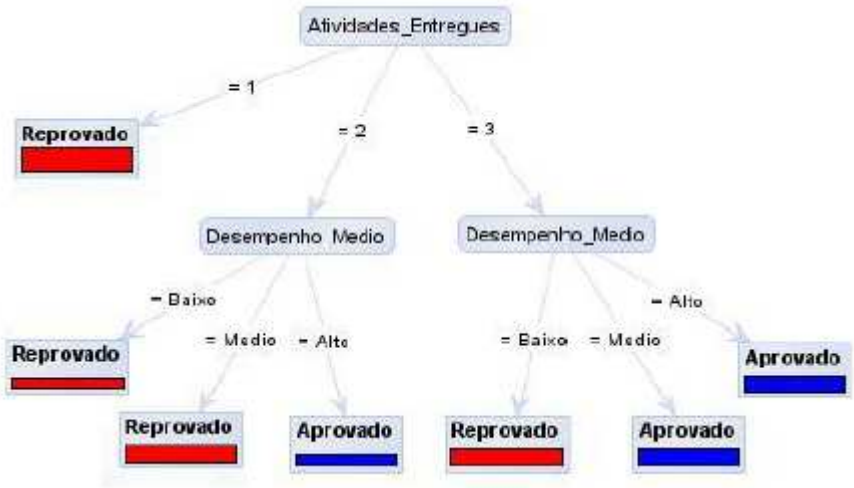
Fonte: Kampff (2009).

Esse exemplo aponta que em três dos treze registros analisados, sempre que um estudante entregou apenas uma atividade, ele foi reprovado. Dos estudantes que entregaram mais de uma atividade, todos os que tiveram desempenho alto foram aprovados e, daqueles que tiveram desempenho médio ou baixo, todos os que entregaram somente duas atividades foram reprovados. Dos estudantes que entregaram três atividades, os que tiveram desempenho médio foram aprovados e os demais foram reprovados. Dessa maneira, um novo registro que for adicionado, com apenas as informações de quantidades de atividades entregues e desempenho, terá o seu atributo de resultado identificado entre aprovado e reprovado.

3.3.3 Árvores de Decisão

Um algoritmo de árvore de decisão constrói, a partir de seus dados de entrada, uma árvore onde cada nó representa um ponto de decisão que o leva a um novo nó ou a uma folha que mostre o resultado previsto. A Figura 7 demonstra graficamente a estrutura de uma árvore de decisão onde os valores dos nós “Atividades_Entregues” e “Desempenho_Médio” determinarão o resultado da decisão tomada, representado pelas folhas “Aprovado” e “Reprovado”.

Figura 7 - Representação gráfica de uma árvore de decisão



Fonte: Kampff (2009).

Destaca-se na classificação por árvores de decisão a boa visualização das características que interferem nas classes (KAMPFF, 2009).

Camilo *et al* (2009) resumem que a técnica de classificação por árvore de decisão funciona como um fluxograma em forma de árvore, onde cada nó (não folha) indica um teste feito sobre um valor (por exemplo, idade > 20). As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe (categoria) a qual o registro pertence. Após ter a árvore de decisão montada, para classificarmos um novo registro, basta seguir o fluxo na árvore (mediante os testes nos nós não-folhas) começando no nó raiz até chegar a uma folha. Pela estrutura que formam, as árvores de decisões podem ser convertidas em regras de classificação.

Kampff (2009) destaca que os algoritmos para construção de árvores de decisão atuam recursivamente durante a fase de aprendizagem, particionando os dados até o nível em que as

folhas representem classes puras ou até que o critério de parada tenha sido atingido com a quantidade estabelecida de casos identificados.

Além da representação gráfica, uma árvore de decisão gerada pode ser representada textualmente, conforme a Figura 8.

Figura 8 - Representação textual de uma árvore de decisão

```
SE Atividades_Entregues = 1 ENTÃO Reprovado
SE Atividades_Entregues = 2 ENTÃO
        SE Desempenho_Medio = Baixo ENTÃO Reprovado
        SE Desempenho_Medio = Medio ENTÃO Reprovado
        SE Desempenho_Medio = Alto ENTÃO Aprovado
SE Atividades_Entregues = 3 ENTÃO
        SE Desempenho_Medio = Baixo ENTÃO Reprovado
        SE Desempenho_Medio = Medio ENTÃO Aprovado
        SE Desempenho_Medio = Alto ENTÃO Aprovado
```

Fonte: Kampff (2009).

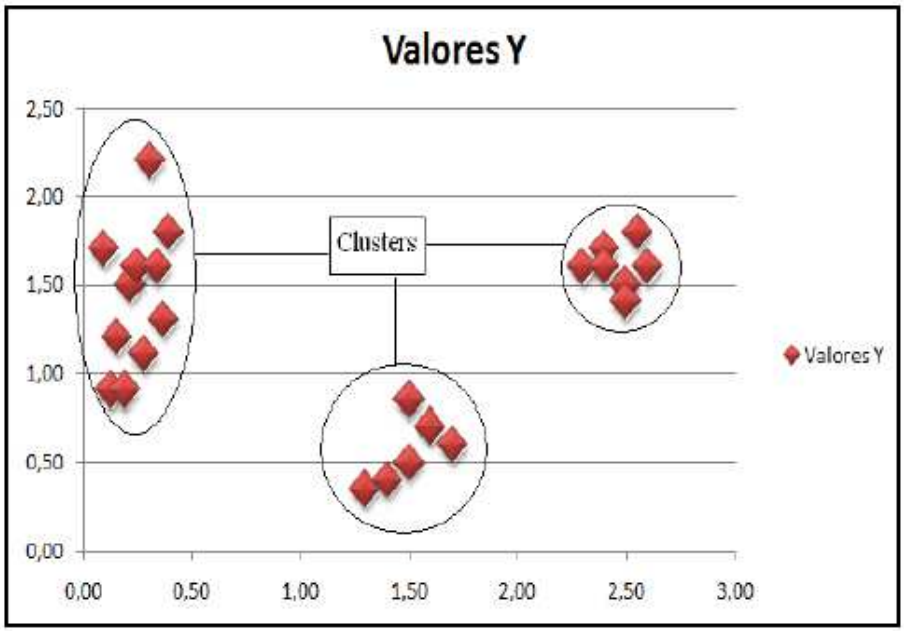
A partir dos exemplos acima, apresentados por Kampff (2009) utilizando a sua base de dados de exemplo, é possível identificar de imediato que o fator mais preponderante na aprovação dos estudantes foi o da quantidade de atividades entregues, posicionado na raiz da árvore.

3.3.4 Agrupamento

As técnicas de agrupamento ou *clustering* consistem em identificar grupos de dados com características semelhantes. Para realizar o agrupamento são estabelecidos critérios para a subdivisão do conjunto principal em subconjuntos. Entre esses critérios destacam-se, segundo Maravalle *et al* (1997 apud KAMPFF, 2009), a homogeneidade e a separação. Chama-se de homogeneidade de um *cluster* o grau de similaridade de seus elementos, e de separação o nível de diferença entre elementos de *clusters* distintos.

O objetivo principal do agrupamento, como pode ser visto na Figura 9, de Pasta (2011), é a partição da base de dados em um número determinado de *clusters*, com instâncias similares.

Figura 9 - Exemplo de visualização de clusters



Fonte: Pasta (2011).

Diferentemente da classificação, no agrupamento nenhuma informação prévia sobre classes é passada ao algoritmo sendo a descoberta dos clusters realizada inteiramente pelo algoritmo, o que pode gerar agrupamentos que, à primeira vista, não fazem sentido a uma análise humana (BERRY ; LINOFF, 1997 *apud* KAMPFF, 2009).

3.4 CONCLUSÃO

No capítulo atual, vimos que a mineração de dados é uma etapa no processo global de KDD que consiste em atividades de pré-processamento, mineração de dados e pós-processamento. Suas técnicas têm sido largamente aplicadas no comércio varejista e em *e-commerce* e começou a ser utilizado em *e-learning*, com resultados promissores (ROMERO *et al*, 2006).

Da mesma forma com que a mineração de dados é utilizada para planejar o estoque de deteminado supermercado em função do comportamento de seus clientes, é possível minerar dados de estudantes para verificar, por exemplo, a relação existente entre aprendizado dos alunos e a metolodoliga pedagógica do curso. Dentro deste contexto, surgiu uma nova área de pesquisa conhecida como “Mineração de Dados Educacionais”, definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais (BAKER *et al*, 2011).

No próximo capítulo, é apresentado o conceito de Mineração de Dados Educacionais, juntamente com uma compilação das principais pesquisas realizadas nessa área.

4 MINERAÇÃO DE DADOS EDUCACIONAIS

Com a proliferação dos cursos à distância no Brasil nos últimos anos, identificou-se uma grande diversidade de modelos pedagógicos suportados por uma gama de ferramentas de apoio pedagógico (ABED, 2011). A Tabela 9 apresenta a distribuição das respostas do Censo ABED 2011, à questão "*No caso dos cursos on-line, o LMS (Learning Management System – Sistema de Gestão Pedagógica e Administrativa da Aprendizagem) usado pela instituição/empresa, em 2010 é um produto [...]*". Observa-se que, enquanto algumas instituições de ensino superior decidiram construir seus próprios Sistemas de Gestão de Aprendizagem (SGA), a maioria tem optado por utilizar sistemas de código aberto e customizá-los às suas necessidades.

Tabela 9 - Distribuição das respostas à questão sobre a natureza dos LMS em instituições de ensino

Resposta	**Frequência**	**%**
Totalmente desenvolvido pela própria instituição	2	6%
Adquirido de uma empresa comercial	0	0%
Locado de uma empresa comercial	7	20%
Gratuito	10	29%
Não informado	16	46%

Fonte: Censo ABED (2011).

Todos os modelos têm em comum o fato de concentrarem suas funcionalidades no suporte automatizado à operacionalização de seus cursos, minimizando os aspectos e necessidades gerenciais.

Dentre os aspectos gerenciais da educação, destaca-se a preocupação das instituições de ensino, sobretudo daquelas que oferecem cursos de Educação a Distância, com os altos índices de evasão dos estudantes. Nos cursos tradicionais, presenciais, a ausência do estudante nas salas de aulas, predominantemente sobre outros fatores, aponta diretamente para uma potencial evasão do mesmo. Na modalidade EAD, este fenômeno da evasão somente é detectado no ato de renovação de matrícula, quando nada ou muito pouco pode ser feito para revertê-lo.

Apesar dos sistemas de gestão de aprendizagem utilizados pelas instituições de ensino produzirem um volume considerável de dados sobre seus cursos e estudantes, relativos à sua participação e operação nos ambientes virtuais, os gestores desses cursos e instituições sofrem com a falta de informações consolidadas que os apoiem em seus processos decisórios. Diante

desse cenário, a utilização de técnicas de mineração pode munir o gestor de cursos de Educação a Distância de informações que o apoiem nas suas decisões.

A área de Mineração de Dados Educacionais (MDE) está orientada ao uso de técnicas de mineração de dados em ambientes educacionais. As técnicas de mineração de dados educacionais são projetadas a partir de diversas fontes de pesquisa, tais como inteligência artificial, psicometria, estatística, visualização da informação e modelagem computacional (BAKER *et al*, 2009).

Alguns trabalhos acadêmicos têm tangenciado o apoio à tomada de decisões em instituições de ensino sendo normalmente orientados a produzir informações ao professor com relação ao processo de aprendizagem de seus estudantes. Entretanto, pouco tem sido estudado sobre a utilização dessas bases de dados e técnicas de mineração no apoio aos aspectos administrativos como, por exemplo, no acompanhamento do fenômeno da evasão de estudantes.

Preliminarmente à definição da abordagem apresentada, uma revisão da literatura foi realizada nesta dissertação, buscando identificar o estado da arte nos trabalhos de investigação aplicados à mineração de dados destinada a detecção e combate à evasão de estudantes em cursos EAD. Dentre os estudos analisados, os trabalhos desenvolvidos por Chen *et al* (1996) e Goebel *et al* (1999) trazem taxonomias sobre mineração de dados sem, no entanto, trazer contribuições a ambientes educacionais.

Um retrato histórico dos trabalhos realizados na área de EAD pode ser encontrado nas taxonomias desenvolvidas nos artigos apresentados na Tabela 10.

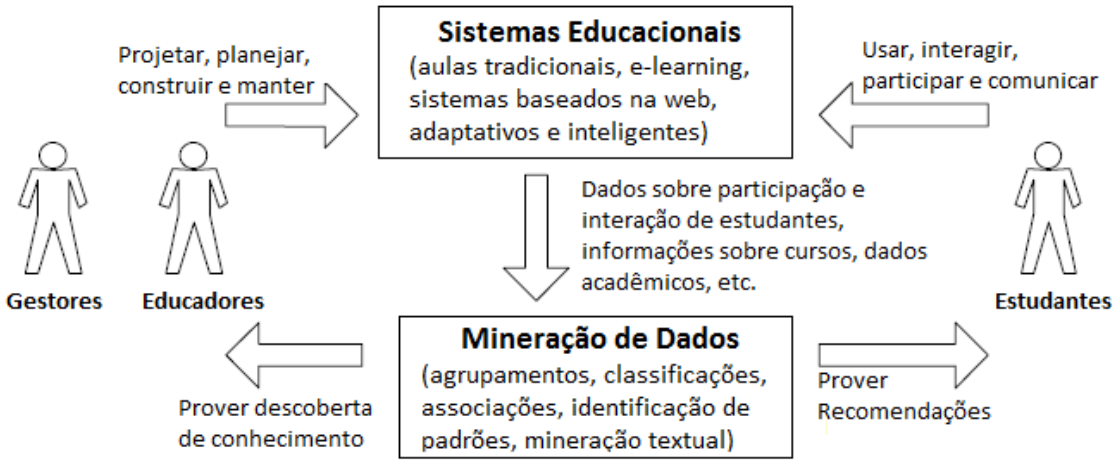
Tabela 10 - Evolução das taxonomias realizadas sobre MDE

Ano	Autores	Referência
1996	CHEN *et al*	Data Mining: an Overview from Database Perspective
1999	GOEBEL *et al*	A Survey of Data Mining and Knowledge Discovery Software Tools
2006	ROMERO *et al*	Educational Data Mining: a Survey from 1995 to 2005
2009	BAKER *et al*	The State of Educational Data Mining in 2009: a Review and Future Visions
2011	BAKER *et al*	Mineração de Dados Educacionais: Oportunidades para o Brasil

Romero *et al* (2006) apresentam um estudo histórico contendo os principais trabalhos sobre mineração educacional de dados no período de 1995 a 2005 onde as principais aplicações na área são classificadas de acordo com o ponto de vista de seu usuário. Nesse

estudo, conforme apresentado na Figura 10, são identificados sistemas sob três pontos de vista diferentes. O primeiro grupo identifica sistemas destinados aos estudantes com o propósito de produzir recomendações de atividades e recursos que auxiliem o processo de aprendizagem. O segundo grupo, sistemas destinados a educadores, produz *feedbacks* sobre a eficácia da estrutura do curso e objetos de aprendizagem, além de classificar os estudantes de acordo com sua necessidade de orientação e reforço. O terceiro e último grupo compreende sistemas destinados à gestão do ensino, buscando parâmetros que orientem a melhoria da eficiência e a maximização dos resultados do negócio.

Figura 10 – O ciclo de aplicações de mineração de dados em sistemas educacionais



Fonte: Romero (2006).

Baker *et al* (2009) estendem o trabalho iniciado por Romero *et al* (2006) destacando o rápido crescimento no volume de pesquisas sobre a mineração de dados educacionais publicadas até o ano de 2009, além da crescente acessibilidade de dados educacionais a um conjunto maior de pesquisadores. Além disso, aponta uma preferência em determinados métodos de MDE, como predição e descoberta com modelos, em contraste com a baixa utilização da mineração de relações.

As pesquisas de caráter mais prático, relacionadas a seguir, apontam uma tendência na utilização de técnicas de mineração de dados em detrimento de pesquisas manuais, sobretudo, devido ao dinâmico e expressivo volume de dados produzidos por meio de ferramentas de gerenciamento de aprendizagem em cursos EAD.

Um estudo de caso foi realizado com o desenvolvimento de um tutorial para utilizações futuras por Romero *et al*, (2007), no qual os autores descrevem de que forma as

diferentes técnicas de mineração de dados aplicadas separadamente ou de forma complementar, podem ser utilizadas para aperfeiçoar cursos online e a aprendizagem de seus estudantes. Além disso, afirmam que a utilização das ferramentas de mineração de dados disponíveis no mercado pressupõe um conhecimento mais técnico, devido à sua complexidade, tornando-se de difícil utilização por educadores e gestores da educação. Dessa forma, é importante o desenvolvimento de aplicações "*user friendly*" e intuitivas, com visualizações dinâmicas de resultados, integradas com os sistemas gerenciadores de aprendizagem, para que os profissionais da educação possam lançar mão de seus recursos nas suas tomadas de decisão.

A utilização da MDE em processos decisórios buscando classificar e agrupar estudantes utilizando técnicas de mineração de dados pode ser encontrada em trabalhos como Pimentel *et al* (2003) e Romero *et al* (2003).

Em Pimentel *et al* (2003), os autores utilizam técnicas de clusterização para a identificação de grupos homogêneos de aprendizes no ensino presencial. Os dados pesquisados foram adquiridos por meio do preenchimento de questionários pelos estudantes, mas o estudo aponta como trabalho futuro a utilização de dados dinamicamente gerados por ferramentas de gerenciamento de aprendizagem.

Em Romero *et al* (2003), uma nova experiência de classificar estudantes utilizando técnicas de mineração de dados é apresentada. No estudo apresentado, diferentes técnicas de mineração de dados para classificação de estudantes foram utilizadas por meio de uma ferramenta de mineração dos dados disponível no *Moodle*, sistema de gerenciamento de aprendizagem de código aberto amplamente utilizado no mercado de cursos à distância. Constatou-se que a utilização de etapas de pré-processamento dos dados, tais como discretização e balanceamento de dados, melhoram consideravelmente o desempenho da maioria dos algoritmos de classificação. Para trabalhos posteriores, Romero *et al* (2003) apontam a mensuração do impacto da qualidade e quantidade dos dados no desempenho dos algoritmos, além da utilização da ferramenta por professores em situações pedagógicas reais, comprovando o seu amplo uso pela comunidade docente.

4.1 MINERAÇÃO DE DADOS EDUCACIONAIS E EVASÃO

O fenômeno da evasão, sobretudo em cursos de educação online, também tem sido alvo de estudos científicos. No trabalho apresentado por Favero *et al* (2006), os autores ressaltam a importância da dialogicidade na construção dos ambientes virtuais de aprendizagem em cursos EAD como fator motivador de seus estudantes, atuando como mitigador da evasão de estudantes. Já Parker (1999) apresenta em que medida fatores como sexo, idade, quantidade de cursos EAD concluídos previamente, poder aquisitivo, e número de horas trabalhadas poderiam prever evasão em cursos de educação à distância.

A mineração de dados surge, então, como uma ferramenta no combate à evasão, tanto de maneira proativa, no planejamento e estruturação de cursos à distância, por meio de seus processos decisórios, quanto de forma preditiva, por meio da monitoração de indicadores e construção de mecanismos de alerta para docentes, discentes e gestores (ROMERO *et al*, 2006).

O trabalho apresentado por Hajra *et al* (2008) sugere a utilização de serviços web, mineração de dados e tecnologia móvel como instrumentos de melhoria da garantia da qualidade na educação. Sua principal contribuição consiste na definição de um modelo combinado de tecnologias voltado à garantia de qualidade na educação, mas que pode ser facilmente adaptado para outros segmentos de mercado. Uma vez que a dimensão “tempo” não foi considerada em seu trabalho, Hajra *et al* (2008) sugerem melhorias no seu modelo com a inclusão deste elemento, assim como por meio de processamentos paralelos de mineração de dados que venham a aperfeiçoar o seu desempenho.

Dekker (2009) busca responder em que nível as técnicas de mineração de dados poderiam ser aplicadas para distinguir, logo a partir do primeiro semestre, quais estudantes têm mais possibilidades de concluir o curso de Engenharia Elétrica. Apesar de apresentar uma análise com cerca de 80% de precisão com base em dados do primeiro semestre do curso, o autor considera o estudo inconclusivo quando busca determinar que dados mais impactam no processo de predição.

Pasta (2011) analisa o perfil de egressos e ingressos de uma instituição de ensino superior utilizando técnicas e ferramentas de mineração de dados. Em suas conclusões, o estudo apresenta uma predominância de estudantes originários de escolas públicas que escolheram a instituição principalmente devido à sua localização e que, após concluírem seus cursos não recomendam a terceiros a realização de um curso na instituição. O autor aponta

como possíveis trabalhos futuros a utilização de outras técnicas de mineração distintas daquelas utilizadas e a disponibilização do recurso, na forma de um aplicativo, para que o próprio gestor o manipule nos seus processos decisórios.

A utilização da mineração de dados como apoio à monitoração de processos em cursos de educação à distância foi utilizada por Campana *et al* (2006). Baseando-se no ambiente virtual de aprendizagem AVAUFES, os autores do trabalho desenvolveram uma camada de agentes propiciando um controle das atividades pelos usuários – professores e tutores de cursos online, disparando alertas e informes de compromissos, atividades, rendimento e cumprimento de prazos pelos estudantes. A mineração de dados foi utilizada para identificar os indicadores e valores que ativam os alertas.

Seguindo a mesma linha de construção de alertas a partir da mineração de dados, Kampff (2009) apresenta uma amostra de 1.564 estudantes que cursaram, no período de cinco anos, a disciplina de Instrumentalização Científica, pertencente ao ciclo de formação básica de uma instituição de ensino superior especializada em cursos à distância com abrangência nacional. O trabalho buscou definir um sistema de alertas dinâmicos, integrado ao SGA desenvolvido e utilizado pela instituição, chamado NetAula, para dar suporte ao corpo docente no acompanhamento do processo de aprendizagem dos seus estudantes, monitorando seus perfis e desempenho notificando os professores de forma que possam atuar preventivamente em iniciativas ou intenções de evasão. O grande diferencial da arquitetura proposta reside no fato de combinar indicadores, com parâmetros configurados pelo professor com outros, pré-estabelecidos com base em mineração de dados, por meio da descoberta de relações e padrões de comportamento de outros estudantes em edições anteriores.

Analisando os trabalhos encontrados, observa-se que as possibilidades de pesquisa na área de mineração de dados educacionais são vastas e abrangentes, podendo compreender desde a análise de bases de dados instantâneos com a extração de informações históricas e situacionais até a geração de alertas dinâmicos disparados pela monitoração de indicadores com o apoio de ferramentas de mineração de dados (BAKER, 2009; BAKER *et al*, 2011; ROMERO *et al*, 2006). Ao extrair o retrato de um momento de uma base de dados educacionais para a aplicação de técnicas e ferramentas de mineração de dados, é possível identificar padrões de comportamento desses dados que poderão auxiliar o gestor no seu processo decisório.

Dimokas *et al* (2008) descrevem a construção de uma solução de *data warehouse* que permite a análise de dados educacionais na Universidade de Thessaloniki, Grécia. Seu modelo organiza na forma de esquemas em estrela, seus fatos e dimensões, priorizando informações

de desempenho dos seus estudantes. Similarmente, Silva *et al* (2002) discutem a utilização de técnicas de *data warehouse* e mineração de dados para auxiliar na avaliação do ensino à distância de alunos matriculados em seus cursos. A maior contribuição, em ambos os casos, traduz-se na representação esquemática das informações relevantes.

Outras possibilidades de pesquisa são detectadas em trabalhos que podem ser evoluídos e desenvolvidos para, uma vez identificados os principais indicadores para o comportamento analisado do público alvo, construir mecanismos de mineração dinâmica dos dados produzidos no dia a dia de maneira que o gestor possa ser alertado sempre que esses indicadores atingirem faixas específicas de valor. Na mineração de dados educacionais, por exemplo, a identificação de indicadores que possam predizer a evasão de estudantes em cursos online, por meio de sua participação nos ambientes virtuais de aprendizagem fornecem importantes subsídios nos processos decisórios dos gestores, sobretudo se seus valores são monitorados através de alertas.

Uma característica comum nos trabalhos analisados consiste na falta de um foco em determinado padrão de comportamento dos atores envolvidos na educação – professores, estudantes ou gestores. Mesmo trabalhos que apontam para um objetivo específico como, por exemplo, para o fenômeno da evasão de estudantes em cursos *online*, terminam por, simplesmente, indicar possibilidades de apoio em processos decisórios, sem apontar impactos nas taxas de retenção de estudantes após a utilização das técnicas de mineração.

Outro possível caminho, apontado pelos estudos de Kampff (2009), Campana *et al* (2006) e Cislaghi (2008), está na utilização da mineração de dados para determinar indicadores diversos – de evasão, de aprendizagem e, com esses indicadores construir mecanismos de alertas dinâmicos para o apoio à tomada de decisões. Ambos os estudos se basearam em sistemas de gerenciamento de aprendizagem de código fechado. Uma possibilidade é reconstruir a experiência desses trabalhos tomando como ponto de partida um SGA de código aberto, como o Moodle, permitindo construir aplicativos gerenciais para seus ambientes virtuais de aprendizagem.

Baker *et al* (2011) apresentaram um quadro de oportunidades da mineração de dados educacionais no Brasil devido ao incentivo governamental à utilização da Educação a Distância no país, alertando para a necessidade de disponibilização dos dados dos estudantes de forma padronizada e estruturada para a comunidade científica brasileira.

Em busca da identificação de uma estrutura padronizada para trabalhos de mineração de dados educacionais, a próxima seção apresenta os trabalhos disponíveis que abordam,

especificamente, a fase de preparação de dados para MDE, com destaque para os modelos e esquemas de dados projetados e as razões para as escolhas de suas entidades e atributos.

4.2 PREPARAÇÃO DE DADOS NA MINERAÇÃO DE DADOS EDUCACIONAIS

Segundo Pyle (1999), a preparação básica dos dados para a mineração compreende a identificação da origem e localização dos dados trabalhados (*data discovery*), a caracterização de quais dados se adequam melhor ao propósito da mineração (*data characterization*), e a montagem de um conjunto de dados para serem minerados (*data set assembly*). Esse conjunto de atividades compõe o ensaio de dados (*data assay*).

Na literatura, especificamente sobre mineração de dados educacionais, é dada grande relevância a propostas de algoritmos de mineração de dados e pouco se justifica a escolha do esquema de dados em que são apoiados estes algoritmos. Esta seção discute os esquemas de dados apresentados para mineração de dados educacionais.

Alguns trabalhos acadêmicos (SILVA *et al*, 2002; FERREIRA JR. *et al*, 2006; SOLODOVŅIKOVA *et al*, 2005; ROMERO *et al*, 2007) têm contribuído com informações decorrentes da aplicação de técnicas de mineração de dados que apoiam a tomada de decisões destinadas ao professor decorrentes do processo de aprendizagem de seus estudantes. A geração de informações que apoiem o processo de decisão administrativo como no acompanhamento do fenômeno da evasão de estudantes em cursos de educação a distância pouco tem sido discutida na literatura mesmo em se tratando de um problema complexo e de impacto para as instituições de ensino.

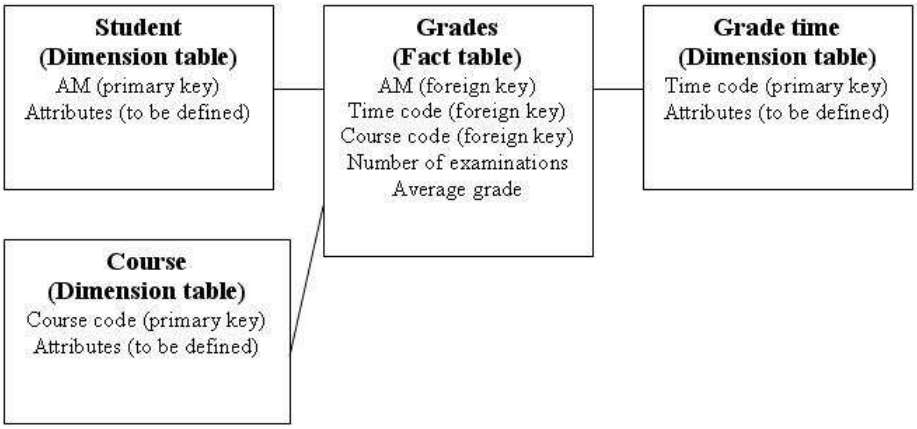
Em pesquisas voltadas para a mineração de dados educacionais (BAKER *et al*, 2011; DIMOKAS *et al*, 2008; SILVA *et al*, 2002; FERREIRA JR. *et al*, 2006; SOLODOVŅIKOVA *et al*, 2005) percebe-se uma preocupação comum em apontar esquemas de dados sem a devida justificativa do uso de seus atributos. São trabalhos que partem de necessidades específicas institucionais, cujo objetivo principal é a avaliação do processo técnico de mineração, porém subestimando o impacto, nos seus resultados, da escolha dos seus grupos de atributos na definição dos fatos e dimensões.

A preocupação em se justificar a escolha de tais atributos decorre do fato de que eles podem influenciar nos resultados obtidos, após o emprego de técnicas de mineração de dados. Na análise de atratividade de um ambiente virtual de aprendizagem, por exemplo, a escolha de um único atributo que indique a quantidade de acessos de cada estudante ao referido

ambiente, sem considerar as atividades realizadas em cada acesso, pode apontar para resultados distorcidos.

Em sua modelagem dimensional, Dimokas *et al* (2008) apresentam suas representações de dados sob dois pontos de vista – a primeira visão orientada aos procedimentos de atribuição de notas aos estudantes e a segunda visão orientada aos processos de conclusão de curso e obtenção de diplomas. Uma tabela de fatos apresenta a quantidade de avaliações e a nota média do estudante no seu curso e período. Tabelas dimensionais são apresentadas para dados dos estudantes, dos cursos e dos períodos cursados, mas o trabalho apenas relaciona os atributos utilizados, sem justificar sua relevância para essas dimensões. A Figura 11 apresenta o esquema em estrela construído para a visão do processo de notas.

Figura 11 - Esquema estrela para a procedure de notas



Fonte: Dimokas (2008).

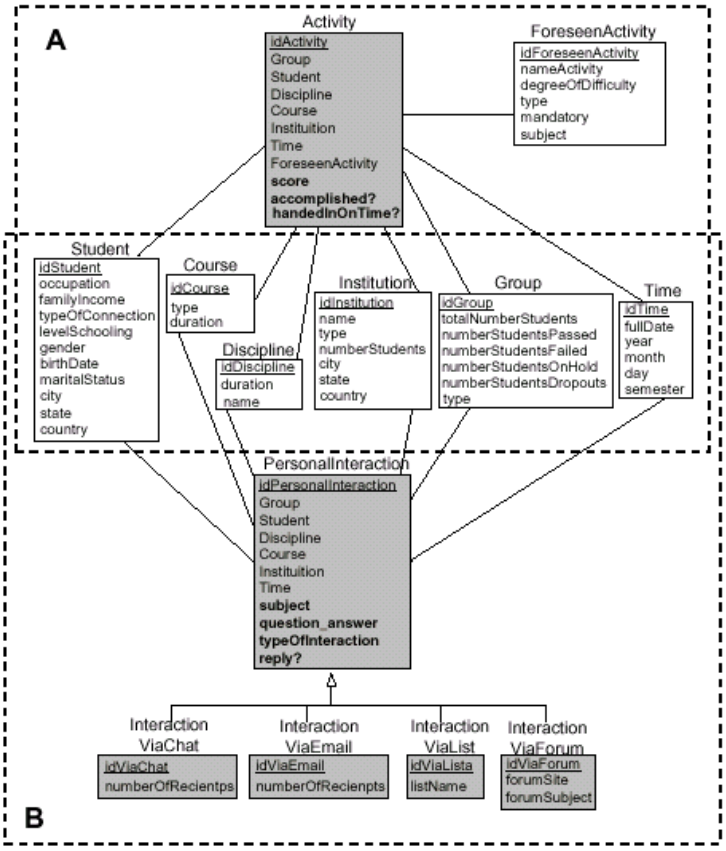
A segunda visão, voltada aos processos de conclusão de cursos, representa o armazenamento de dados de concluintes, relacionando os atributos sem justifica-los.

Utilizando outra abordagem, Silva *et al* (2002) propõem a utilização de recursos de *data warehouse* e mineração de dados na avaliação contínua de cursos *online*. O foco de sua análise concentra-se nos níveis de interação dos estudantes com os objetos de aprendizagem disponibilizados no curso indicando, basicamente, dois grupos de informações para seu diagnóstico – sobre a utilização do material disponibilizado no curso e sobre sua comunicação estabelecida, entre pessoas – colegas, professores, tutores e com os materiais didáticos disponíveis – conteúdos, testes, exercícios.

A Figura 12 apresenta os seus esquemas em estrela, indicando as tabelas-fato sob a perspectiva de interação interpessoal (B) e com as atividades (A), utilizando o mesmo

conjunto de tabelas dimensionais. As caixas em cinza representam tabelas-fato e as demais representam as tabelas de dimensão a partir das quais se deseja armazenar os valores que determinam as medidas das tabelas de fatos. Esse trabalho também não justifica a utilização do seu conjunto de atributos.

Figura 12 - Esquema estrela para as procedures



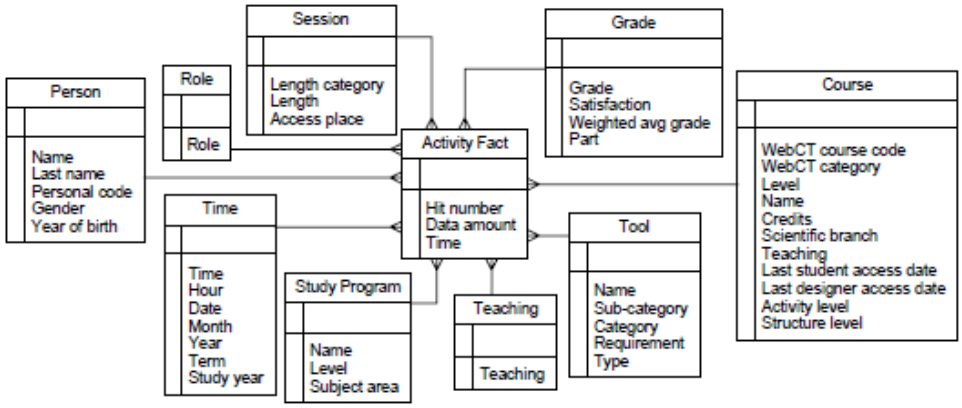
Fonte: Silva (2002).

Apresentando esquemas, sem as justificativas pela escolha de seus atributos, Solodovņikova *et al* (2005) utilizam no seu *data warehouse* três representações em estrela com tabelas de fatos contendo, respectivamente, a estrutura do curso analisado, com suas ferramentas e recursos; o grau de participação dos estudantes nos cursos e as atividades realizadas pelos estudantes durante os cursos. Todas as representações utilizam tabelas dimensionais de tempo, curso e ferramentas, complementadas com outras tabelas de dimensões específicas em cada dimensão.

O diferencial de seu trabalho está no nível de detalhamento de seus atributos. Cada atributo possui uma descrição dos tipos de valores possíveis, orientando o trabalho de extração de dados nos processos de tomada de decisão.

A Figura 13 apresenta a representação do esquema que aborda as atividades dos estudantes no curso.

Figura 13 - Esquema estrela projetado representando as atividades dos estudantes



Fonte: Solodovņikova (2005).

A tabela fato *Activity Fact* contem atributos referentes à utilização dos recursos disponíveis no curso pelo aluno. As demais tabelas (dimensionais) representam:

- *Time* - tabela dimensional padrão para *data warehouse*;
- *Person* – dados pessoais dos envolvidos nos cursos;
- *Role* – representa os papeis que os envolvidos podem assumir em relação aos cursos – estudante, professor, autor.
- *Grade* – contem as notas dos estudantes, além de atributos como grau de satisfação do estudante e parte do curso relacionada com a nota;
- *Course* – dados relacionados às disciplinas, tais como o ano do curso em que é ofertada (*level*), a quantos créditos corresponde (*credits*), entre outros.
- *Teaching* – contem, apenas, um atributo indicando se a disciplina está sendo ofertada no momento ou não.
- *Session* – atributos referentes às conexões com o ambiente virtual, realizadas pelos estudantes - endereço IP do computador do aluno, duração da conexão e classificação da conexão (curta, média, longa, muito longa).
- *Tool* – informações sobre os recursos do LMS, como sua obrigatoriedade (*requirement*) e um indicador se é estática ou dinâmica (*type*).

- *Study Program* – representa programas de estudo disponibilizados pela instituição.

4.3 CONSIDERAÇÕES SOBRE OS MODELOS ANALISADOS

Os estudos realizados, tanto os relacionados com mineração de dado educacionais, quanto aqueles que abordam conceitualmente a educação a distância, apresentam representações de fatores que interferem (ou podem vir a interferir) no processo de aprendizagem ou na educação como negócio. De uma maneira geral, enquanto os modelos conceituais indicam claramente eventos e características que estão diretamente relacionadas com a permanência do estudante nos seus cursos, os modelos e esquemas de dados das pesquisas sobre MDE apresentam atributos distintos e, muitas vezes, divergentes dos modelos conceituais, denotando uma ausência de conexão entre os dois grupos (e entre os próprios trabalhos sobre MDE).

Da parte dos modelos apresentados nos trabalhos sobre mineração de dados educacionais, encontramos uma variedade de proposições de análise, voltadas ao problema específico de cada instituição analisada sem, no entanto, o estabelecimento de vínculos com modelos conceituais. Entidades como Estudante, Curso ou Disciplina estão presentes em praticamente todos os modelos, associadas com outras tabelas, de interesse particular da pesquisa, como *PersonalInteraction* (interações pessoais) do esquema de Silva *et al* (2002), *Grades* (notas) do esquema apresentado por Dimokas *et al* (2008) ou *Tools* (ferramentas) do esquema construído por Solodovnikova *et al* (2005).

Mesmo entidades que constam em mais de um trabalho, possuem poucas correlações na sua composição. Tomemos como exemplos as entidades *Course* (disciplina) e *Student* (estudante). A primeira está presente nos três modelos citados, porém com diferenças significativas. Enquanto o modelo de Solodovnikova *et al* (2005) detalha 11 atributos na entidade, o esquema de Silva *et al* (2002) identifica somente o tipo e a duração como seus atributos e Dimokas *et al* (2008) nem chega a detalhar a entidade, apontando uma chave primária e a expressão *to be defined* (a definir) para seus atributos. Através da Tabela 11 podemos verificar a baixa similaridade da entidade nos três modelos analisados.

Tabela 11 - Atributos da entidade *Course* em três esquemas para MDE

Solodovnikova *et al* (2005)	Silva *et al* (2002)	Dimokas *et al* (2008)
WebCT course code	*idCourse*	CourseCode
WebCT category	*type*	*Attributes to be defined*
Level	*duration*	
Name		
Credits		
Scientific Branch		
Teaching		
Last student access date		
Last designer access date		
Activity level		
Structure level		

Mesmo uma entidade comum para Mineração de Dados Educacionais, como Estudante, não possui uma abordagem consensual nos esquemas analisados. Silva *et al* (2002) apresenta um conjunto de 11 atributos na entidade, aparecendo entre eles, gênero (*gender*) e data de nascimento (*birthDate*). O esquema de Solodovnikova *et al* (2005), por sua vez, apresenta o estudante como um papel (*role*) da entidade Pessoa (*Person*) que possui apenas 5 atributos, incluindo gênero (*Gender*) e ano de nascimento (*Year of Birth*) ao invés de data de nascimento. A ausência de justificativas evidentes nos trabalhos para optar por ano de nascimento ou data de nascimento em um ou em outro modelo, assim como a escolha de um ou outro conjunto de entidades e atributos não implica em ausência de sentido, mas dificulta a definição de padrões e comparações entre as abordagens.

4.4 CONCLUSÃO

Em mais de dez anos de estudos, desde que foi introduzido o conceito da mineração de dados educacionais, vários pesquisas têm sido realizadas com diferentes abordagens e propósitos.

Construções de *data warehouses* educacionais, aplicações de técnicas de mineração de dados em bases educacionais com comparações de desempenho de seus algoritmos, definições de indicadores de evasão a partir de análises de resultados de técnicas de mineração de dados, estudos de adaptabilidade de algoritmos, aplicabilidade de técnicas de mineração na aferição do aprendizado são algumas das linhas de pesquisa recorrentes e,

mesmo os referidos trabalhos de aplicabilidade encontrados procuram respostas para situações-problema de instituições de ensino específicas.

A partir dessa constatação, o presente trabalho procura ocupar uma lacuna identificada, propondo um modelo de dados que favoreça pesquisas com técnicas de mineração de dados em situações-problema transversais a diversas instituições de ensino. Busca-se propor um DER (digrama de entidades e relacionamentos) comum para mineração de dados educacionais que permita a sua utilização por pesquisadores em suas investigações, independentemente das instituições pesquisadas.

No próximo capítulo, será apresentada uma proposta de modelo de dados para uso em pesquisas de mineração de dados educacionais, que justifica a utilização de suas entidades e atributos e busca prover uma maior homogeneidade em futuros trabalhos, além de proporcionar uma convergência com os principais modelos conceituais discutidos.

5 MODELO PARA MINERAÇÃO DE DADOS EDUCACIONAIS

O presente trabalho tem por objetivo a proposição de um modelo de dados para a utilização em pesquisas de mineração de dados educacionais. A partir dos fatores apontados pelos modelos conceituais estudados (capítulo 2) e das necessidades para a aplicação em técnicas de mineração de dados (capítulo 3), foram definidas entidades e atributos que proporcionem um maior conjunto de pesquisas de mineração de dados educacionais, organizados no esquema e o modelo de dados propostos nas seções a seguir.

5.1 DEFININDO UM ESQUEMA DE DADOS

A contextualização dos atributos utilizados nos estudos de mineração de dados se faz necessária na medida em que expõe a estratégia abordada na busca pela solução do problema pesquisado. O estudo atual aponta uma baixa preocupação, nas pesquisas analisadas, com a contextualização dos dados trabalhados, o que resulta em pesquisas específicas para seus cursos e instituições de ensino, de baixa adaptabilidade para outros ambientes e cenários do ponto de vista da gestão de negócios.

Para construir e propor um conjunto de atributos genéricos e úteis em aplicações de mineração de dados educacionais, devidamente justificados, tomou-se como ponto de partida o modelo conceitual revisado de Tinto (2006), apresentado na Figura 9 no capítulo anterior, uma vez que o mesmo já foi extensamente discutido na literatura sobre gestão da educação, segundo McCubbin (2003).

O modelo define os principais pontos de observação considerados importantes em estudos sobre evasão de estudantes – atributos do estudante anteriores ao seu ingresso, informações sobre seu comprometimento em graduar-se e seu esforço de estudo, a sua experiência com a instituição, tanto no aspecto acadêmico quanto na sua integração social e, finalmente, seus resultados na aprendizagem. São considerados atributos relacionados aos estudantes, como indivíduos, tais como:

- Sexo, idade, características e limitações físicas;
- Dados relativos à localização geográfica de sua residência;
- Dados sobre seu histórico familiar, como estado civil, profissão e formação dos pais, renda familiar;

- Dados sobre seu histórico escolar e profissional, como o seu grau de escolaridade, sua origem escolar (se proveniente de instituição pública ou privada), eventuais graduações anteriores, situação laboral.

O comprometimento do estudante com o objetivo de se graduar inicia no seu processo de ingresso na instituição, acompanhando-o durante todo o seu curso. Podemos identificar como atributos desse grupo o tipo de ingresso no curso (se veio através de uma transferência, como portador de diploma ou através de concurso), o resultado obtido (caso tenha ingressado por concurso), a sua prontidão em matricular-se, uma vez aprovado. Durante o curso, sua frequência de participação nas disciplinas cursadas e atividades, sua prontidão em pagar as mensalidades e seu grau de adimplência são exemplos de atributos que registram o seu esforço no curso.

Do ponto de vista da experiência institucional, podemos identificar características do curso escolhido tais como:

- A sua natureza (bacharelado, licenciatura, graduação tecnológica), a área da ciência na qual o curso está inserido, a modalidade do curso;
- Sua duração, sua periodicidade de integralização, a quantidade de disciplinas cursadas simultaneamente;
- Características físicas das instalações frequentadas, tais como a sua localização geográfica, quantidade e qualidade dos recursos disponíveis (ar condicionado, cadeiras, projetores, acesso a *internet*, laboratórios);
- Características do estudante no curso, como:
    - Seu vínculo financeiro (se pagante ou bolsista), seu grau de adimplência, sua prontidão em quitar os débitos;
    - Sua frequência nos mecanismos disponibilizados, sejam eles em sala de aula ou em ambientes virtuais, seu nível de participação nas atividades e recursos de interação;
    - Participação em eventos extracurriculares;
    - Seus resultados acadêmicos, como quantidade de disciplinas pendentes (reprovações) e disciplinas concluídas, quantidade de períodos (semestres, trimestres) concluídos.

A definição de um modelo de dados comum para mineração de dados na área educacional deve pressupor as suas possíveis utilizações em diferentes situações-problema e instituições. Um modelo, para ser genérico, deve conter os elementos mínimos necessários para ser utilizado em pesquisas e trabalhos que busquem investigar questões, não apenas acadêmicas, mas também de caráter administrativas dos estudantes e instituições analisados.

O esquema de dados proposto estratifica os atributos dos estudantes de acordo com sua natureza no problema abordado, ao tempo em que são apontadas as principais dimensões de qualidade que os dados armazenados nesses atributos devem possuir para proporcionar os melhores resultados nos trabalhos de mineração realizados.

A seguir, são apresentados os atributos dos estudantes segmentados por sua natureza:

- Atributos relativos à forma de ingresso na instituição, tais como tipo de ingresso e nota obtida ao ingressar;
- Atributos socioeconômicos, como sexo, estado civil, faixa etária, nível escolar, renda familiar;
- Atributos financeiros, que indicam a relação e a situação financeira dos estudantes com a sua instituição de ensino;
- Atributos acadêmicos, que revelam a atuação e o desempenho dos estudantes nos seus cursos;

Atributos Relativos ao Ingresso no Curso - Pouco considerados nas pesquisas realizadas para mineração de dados educacionais, os dados sobre o ingresso do estudante no curso podem revelar informações importantes em processos de tomada de decisão de gestores acadêmicos ou administrativos de instituições de ensino superior. Muitas vezes, por exemplo, dificuldades de desempenho acadêmico de estudantes podem ter origem no grau de facilitação que tiveram ao ingressaram nos seus cursos, como um baixo nível de dificuldade nas provas de vestibular ou uma matrícula sem prova ou teste de aptidão.

A definição dos atributos relativos ao ingresso dos estudantes no curso/instituição para o esquema proposto, apresentados na Tabela 12, toma por base as informações a respeito do tipo de ingresso na instituição. Dentre as possibilidades existentes tem-se: prova de vestibular, prova de ENEM (exame nacional do ensino médio), matrícula especial para portadores de diploma superior, transferência de outro curso na mesma instituição (interna) ou de outra instituição (externa). Além disso, deve-se registrar o momento cronológico do ingresso do estudante para as análises.

Em algumas das possibilidades de estudo que utilizem o modelo proposto, a informação se o curso no qual está matriculado o aluno foi a sua opção principal pode ter relevância.

Para estudantes que tenham ingressado através de um concurso (prova), deve-se buscar o registro de seu desempenho nessa prova. Para uma aferição homogênea, propõe-se um registro percentual da nota obtida sobre a nota máxima possível na avaliação.

Ainda na classe de atributos relativos ao ingresso, a predisposição do estudante em efetivar a sua matrícula pode apontar a sua motivação em frequentá-lo. Para aferir informações a esse respeito, o esquema propõe a utilização de um atributo que indique o nível de antecipação do estudante na efetivação de sua matrícula no curso/instituição. Para tanto, deve-se calcular e registrar a quantidade de dias resultante da subtração entre a data final do prazo para matrícula e a data da realização da matrícula – maiores antecipações terão valores mais altos enquanto que eventuais matrículas após o prazo (em casos em que a instituição permita) apresentarão valores negativos.

Tabela 12 - Atributos de Ingresso no Curso

Atributo	Descrição
Data de Ingresso	Identificação temporal relativa ao ingresso do estudante no curso/instituição.
Tipo de Ingresso	Opção entre: Prova, ENEM, Matrícula Especial, Transferência Externa ou Transferência Interna.
Nota Obtida	Valores de 0 a 100. Para Ingresso por Prova ou ENEM.
Antecipação de Matrícula	Quantidade de dias de antecipação entre o dia da matrícula e o prazo final.
Opção do Curso	Indicação se o curso matriculado foi a primeira opção do estudante na inscrição.

Dados relativos a ingresso podem ser úteis na definição de campanhas publicitárias voltadas à captação de estudantes, estabelecendo segmentações do público-alvo, na oferta de novos cursos ou produtos educacionais com características que venham a atender a esse cliente em potencial, além de fornecer informações bastante relevantes sobre predisposições de estudantes em permanecer no curso em que se matriculou.

Atributos Socioeconômicos - Os dados de natureza socioeconômica estão entre os mais frequentes em análise de perfis nas mais diversas áreas. No contexto educacional, podem contribuir para inúmeras inferências nos processos decisórios. Informações sobre Sexo, Idade e Estado Civil, por exemplo, do estudante podem estabelecer perfis comportamentais para ações segmentadas. O esquema proposto oferece esses atributos, assumindo valores específicos dentre opções pré-estabelecidas – Sexo, masculino ou feminino; Estado Civil, solteiro, casado, separado ou viúvo.

São propostos, também, atributos que indiquem o grau de escolaridade do estudante (2º grau completo, curso técnico, graduado ou pós-graduado), assim como a sua procedência escolar, indicando se veio de uma instituição pública ou privada.

No que se refere à situação econômica do estudante são propostos dois atributos. O primeiro, indicando a renda referencial da família do estudante, em valores discretos de faixas de salários mínimos e outro, indicando a situação laboral do estudante, informando se o mesmo está desempregado, empregado, se é autônomo ou empresário.

Ainda no grupo de atributos socioeconômicos, são definidos no esquema proposto indicações de município e bairro residencial, comercial e da localização das instalações que frequenta, sejam num *campus* ou polo de apoio presencial para cursos EAD. Todos esses atributos devem ter seus dados tratados de maneira que seus valores sejam informados a partir de listas de valores pré-estabelecidos, evitando formas de preenchimento distintas para a mesma informação. A Tabela 13 apresenta a lista de atributos socioeconômicos do esquema proposto.

Tabela 13 - Atributos Socioeconômicos

Atributo	Descrição
Sexo	Opção entre: Masculino ou Feminino
Estado Civil	Opção entre: Solteiro, Casado, Viúvo, Separado.
Data de Nascimento	Dia, mês e ano do nascimento.
Escolaridade	Opção entre: 2o grau, técnico, graduado, pós-graduado.
Instituição de Origem	Opção entre: Pública ou Privada
Renda Familiar	Em quantidade de salários mínimos
Situação Laboral	Opção entre: nunca trabalhou, empregado, desempregado, autônomo, empresário.
Cidade Residencial Bairro Residencial	Opções de dados textuais a partir de listas pré-estabelecidas. Atributos sobre a localização da residência do estudante.
Cidade Comercial Bairro Comercial	Opções de dados textuais a partir de listas pré-estabelecidas. Atributos sobre a localização do trabalho do estudante.
Cidade Estudo Bairro Estudo	Opções de dados textuais a partir de listas pré-estabelecidas. Atributos sobre a localização das instalações da instituição de ensino frequentadas pelo estudante.

Atributos Financeiros - Mais atraentes para gestores de cursos em instituições de ensino, por seu caráter eminentemente administrativo, valores de natureza financeira pouco têm sido utilizados nas pesquisas estudadas, cuja natureza destina-se mais para questões acadêmicas.

O esquema proposto, conforme relacionado na Tabela 14, apresenta um conjunto de quatro atributos de natureza financeira com capacidade de produzir informações necessárias a

processos de tomada de decisão de gestores em instituições de ensino. Inicialmente propõe-se o armazenamento da indicação do tipo de vínculo financeiro entre o estudante e a instituição, dentre as opções “pagante” ou “bolsista”, propondo, ainda uma subdivisão deste último entre bolsa PROUNI, FIES ou bolsa da própria instituição.

Um possível indicador do comportamento do estudante, de natureza financeira, está no seu hábito de pagamento. Sugere-se o cálculo e armazenamento do tempo médio, em dias, de antecipação no pagamento das mensalidades dos estudantes. Definimos como valor de antecipação à subtração entre a data de vencimento da mensalidade e a data do seu pagamento. Dessa forma, pagamentos antecipados terão valores positivos e pagamentos atrasados, valores negativos. Calcula-se e armazena-se a média histórica desse valor para cada estudante analisado.

Atributos para armazenamento do nível de endividamento dos estudantes também são sugeridos no esquema – um primeiro quantitativo, que calcule e armazene a quantidade de mensalidades vencidas e não pagas de cada estudante analisado no momento atual e um segundo de natureza mais qualitativa indicando, do montante do valor monetário faturado a cada estudante, qual o percentual pendente, em débito.

Tabela 14 - Atributos Financeiros

Atributo	Descrição
Vínculo Financeiro	Opção entre: pagante, bolsista PROUNI, bolsista FIES, bolsista interno.
Antecipação média de Mensalidade.	Quantidade média de dias de antecipação entre o dia de pagamento e o vencimento da mensalidade.
Pendências Financeiras	Quantidade de mensalidades vencidas e não pagas.
Índice de Débito	Total Pago sobre Total Devido.

Análises financeiras, assim como análises socioeconômicas, podem trazer à tona características que definem perfis de estudantes em um curso ou em uma instituição de ensino. Se analisados em conjunto com outras dimensões de atributos, valores financeiros podem determinar relações de “causa e efeito” entre atributos que, a princípio, não aparentam estar relacionados mostrando, por exemplo, a relação entre o nível de endividamento de estudantes com seu desempenho acadêmico ou, ainda, o que influencia mais frequentemente na decisão de estudantes em evadir de seus cursos, se seu desempenho acadêmico ou sua dificuldade em cumprir com suas obrigações financeiras.

Atributos Acadêmicos - Os dados de natureza acadêmica estão presentes na maioria dos estudos realizados em trabalhos de mineração de dados educacionais uma vez que trazem,

normalmente, informações diretas sobre o aprendizado dos estudantes. Desempenho e participação em atividades acadêmicas são os principais elementos nessas análises.

O esquema proposto traz, além desses, sugestões de atributos que podem acrescentar novas análises ao viés acadêmico. Assim como no grupo de atributos relativos ao ingresso no curso, o grupo de atributos acadêmicos demanda o registro de um identificador temporal do momento em que os dados estão sendo registrados de forma a permitir separar informações em momentos distintos.

A utilização de um atributo que indique a situação acadêmica do estudante, entre matriculado ou evadido (cancelado, trancado ou em abandono), é proposta no esquema em questão permitindo inferir sobre estudantes ativos ou inativo.

A caracterização do curso também pode fazer emergir relações importantes nos dados analisados. O nosso modelo propõe as classificações de tipo entre bacharelado, licenciatura e tecnológico; de modalidade, entre presencial ou à distância – EAD; e de escola, entre ciências exatas/tecnológicas, biológicas/de saúde e humanas/sociais.

Três indicadores quantitativos são propostos para o momento do recorte dos dados dos estudantes analisados – um para a quantidade de disciplinas cursadas no período, outro para a quantidade de disciplinas pendentes com alguma reprovação prévia e um terceiro que permita armazenar a quantidade de períodos já concluídos, independente se o curso é integralizado por ano, semestre ou trimestre.

Em relação aos indicadores de participação e desempenho, são propostos atributos que traduzam de maneira linear e independente das disciplinas cursadas, o comportamento do grupo analisado. Propõe-se o registro dos valores limítrofes de participação e desempenho com o maior e o menor percentual de frequência dentre as disciplinas cursadas, assim como a menor e a maior nota global obtida entre essas disciplinas. A relação de atributos acadêmicos incluídos no esquema proposto pode ser observada na Tabela 15.

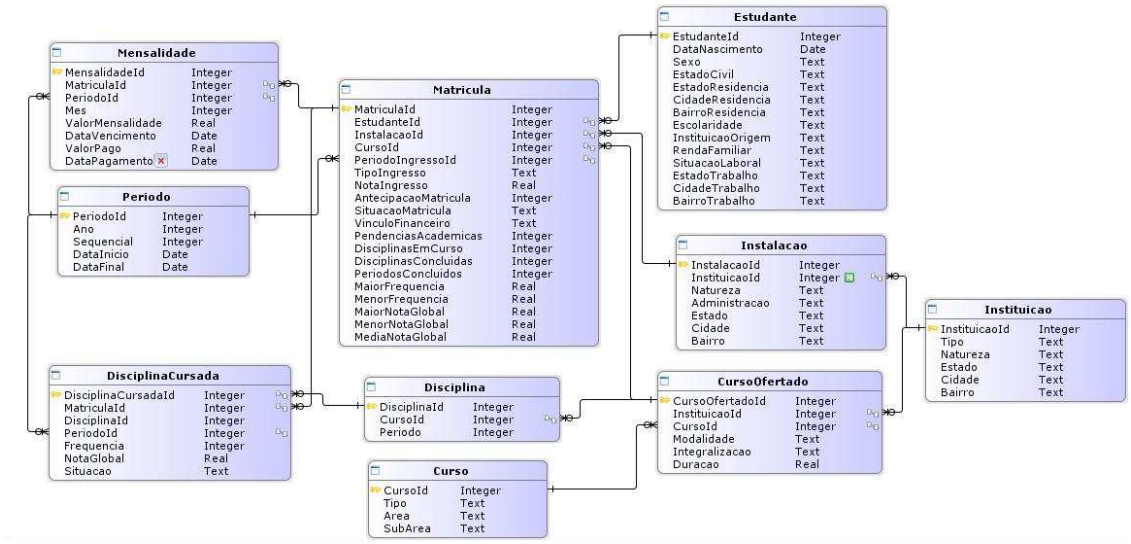
Tabela 15 - Atributos Acadêmicos

Atributo	Descrição
Data Situação	Identificação temporal relativa ao momento da consulta dos dados analisados.
Situação Acadêmica	Opção entre: Matriculado, Trancado, Cancelado, em Abandono.
Tipo do Curso	Opção entre Bacharelado, Licenciatura e Tecnológico.
Modalidade do Curso	Opção entre EAD e Presencial
Escola do Curso	Opção entre Exatas e Tecnológicas, Biológicas e Saúde, Humanas e Sociais;
Pendências Acadêmicas	Quantidade de disciplinas pendentes (reprovações).
Disciplinas Cursadas	Quantidade de disciplinas cursadas no período analisado.
Períodos Concluídos	Quantidade de períodos (semestres, trimestres) concluídos.
Maior Frequência	Maior percentual de frequência entre as disciplinas cursadas
Menor Frequência	Menor percentual de frequência entre as disciplinas cursadas.
Maior Nota Global	Maior nota global relativa dentre as disciplinas cursadas
Menor Nota Global	Menor nota global relativa dentre as disciplinas cursadas.
Nota Média Global	Média das notas globais alcançadas nas disciplinas cursadas.

5.2 UM MODELO DE DADOS PARA MINERAÇÃO DE DADOS EDUCACIONAIS

A partir das considerações feitas, foi desenvolvido um modelo de dados para a construção de *data sets* voltados à Mineração de Dados Educacionais, representado pelo Diagrama de Entidade e Relacionamento (DER) da Figura 14.

Figura 14 - Modelo de Dados proposto para estudos com Mineração de Dados Educacionais.



No modelo de dados proposto, são utilizadas nove entidades-base que representam os principais elementos e atributos apontados nos modelos conceituais discutidos. Com a sua utilização é possível montar uma base de dados homogênea que permita a realização de

consultas e a aplicação de técnicas de mineração de dados, de maneira independente dos softwares utilizados pelas instituições de ensino analisadas e do seu nível de integração.

Instituição – Representa os atributos diretamente relacionados a cada Instituição de Ensino Superior (IES) analisada (Tabela 16).

Os atributos Tipo e Natureza permitem categorizar as instituições de ensino analisadas de acordo, respectivamente, com suas prerrogativas acadêmicas e sua natureza jurídica. De acordo com o Decreto 5.773/06 do Ministério da Educação, uma instituição de ensino superior, para funcionar, deve ser credenciada como faculdade, centro universitário ou universidade, cada qual com diferentes condições de autonomia e operação. Da mesma forma, uma instituição pode ser classificada de acordo com sua natureza jurídica administrativa como pública - gratuita e mantida pelo poder público (Federal, Estadual ou Municipal); ou privada, seja com fins lucrativos ou não.

Igualmente importantes, para efeito dos objetivos do modelo, são os atributos que indicam a localização geográfica da instituição de ensino representadas na entidade pelos campos Estado (Unidade Federativa), Cidade e Bairro. Em um modelo normalizado, um único indicador dentre esses atributos seria suficiente para a localização da instituição uma vez que a relação entre os elementos já traria consigo a informação completa – um bairro é, naturalmente, parte de uma cidade que, por sua vez, está localizada em um Estado.

Tabela 16 - Atributos da Entidade Instituição.

	Atributo	Tipo	Observações
PK	IdInstituicao	Inteiro	Auto numerado. Chave Primária da tabela.
	Tipo	Texto	Classificação da instituição quanto ao seu porte - “faculdade”, “centro universitário” ou “universidade”.
	Natureza	Texto	Natureza Jurídica da instituição - “pública” ou “privada”.
	Estado	Texto	Estado onde se localiza a instituição de ensino.
	Cidade	Texto	Cidade onde se localiza a instituição de ensino.
	Bairro	Texto	Bairro onde se localiza a instituição de ensino.

Instalação – Representa os atributos inerentes a cada instalação física das Instituições de Ensino Superior analisadas, conforme apresentado na Tabela 17. Uma única Instituição de Ensino pode operar em diversos prédios ou instalações.

Para efeito de classificação, uma instalação física de uma instituição de ensino pode ser integrante de seu *campi* ou uma instalação externa como um polo à parte. Pode, também, ser própria, administrada pela própria instituição ou terceirizada, administrada por outra pessoa jurídica que mantem contrato com a instituição de ensino com esse objetivo.

Similarmente à entidade Instituição, devemos registrar a localização física de cada uma de suas instalações, entendendo que, na maioria das vezes, são para esses endereços que os estudantes se dirigem para a realização de suas atividades presenciais, o que pode influenciar, de alguma maneira, características analisadas como seu desempenho no curso ou mesmo a sua permanência.

Tabela 17 - Atributos da Entidade Instalação.

	Atributo	Tipo	Observações
PK	IdInstalacao	Inteiro	Auto numerado. Chave Primária da tabela.
FK	IdInstituicao	Inteiro	Chave estrangeira que relaciona cada Instalação com a sua Instituição de ensino.
	Natureza	Texto	Classificação da instalação quanto à sua natureza - “campus” ou “polo”.
	Administracao	Texto	Classificação da instalação quanto à sua gestão - “própria” ou “terceirizada”.
	Estado	Texto	Estado onde se localiza a instalação da IES.
	Cidade	Texto	Cidade onde se localiza a instalação da IES.
	Bairro	Texto	Bairro onde se localiza a instalação da IES.

Curso – Representa os atributos inerentes aos diversos cursos existentes nas diversas instituições de ensino.

Os Cursos de Graduação atualmente ofertados no Brasil são categorizados em bacharelados, que proporcionam a formação necessária para o exercício das profissões; licenciaturas, que habilitam para o exercício da docência em educação básica (até o ensino médio); e graduação tecnológica, cursos de nível superior, de menor duração de caráter tecnológico e específico.

Utilizando a classificação da CAPES (Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), inserimos a identificação de área e subárea do curso, uma vez que existe uma grande diversidade de nomenclaturas de cursos oferecidos, dificultando a sua comparação. A Tabela de Áreas de Conhecimento da CAPES dispõe de oito grandes áreas: Ciências Exatas e da Terra, Ciências Biológicas, Engenharias, Ciências da Saúde, Ciências Agrárias, Ciências Sociais Aplicadas, Ciências Humanas e Linguística, Letras e Artes, cada qual com suas divisões.

A Tabela 18 apresenta a relação dos atributos da entidade curso, com suas informações complementares.

Tabela 18 - Atributos da Entidade Curso.

	Atributo	Tipo	Observações
PK	IdCurso	Inteiro	Auto numerado. Chave Primária da tabela.
	Tipo	Texto	Classificação do curso quanto à sua natureza – "Bacharelado", "Licenciatura" ou "Tecnológico".
	Area	Texto	Classificação do Curso quanto a sua Área de Conhecimento - "Exatas", "Biológicas", "da Saúde", "Agrárias", "Sociais", "Humanas", além de "Engenharias" e "Linguística, Letras e Artes".
	Subarea	Texto	Classificação segundo a CAPES.

CursoOfertado – Representa os atributos inerentes aos diversos cursos oferecidos por cada uma das instituições de ensino.

Os diversos cursos existentes podem ser ofertados de maneira distinta em cada instituição de ensino superior. A entidade CursoOfertado visa identificar o formato desses cursos nas instituições analisadas (vide tabela 19). Um mesmo modelo pode ser utilizado por vários cursos em diversas instituições.

Um curso pode ser ofertado na modalidade presencial (tradicional) ou através de educação a distância – EAD. Pode ter suas disciplinas integralizadas com uma frequência anual, semestral ou trimestral, além de possuírem durações distintas. Um curso de Pedagogia, por exemplo, pode ser ofertado na modalidade presencial, integralizado em seis semestres ou, à distância, integralizado em 14 trimestres, na mesma instituição ou em instituições distintas.

Tabela 19 - Atributos da Entidade CursoOfertado.

	Atributo	Tipo	Observações
PK	IdCursoOfertado	Inteiro	Auto numerado. Chave Primária da tabela.
FK	IdInstituicao	Inteiro	Chave estrangeira que relaciona cada Curso Ofertado com a sua Instituição de ensino.
FK	IdCurso	Inteiro	Chave estrangeira que relaciona cada Curso Ofertado com um dos cursos cadastrados.
	Modalidade	Texto	Classificação da instalação quanto à sua modalidade - "Presencial" ou "EAD".
	Integralizacao	Texto	Classificação do Curso quanto a sua frequência de integralização – "Anual", "Semestral" ou "Trimestral".
	Duracao	Inteiro	Quantidade mínima de "períodos" do curso. A identificação do tipo de período está relacionada com a sua frequência de integralização.

Estudante – Representa os atributos que caracterizam cada Estudante como indivíduo (Tabela 20).

O modelo propõe o registro dos dados pessoais dos estudantes analisados, identificando sua data de nascimento, para cálculo da sua idade, seu sexo e estado civil. São

inseridos, também, atributos para registro das localizações de seu domicílio (Estado, Cidade e Bairro residenciais) e de seu trabalho, caso existam (Estado, Cidade e Bairro profissionais). São propostos os registros da sua renda familiar, em quantidades de salários mínimos, como referência homogênea e da sua situação laboral, indicando se o mesmo está economicamente ativo. Nesse caso, são sugeridas as opções "*nunca trabalhou*", "*desempregado*", "*empregado*" assalariado, profissional "*autônomo*" ou "*empresário*".

O histórico escolar do estudante é registrado através dos atributos instituição de origem, que informa se o mesmo é proveniente de uma instituição pública ou privada; e escolaridade, indicando o nível de formação do estudante.

Tabela 20 - Atributos da Entidade Estudante.

	Atributo	Tipo	Observações
PK	IdEstudante	Inteiro	Auto numerado. Chave Primária da tabela.
	DataNascimento	Data	Data de Nascimento do Estudante, para aferição de sua idade.
	Sexo	Texto	"Masculino" ou "Feminino"
	EstadoCivil	Texto	"Solteiro", "Casado", "Viúvo", "Separado".
	EstadoResidencia	Texto	Estado onde habita o estudante.
	CidadeResidencia	Texto	Cidade onde habita o estudante.
	BairroResidencia	Texto	Bairro onde habita o estudante.
	Escolaridade	Texto	Nível de escolaridade do estudante - "2º grau completo", "técnico", "graduado", "pós-graduado".
	InstituicaoOrigem	Texto	Natureza Jurídica da instituição de ensino de onde o estudante concluiu seu ensino médio - "Pública" ou "Privada".
	RendaFamiliar	Real	Quantidade de salários mínimos equivalente à renda familiar do estudante.
	SituacaoLaboral	Texto	Classificação do estudante quanto à sua relação com o trabalho - "nunca trabalhou", "empregado", "desempregado", "autônomo", "empresário".
	EstadoTrabalho	Texto	Estado onde o estudante trabalha.
	CidadeTrabalho	Texto	Cidade onde o estudante trabalha.
	BairroTrabalho	Texto	Bairro onde o estudante trabalha.

Periodo – A entidade *Periodo*, apresentada na tabela 21, surge com a necessidade de pontuar temporalmente os eventos acadêmicos ou administrativos acontecidos.

Com uma diversidade de formas de integralização dos diversos cursos e instituições de ensino, optou-se por registrar, para cada período letivo, o ano em que ocorre e um sequencial numérico que o posicione no ano. Dessa forma, caso se trate de um curso ofertado semestralmente, uma representação com os valores 2013 para ano e 1 para o sequencial, indicarão o primeiro semestre de 2013 e, caso o curso seja trimestral, o primeiro trimestre do

mesmo ano e assim sucessivamente. De forma complementar, são identificadas as datas de início e final do período letivo.

Tabela 21 - Atributos da Entidade Período

	Atributo	Tipo	Observações
PK	IdPeriodo	Inteiro	Auto numerado. Chave Primária da tabela.
	Ano	Inteiro	Ano em que ocorre o período.
	Sequencial	Inteiro	Sequencial que indica em que período do ano o evento ocorre.
	DataInicio	Data	Data em que se iniciou o período em questão.
	DataFim	Data	Data em que se encerrou o período em questão.

Matricula – Representa os atributos que caracterizam a relação do estudante com o curso no qual está matriculado, compreendendo os dados relativos ao seu ingresso no curso, além de dados consolidados sobre sua vida acadêmica e financeira, conforme listagem da tabela 22.

Uma matrícula está associada a um estudante, num curso, em determinado período letivo. Além disso, a matrícula associa o estudante/curso à instalação física da instituição de ensino que frequenta.

Os dados relativos ao ingresso do estudante no curso são representados nessa entidade, com a indicação do período letivo em que o estudante ingressou no curso (IdPeriodoIngresso); a forma de ingresso (TipoIngresso), entre "*Prova*" de vestibular, Prova do Exame Nacional do Ensino Médio - "*ENEM*", "*Matrícula Especial*" para portadores de diploma de curso superior, "*Transferência Externa*" de outra instituição ou "*Transferência Interna*" de outro curso, na mesma instituição de ensino. Para os estudantes que ingressaram através de Prova (Vestibular ou ENEM), propõe-se armazenar os dados relativos à sua nota na avaliação realizada, homogeneizando seus valores através de um percentual referente à nota obtida em relação à pontuação máxima possível para a avaliação. Também é considerado um atributo "de Ingresso" o tempo, em dias, de antecipação da matrícula em relação ao prazo máximo para matrículas no período está representado na entidade.

A condição atual do estudante no curso pode ser identificada através do atributo Situação da Matrícula, que indica se o mesmo está "*matriculado*" regularmente, se evadiu sem aviso ("*em abandono*"), temporariamente ("*trancado*") ou definitivamente "*cancelado*"). O seu vínculo financeiro com a instituição pode ser muito útil em análises comportamentais do estudante, pois entende-se que um estudante "*pagante*" tem um comprometimento diferenciado de bolsistas, sejam esses de programas governamentais, como o PROUNI

(Programa Universidade para Todos) e o FIES (Fundo de Financiamento ao Estudante do Ensino Superior), sejam de programas institucionais próprios.

Principal componente para estudos com MDE, a entidade Matricula agrega atributos derivados de outras entidades, facilitando a sua utilização em *data warehouses* e outros *data sets*:

- Pendências Acadêmicas – visa armazenar, de forma consolidada, a quantidade de disciplinas nas quais o estudante não obteve aprovação no curso.
- Disciplinas em Curso – armazena a quantidade de disciplinas cursadas por cada estudante no período da análise.
- Disciplinas Concluídas – registra, de forma complementar às anteriores, quantas disciplinas já foram cursadas e concluídas com êxito por cada estudante.
- Períodos Concluídos – Similarmente ao atributo DisciplinasConcluidas, registra o progresso do estudante através da quantidade de períodos (semestres, trimestres, etc.) já concluídos.
- Maior Frequência – Maior frequência percentual de participação dentre as disciplinas cursadas.
- Menor Frequência – Menor frequência percentual de participação dentre as disciplinas cursadas.
- Maior Nota Global – Maior nota global dentre as disciplinas cursadas.
- Menor Nota Global – Menor nota global dentre as disciplinas cursadas.
- Nota Global Média – Média das notas globais alcançadas nas disciplinas cursadas.

Tabela 22 - Atributos da Entidade Matricula

	Atributo	Tipo	Observações
PK	IdMatricula	Inteiro	Auto numerado. Chave Primária da tabela.
FK	IdEstudante	Inteiro	Chave estrangeira que relaciona cada Matrícula com um dos Estudantes.
FK	IdCurso	Inteiro	Chave estrangeira que relaciona cada Matrícula com um dos cursos ofertados.
FK	IdInstalacao	Inteiro	Chave estrangeira que relaciona cada Matrícula com a instalação onde o curso é ofertado.
FK	IdPeriodoIngresso	Inteiro	Chave estrangeira que relaciona cada Matrícula com o período em que o estudante iniciou seu curso.
	TipoIngresso	Texto	Classificação da Matrícula quanto ao seu ingresso no curso – “Prova”, “ENEM”, “Matrícula Especial”, “Transferência Externa” ou “Transferência Interna”.
	NotaIngresso	Inteiro	Valores de 0 a 100, representando o percentual de acerto do estudante no concurso. Para Ingresso por Prova ou ENEM.
	AntecipacaoMatricula	Inteiro	Diferença, em dias, entre o dia da matrícula do estudante e o dia de encerramento do prazo para matrículas.
	SituacaoMatricula	Texto	Classificação da Matrícula quanto ao seu status – “Matriculado”, “Trancado”, “Cancelado”, “em Abandono”.
	VinculoFinanceiro	Texto	Classificação da Matricula quanto à relação financeira do estudante com a instituição de ensino – “Pagante”, “Bolsista PROUNI”, “Bolsista FIES”, “Bolsista Outros”.
	PendenciasAcademicas	Inteiro	Consolidação da quantidade de disciplinas reprovadas pelo estudante no curso.
	DisciplinasEmCurso	Inteiro	Consolidação da quantidade de disciplinas cursadas pelo estudante no curso.
	DisciplinasConcluidas	Inteiro	Consolidação da quantidade de disciplinas já concluídas pelo estudante no curso.
	PeriodosConcluidos	Inteiro	Consolidação da quantidade de períodos (semestres, trimestres) concluídos pelo estudante no curso.
	MaiorFrequencia	Real	Consolidação do maior percentual de frequência entre as disciplinas cursadas.
	MenorFrequencia	Real	Consolidação do menor percentual de frequência entre as disciplinas cursadas.
	MaiorNotaGlobal	Real	Consolidação da maior nota global relativa dentre as disciplinas cursadas.
	MenorNotaGlobal	Real	Consolidação da menor nota global relativa dentre as disciplinas cursadas.
	NotaMediaGlobal	Real	Consolidação da média das notas globais alcançadas nas disciplinas cursadas.

Mensalidade – Representa os atributos que caracterizam os eventos financeiros mensais dos estudantes nos cursos (Tabela 23).

Cada registro de mensalidade ou parcela está associado a uma matrícula em um período letivo. O conjunto de atributos da entidade é formado pelo mês de referencia da mensalidade, pelas datas de vencimento e pagamento, assim como pelo valor da mensalidade (cobrado) e pelo valor efetivamente pago.

Tabela 23 - Atributos da Entidade Mensalidade.

	Atributo	Tipo	Observações
PK	IdMensalidade	Inteiro	Auto numerado. Chave Primária da tabela.
FK	IdMatricula	Inteiro	Chave estrangeira que relaciona cada Mensalidade com uma Matrícula.
FK	IdPeriodo	Inteiro	Chave estrangeira que relaciona cada Mensalidade com um Período Letivo.
	Mês	Inteiro	Mês de referencia da parcela.
	DataVencimento	Data	Data de vencimento da parcela.
	ValorMensalidade	Real	Valor cobrado.
	DataPagamento	Data	Data do pagamento da parcela.
	ValorPagamento	Real	Valor efetivamente pago.

Disciplina – Para efeito do modelo proposto, cujo objetivo tem o estudante como foco principal, optou-se por tratar as disciplinas somente como um repositório para os atributos que registram a integração acadêmica dos estudantes. Não foram considerados atributos sobre a natureza ou o conteúdo das disciplinas.

A função da entidade é registrar a organização das disciplinas na matriz, ao longo de sua integralização. Além da associação da disciplina com o curso do qual faz parte, é incluído o atributo que indica em que período do curso a disciplina ocorre em sua matriz. Os atributos da entidade Disciplina são apresentados na Tabela 24.

Tabela 24 - Atributos da Entidade Disciplina.

	Atributo	Tipo	Observações
PK	IdDisciplina	Inteiro	Auto numerado. Chave Primária da tabela.
FK	IdCurso	Inteiro	Chave estrangeira que relaciona cada Disciplina com um Curso.
	Periodo	Inteiro	Indicador numérico do período (semestre, trimestre) em que a disciplina é ofertada no curso.

DisciplinaCursada – Entidade que relaciona as disciplinas dos cursos aos estudantes e suas matrículas. Seus atributos e características estão apresentados na tabela 25.

Cada registro de “disciplina cursada” está associado a uma matrícula, a uma disciplina em um período letivo. Um estudante pode cursar a mesma disciplina mais de uma vez, em períodos letivos, uma vez que não tenha obtido aprovação. Da mesma maneira, a cada período letivo, um estudante geralmente é relacionado a várias disciplinas.

Em cada disciplina cursada, são registrados dados da frequência de participação do estudante, sua nota global, em valores percentuais sobre a nota máxima possível e a situação do estudante matriculado na disciplina no período – a disciplina, no período, pode já ter sido

"*concluída*" com sucesso, pode ter sido "*dispensada*" por já ter sido cursada em outra instituição, pode ter sido "*reprovada*" quando o estudante não obtiver sucesso em sua aprovação e pode estar "*em curso*" no momento da extração dos dados.

Tabela 25 - Atributos da Entidade DisciplinaCursada.

	Atributo	Tipo	Observações
PK	IdDisciplinaCursada	Inteiro	Auto numerado. Chave Primária da tabela.
FK	IdMatricula	Inteiro	Chave estrangeira que relaciona cada Disciplina Cursada com um Estudante.
FK	IdDisciplina	Inteiro	Chave estrangeira que relaciona cada Disciplina Cursada com uma Disciplina do Curso.
FK	IdPeriodo	Inteiro	Chave estrangeira que relaciona cada Disciplina Cursada com um Período Letivo do Curso.
	Frequencia	Real	Frequência de participação do Estudante na Disciplina.
	NotaGlobal	Real	Nota Global do Estudante na Disciplina, no Período.
	Situação	Texto	Indicador da situação do estudante com a Disciplina – "Em Curso", "Concluída", "Dispensada", "Reprovada".

5.3 CONCLUSÃO

Trabalhos de mineração de dados educacionais carecem de uma estrutura padronizada e homogênea para pesquisadores e gestores, que permitam análises comparativas entre trabalhos distintos e, sobretudo, uma maior convergência nas suas abordagens. O modelo de dados proposto no trabalho atual visa iniciar o preenchimento da lacuna identificada por Baker *et al* (2011) na padronização e estruturação de dados educacionais para a comunidade científica brasileira.

O modelo proposto busca estabelecer conceitos comuns na análise de situações problema educacionais, além de possuir uma estrutura que permite maior produtividade na seleção de dados para mineração, uma vez que seu conjunto de entidades e atributos, devidamente justificados, facilita a identificação dos fatores mais relevantes nas diversas situações problemas analisadas.

O próximo capítulo apresentará a utilização do modelo de dados proposto em cenários reais, avaliando a sua aplicabilidade.

6 ESTUDO DE CASO – APLICAÇÃO DO MODELO PROPOSTO

6.1 INTRODUÇÃO

A proposição de um modelo de dados para a utilização em mineração de dados educacionais necessita ter sua aplicabilidade posta à prova em situações reais. O estudo de caso descrito neste capítulo parte de um cenário real em uma instituição de ensino superior, e procura demonstrar, não apenas a aplicabilidade do modelo proposto, como também as vantagens em utilizá-lo em detrimento à opção de partir sem uma estrutura específica para estudos de mineração de dados educacionais.

No decorrer deste capítulo são apresentados o planejamento, execução e as conclusões do estudo de caso.

6.2 PLANEJAMENTO

A seção atual apresenta o planejamento realizado para a execução do estudo de caso organizado para avaliar o modelo proposto. Composto pelas definições dos objetivos (global, da medição e do estudo) das questões e hipóteses que apontem indícios de aplicabilidade do modelo.

6.2.1 Objetivo Global

Caracterizar a aplicabilidade do modelo proposto em um problema real.

6.2.2 Objetivo da Medição

Aplicando o modelo proposto em um cenário real de dados educacionais, caracterizar:

- Se existe um subconjunto de atributos do modelo que possibilite a identificação de fatores de evasão de estudantes;
- Se a utilização do modelo proposto reduz o esforço de identificação de atributos relevantes à identificação de fatores de evasão de estudantes, em comparação à utilização de outro modelo;

- Se a utilização do modelo proposto reduz o esforço de preparação de dados para a aplicação em técnicas de mineração de dados, em comparação à utilização de outro modelo.

6.2.3 Objetivo do Estudo

Analisar a aplicabilidade do modelo de dados proposto
Com o propósito de caracterizar as suas vantagens
Com respeito a utilização de técnicas de Mineração de Dados Educacionais.
Do ponto de vista do gestor de cursos de graduação na modalidade EAD
No contexto da identificação de fatores que levam à evasão de estudantes.

6.2.4 Questões

Q1: O modelo proposto contempla atributos relevantes e suficientes para a identificação de fatores de evasão?

Métrica: Identificação de fatores de evasão através de técnicas mineração de dados.

Q2: O uso do modelo reduz o esforço de identificação de atributos relevantes à identificação de fatores de evasão de estudantes, em comparação à sua não utilização?

Métrica: Esforço (tempo) necessário para a seleção de atributos relevantes para o problema analisado.

Q3: O uso do modelo reduz o esforço de preparação de dados para a aplicação em técnicas de mineração de dados, em comparação à sua não utilização?

Métrica: Esforço (tempo) necessário para a preparação de dados para a aplicação em técnicas de mineração de dados.

6.2.5 Definição das Hipóteses

Hipótese nula (H0): A utilização do modelo proposto não possui nenhum impacto na aplicação de técnicas de mineração de dados educacionais.

$I_{Análise}$: Indícios de evasão de estudantes identificados pela análise estatística direta dos dados utilizados.

$I_{Mineração}$: Indícios de evasão de estudantes identificados através da utilização de técnicas de mineração de dados populadas com o modelo proposto.

E_{Direto}: Esforço necessário para a identificação e seleção de atributos relevantes à identificação de fatores associados a evasão de estudantes através da análise direta das tabelas dos bancos de dados da instituição de ensino.

E_{Modelo}: Esforço necessário para a identificação e seleção de atributos relevantes à identificação de fatores associados a evasão de estudantes através da utilização do modelo proposto.

$E_{DataBase}$: Esforço necessário para a preparação dos dados para a mineração a partir das tabelas dos bancos de dados da instituição de ensino.

$E_{DataSet}$: Esforço necessário para a preparação dos dados para a mineração a partir de um *data set* preparado de acordo com o modelo proposto.

H0: $(I_{Análise} \cap I_{Mineração} = \emptyset) \wedge (E_{Direto} = E_{Modelo}) \wedge (E_{DataBase} = E_{DataSet})$

Hipótese alternativa (H1): A utilização de um subconjunto de atributos do modelo proposto em técnicas de mineração de dados propicia a identificação de indícios de evasão de estudantes.

H1: $I_{Análise} \cap I_{Mineração} \neq \emptyset$

Hipótese alternativa (H2): A utilização do modelo de dados proposto reduz o esforço necessário, em horas, para a seleção de atributos relevantes à identificação de fatores que levam a evasão de estudantes.

H2: $E_{Direto} > E_{Modelo}$

Hipótese alternativa (H3): A utilização do modelo de dados proposto reduz o esforço necessário, em horas, para a preparação dos dados para a aplicação de técnicas de mineração de dados.

H3: $E_{DataBase} > E_{DataSet}$

6.2.6 Descrição da Instrumentação

Para este estudo de caso, foram necessários os seguintes instrumentos:

- Ferramentas de consultas nativas dos sistemas gerenciadores de bancos de dados PostgreSQL e SQL Server;
- Ferramentas de escritório para a manipulação de dados – Microsoft Excel e Microsoft Access;
- Ferramenta de modelagem de dados DBWrench;
- WEKA – Ferramenta para mineração de dados.
- Convite aos pesquisadores para participação no Estudo de Caso (ANEXO A).
- Documento de Orientações aos pesquisadores para participação no Estudo de Caso (ANEXO B).
- Formulário para coleta de informações do Estudo de Caso (ANEXO C).

6.2.7 Seleção do Contexto

Este estudo está inserido no seguinte contexto:

- **O processo:** *online*. Todos os processos, em todas as suas etapas serão operados e monitorados pelos participantes do experimento. As medições serão realizadas durante a atuação da equipe.
- **Os participantes:** O próprio pesquisador, autor desse trabalho. Analistas de sistemas convidados.
- **Realidade:** situação real.
- **Generalidade:** específico. A aplicabilidade do modelo de dados proposto será analisada dentro de um problema real de uma instituição de ensino, com o objetivo de identificar indícios de evasão de estudantes em alguns de seus cursos.

6.2.8 Seleção dos Indivíduos

A escolha dos participantes será baseada em princípios não probabilísticos onde os participantes serão determinados por conveniência.

6.2.9 Variáveis

6.2.9.1 Variáveis independentes

1. Indícios de evasão de estudantes identificados pela análise estatística direta dos dados utilizados.
2. Ferramentas de consultas nativas dos sistemas gerenciadores de bancos de dados PostgreSQL e SQL Server.
3. Ferramentas de escritório para a manipulação de dados – Microsoft Excel e Microsoft Access.
4. WEKA – Ferramenta para mineração de dados.

1.1.1.1 VARIÁVEIS DEPENDENTES

1. Indícios de evasão de estudantes identificados através da utilização de técnicas de mineração de dados populadas com o modelo proposto.
2. Esforço necessário para a identificação e seleção de atributos relevantes à identificação de fatores associados a evasão de estudantes através da utilização do modelo proposto.
3. Esforço necessário para a preparação dos dados para a mineração a partir de um *data set* preparado de acordo com o modelo proposto.

6.2.10 Projeto do Estudo de Caso

Para as hipóteses alternativas formuladas, serão utilizados os seguintes padrões de projeto:

H1: $\mathbf{I_{Análise} \cap I_{Mineração} \neq \emptyset}$

- **Fator:** Utilização ou não do modelo proposto.
- **Tratamento:** Análise estatística direta dos dados disponíveis ($\mathbf{I_{Análise}}$), buscando identificar indícios de evasão e análise dos resultados das aplicações de mineração de dados alimentadas a partir do modelo proposto ($\mathbf{I_{Mineração}}$) .
- **Comentários:** Extraídos de um *data set* gerado a partir do modelo de dados proposto, os dados referentes ao conjunto de estudantes evadidos

serão analisados estatisticamente à procura de informações de maior ocorrência. Em seguida, serão realizados dois (2) casos de aplicação de mineração de dados, cujos resultados serão comparados com os resultados da análise direta.

- **Caso 1:** Aplicação de um algoritmo de *clustering* a partir dos dados referentes aos ingressos dos estudantes nos períodos analisados.
- **Caso 2:** Aplicação de um algoritmo de classificação, com a geração de uma árvore de decisão, a partir dos dados referentes aos ingressos dos estudantes nos períodos analisados.

H2: $E_{Direto} > E_{Modelo}$

- **Fator:** Utilização ou não do modelo proposto.
- **Tratamento:** Seleção dos atributos relevantes à identificação de indícios de evasão, utilizando o modelo proposto (**E_{Modelo}**) e diretamente, pelas bases de dados disponíveis (**E_{Direto}**).
- **Comentários:** Serão convidados 06 (seis) analistas de sistemas sem conhecimento prévio dos modelos de dados para realizar, aleatoriamente, os dois tratamentos para o fator esforço.

H3: $E_{DataBase} > E_{DataSet}$

- **Fator:** Utilização ou não do modelo proposto.
- **Tratamento:** Aferição do tempo de preparação dos arquivos de entrada para as operações de mineração de dados dos casos 1 e 2, utilizando o modelo proposto (**$E_{DataSet}$**) e diretamente, pelas bases de dados disponíveis (**$E_{DataBase}$**).
- **Comentários:** O mesmo participante realizará os 04 (quatro) tratamentos para o fator esforço.

6.3 EXECUÇÃO

Nesta seção são apresentadas: (1) a descrição dos processos de identificação e seleção de atributos com a utilização do modelo proposto, assim como sem a sua utilização,

comparando ambos os esforços; (2) a descrição dos processos de preparação dos arquivos de entrada de dados para a mineração, comparando os esforços da realização com o modelo e sem o modelo proposto; (3) a realização de aplicações de mineração de dados para os casos 1 e 2.

6.3.1 Identificação e Seleção dos Atributos para a Mineração

O sucesso de uma aplicação de mineração de dados depende, além da escolha correta do conjunto amostral dos dados a analisar, da correta identificação dos atributos a serem investigados no processo. O conhecimento prévio do problema a ser analisado facilita a seleção do conjunto de atributos mais importante na abordagem. A hipótese alternativa H2 do presente estudo afirma que a utilização do modelo de dados proposto reduz o esforço necessário à identificação e seleção de atributos relevantes à identificação de indícios de evasão de estudantes.

Para a verificação dessa hipótese, foram elaborados dois conjuntos de documentos contendo, em cada um, uma carta de apresentação à experiência, explicações conceituais sobre o tema da evasão, um formulário para registro da experiência e um modelo de dados. O primeiro conjunto continha o modelo de dados proposto, com descrições sobre suas entidades e atributos, enquanto que o segundo conjunto apresentava um diagrama simplificado da base de dados do sistema de processo seletivo da instituição analisada. Uma cópia de cada documento pode ser encontrada anexada ao apêndice dessa dissertação (ANEXO A – Convite aos pesquisadores para participação no Estudo de Caso; ANEXO B – Orientações aos pesquisadores para participação no Estudo de Caso e ANEXO C – Formulário para coleta de informações do Estudo de Caso).

Foram convidados seis analistas de sistemas sem conhecimento prévio dos modelos de dados e, aleatoriamente, um dos conjuntos foi entregue para cada participante, com o objetivo de que de posse, unicamente, do material apresentado, pudessem identificar os atributos mais relevantes ao tema da evasão de estudantes.

6.3.1.1 Caracterização dos participantes

Todos os analistas de sistemas que participaram da atividade possuem experiência em desenvolvimento de software e análise de diagramas de entidade e relacionamento. Trata-se de um grupo entre 25 e 44 anos de idade com pelo menos 1 ano de experiência em

desenvolvimento de sistemas e modelagem de dados. O participante mais experiente possui mais de 25 anos de atividade profissional. Apenas um dos participantes não possuía curso superior completo e apenas um deles nunca teve nenhum contato com Educação a Distância.

6.3.1.2 Análise dos Resultados

Cada participante trabalhou exclusivamente com o material disponibilizado, sem que houvesse qualquer limitação ao tempo da experiência. A documentação fornecida orientava expressamente os seus usuários nas etapas a serem seguidas para um melhor resultado:

1) Leitura das Orientações para a Realização do Trabalho que caracterizam o cenário do problema a ser analisado com a utilização dos modelos de dados;
2) Medição do horário de início do trabalho;
3) Análise do Modelo de Dados fornecido no sentido de identificar os 10 principais atributos que estejam relacionados, direta ou indiretamente, com a evasão de estudantes;
4) Medição do horário de conclusão do trabalho;
5) Preenchimento do Perfil do Pesquisador no formulário anexo, descrevendo brevemente o seu perfil profissional;
6) Preenchimento das informações de Aferição dos Resultados no formulário, descrevendo o esforço necessário e suas impressões sobre o trabalho.

A seção de Aferição dos Resultados do formulário apresentava uma tabela em branco com 10 linhas para o preenchimento dos atributos identificados e, para cada atributo, o nome da entidade da qual faz parte, além de uma breve justificativa da sua escolha.

Quatro participantes analisaram o material com o modelo proposto enquanto que três pesquisadores analisaram apenas o modelo original. Um participante analisou ambos os modelos, totalizando sete análises do conjunto.

O tempo médio das análises do material com o modelo da base de dados original foi de 33 minutos, maior que o tempo para a análise do modelo proposto, de 25 minutos. Os tempos máximo e mínimo das análises também foram menores sobre o material relativo ao modelo proposto, conforme assinalado na tabela 26.

Tabela 26 - Comparativos dos tempos mínimo, médio e máximo das análises dos modelos

	Modelo Proposto	Modelo Original
Tempo mínimo	00:15	00:25
Tempo médio	00:25	00:33
Tempo máximo	00:43	00:48

Todos os participantes alegaram dificuldades na compreensão dos significados das entidades e atributos do modelo original e consequentemente na análise de sua relevância sobre o tema, situação reforçada pelo fato de apenas uma das três análises desse documento ter completado o conjunto de dez atributos solicitados. Chama a atenção também a baixa reincidência dos atributos desse grupo nos trabalhos. Do total de vinte e quatro atributos identificados pelos pesquisadores como relevantes, apenas dois deles aparecem em mais de uma lista.

Os pesquisadores que trabalharam com o modelo de dados proposto identificaram vinte e seis atributos distintos considerados como relevantes na análise de evasão de estudantes. Desse conjunto, dez atributos aparecem em mais de uma lista, sendo que um deles consta em todas as quatro listas geradas.

A experiência realizada indica que mesmo havendo uma variação do tempo decorrido de acordo com a vivência e dedicação de cada pesquisador, a utilização do modelo proposto na análise do tema sugerido reduziu o esforço para a identificação e seleção dos atributos, confirmando a Hipótese Alternativa H2, que afirma que o esforço para a identificação dos atributos sem a utilização do modelo proposto (**E_{Direto}**) é maior do que o esforço com a sua utilização (**E_{Modelo}**).

Uma constatação colateral encontrada aponta que a utilização do modelo proposto resultou numa maior homogeneidade dos atributos encontrados o que, por sua vez, indica mais facilidade nas análises. Além disso, todos os participantes que analisaram o modelo original alegaram dificuldades na identificação pela falta de maiores esclarecimentos.

6.3.2 Preparação de Dados Para a Mineração

Uma vez identificado o conjunto de atributos pertinentes à análise do problema a ser investigado com a mineração, é necessário construir um *data set* contendo os dados de entrada para os algoritmos utilizados. Esses dados, geralmente são provenientes dos bancos de dados

de uma ou mais instituições analisadas podendo inclusive ser oriundos de diferentes plataformas tecnológicas (Oracle, SQL, PostGreSQL, etc.). A etapa de preparação de dados para a mineração consiste, justamente, na construção desse conjunto de dados, devidamente tratado, que irá alimentar o algoritmo de mineração utilizado nas sua análises.

A Hipótese Alternativa H3 do presente estudo afirma que a utilização do modelo de dados proposto reduz o esforço necessário, em horas, para a preparação dos dados para a aplicação de técnicas de mineração de dados. Para a verificação dessa hipótese uma amostra de dados reais será definida, a partir da qual serão realizadas quatro extrações e tratamentos de dados com a geração, em cada uma, de um *data set* para mineração. Duas extrações serão realizadas sem a utilização do modelo de dados proposto, gerando um arquivo que será utilizado como entrada para um algoritmo de classificação e um arquivo para um algoritmo de clusterização. As outras duas extrações, realizadas a partir do modelo de dados proposto, deverá gerar arquivos idênticos aos anteriores. O impacto da utilização do modelo proposto será aferido com base nos tempos de cada processo.

Para os casos estudados, foram utilizados dados de uma única instituição de ensino superior com oferta de cursos na modalidade EAD. A instituição em questão possui um conjunto de aplicações, próprias e de terceiros, para as diferentes finalidades acadêmico-administrativas do processo educacional – processo seletivo, gestão acadêmica dos estudantes, gestão financeira, além do seu Sistema de Gerenciamento de Aprendizagem, que produz dados sobre a participação do estudante nos seus cursos.

Cada aplicação possui seu sistema de gerenciamento de banco de dados específico, apesar de existirem pontos de integração indireta (por processamento batch) entre os mesmos. Além disso, não há um padrão geral para nomenclatura de entidades, atributos e relacionamentos, nem documentações institucionais atualizadas para a orientação de desenvolvedores, uma vez que as documentações existentes são pontuais e específicas, construídas pelos analistas de sistemas responsáveis por cada aplicação. A falta de padronização nos bancos de dados se dá, principalmente, devido à ausência de um administrador de dados dedicado à função e à utilização de aplicações comerciais em conjunto com as aplicações desenvolvidas internamente.

As aplicações institucionais são complementadas com sistemas de menor porte desenvolvidos, muitas vezes, pelas equipes setoriais com a utilização de dados provenientes de mais de uma base de dados como fonte. A partir de necessidades identificadas pelos setores, são geradas, também, consultas que integram dados de diferentes fontes.

6.3.2.1 Análise Estatística dos dados obtidos.

Na realização do estudo de caso, foram analisadas as matrículas de calouros e veteranos de cursos EAD da instituição do primeiro período letivo do ano de 2011 ao último período letivo de 2012 e, sempre que identificado que uma matrícula de um período letivo não constava no período imediatamente seguinte, caracterizava-se uma evasão. Assim, foram identificadas 3.958 situações de evasão divididas nos períodos letivos de acordo com a representação da tabela 27.

Tabela 27 - Distribuição dos evadidos identificados na amostra, por período letivo

Período Letivo	Quantidade
2011.1 – 2011.2	1.127
2011.2 – 2012.1	1.089
2012.1 – 2012.2	1.742

Todos os dados de identificação da amostra foram armazenados em um único repositório (Microsoft Access) e iniciou-se, então, a busca pelos demais dados desses estudantes, seguindo os conjuntos identificados no capítulo *4.1 Definindo um Esquema de Dados*. As tabelas 28 a 31 apontam a taxa de sucesso na obtenção dos dados, tendo como referência os 3.958 registros (100%).

Tabela 28 - Taxas de Sucesso na obtenção dos dados de Ingresso dos estudantes

Atributo	**Taxa de Sucesso**
Tipo de Ingresso	100%
Nota Obtida	100%
Antecipação de Matrícula	68,47%
Opção do Curso	100%

Tabela 29 - Taxas de Sucesso na obtenção dos dados Sócio-Econômicos dos estudantes

Atributo	Taxa de Sucesso
Sexo	100%
Estado Civil	0%
Idade	100%
Escolaridade	0%
Instituição de Origem	35,45%
Renda Familiar	0%
Situação Laboral	0%
Cidade Residencial	0%
Bairro Residencial	0%
Cidade Comercial	0%
Bairro Comercial	0%
Cidade Polo	100%
Bairro Polo	100%

Tabela 30 - Taxas de Sucesso na obtenção dos dados Financeiros dos estudantes

Atributo	Taxa de Sucesso
Vínculo Financeiro	0%
Antecipação Média de Mensalidade.	0%
Pendências Financeiras	0%
Índice de Débito	0%

Tabela 31 - Taxas de Sucesso na obtenção dos dados Acadêmicos dos estudantes

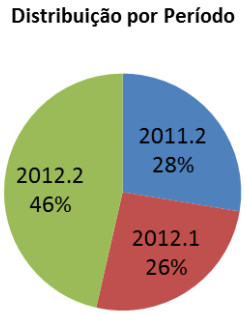
Atributo	Taxa de Sucesso
Tipo do Curso	100%
Modalidade do Curso	100%
Escola do Curso	100%
Pendências Acadêmicas	0%
Disciplinas Cursadas	0%
Períodos Concluídos	100%
Maior Frequência	0%
Menor Frequência	0%
Maior Nota Global	0%
Menor Nota Global	0%
Nota Média Global	0%

Diante da dificuldade na obtenção dos dados e considerando que a maior concentração de evasão se encontra no primeiro período (semestre) dos cursos (segundo censo da ABED – Associação Brasileira de Educação à Distância e segundo a própria base analisada), o

conjunto de dados foi restrito a estudantes de primeiro semestre, com 1.866 registros que atendem a essa condição (aproximadamente 47% do total de evadidos), melhorando as taxas de sucesso.

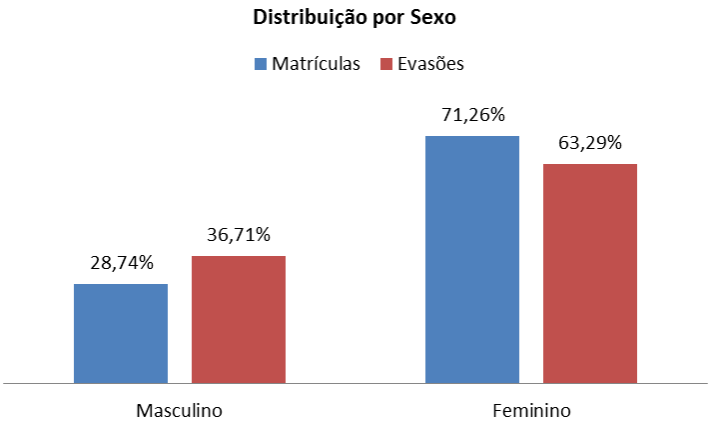
Do ponto de vista dos períodos letivos, conforme a Figura 15, os dados da amostra estão distribuídos de maneira uniforme, em consonância com o conjunto total dos registros. Essa constatação aponta uma baixa relação entre o período letivo do estudante e o seu período de ingresso.

Figura 15 - Distribuição da amostra de evadidos por período de ingresso



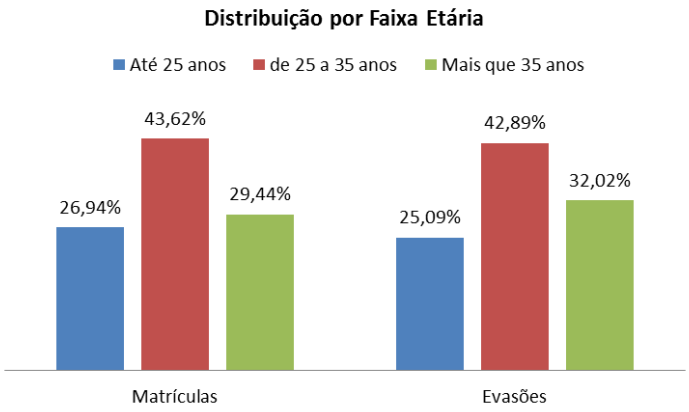
Analisando os aspectos de sexo e idade dos indivíduos da amostra, conforme demonstrado nas Figuras 16 e 17, constata-se também uma baixa relação entre esses atributos e o fenômeno da evasão, uma vez que a maioria de ambos os conjuntos é composta de mulheres (63%) e de pessoas com mais de 25 anos (75%). Encontra-se uma leve tendência a evadir de estudantes do sexo masculino que, no conjunto de matriculados correspondem a cerca de 29% e, no grupo de evadidos tem sua proporção elevada para 37%.

Figura 16 - Comparação das distribuições por sexo



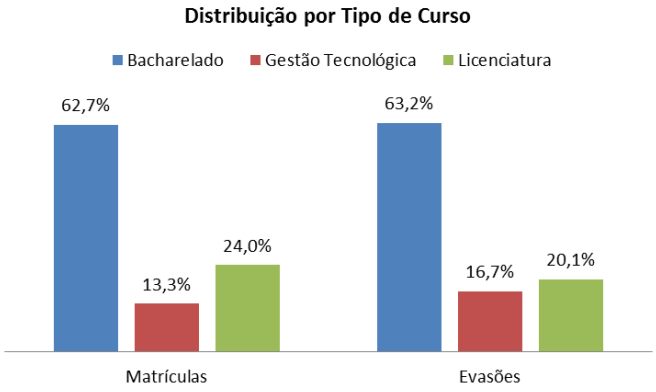
A Figura 17 aponta que seguindo o que acontece em relação a sexo, estudantes com mais de 35 anos tem sua proporção aumentada em cerca de 3 pontos percentuais quando comparamos a amostra de evadidos com a população de matriculados.

Figura 17 - Comparação das distribuições por faixa etária



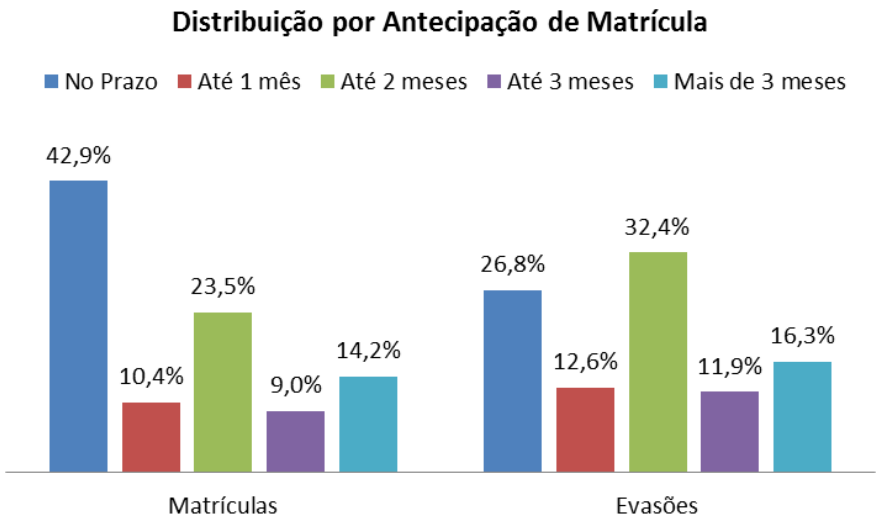
Quando comparamos os dados por tipo de curso, verificamos uma pequena tendência a evadir em estudantes de cursos de Graduação Tecnológica, como pode ser verificado na Figura 18.

Figura 18 - Distribuição da amostra de evadidos por Tipo de Curso



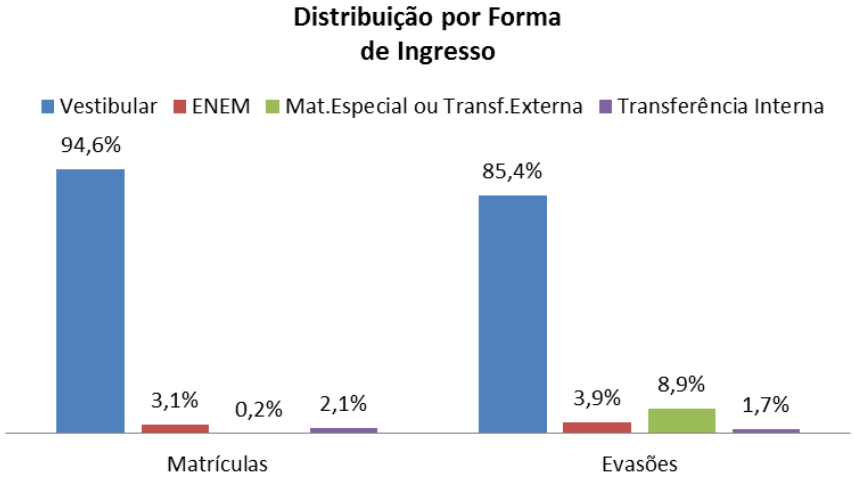
Os dados de antecipação de matrícula, demonstrados na Figura 19, já nos revelam uma maior tendência a evadir de estudantes que realizam suas matrículas com mais de uma semana de antecipação (soma das ocorrências de matrícula antecipadas), uma vez que sua proporção na amostra de evadidos aumenta em quase 16 pontos percentuais (de 57,1% para 73,2%) em relação ao montante de matriculados. Por outro lado, estudantes que realizam suas matrículas no prazo (em até uma semana de antecipação) apresentam uma redução equivalente de participação no grupo de evadidos.

Figura 19 - Distribuição da amostra de evadidos por Antecipação de Matrícula



Pelo aspecto da forma de ingresso, demonstrado na Figura 20, destaca-se uma considerável tendência a evadir em estudantes provenientes de processos de matrícula especial (portadores de diploma superior) e transferência externa, que têm uma participação de 0,19% na população dos matriculados analisados e, na amostra de evadidos, passa a contribuir com quase 10% do total.

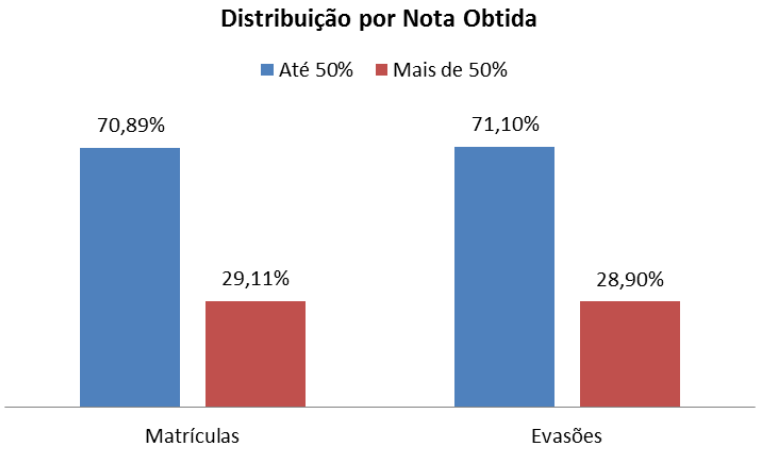
Figura 20 - Distribuição da amostra de evadidos por Forma de Ingresso



Apesar do contingente de estudantes que ingressaram por prova (Vestibular ou ENEM) com nota de até 50% do total representar quase ¾ do total de evadidos que ingressaram por prova, constatamos que ingressar no curso através de prova, com um resultado apenas suficiente para a aprovação não constitui indicativo de evasão, uma vez que

a população total de matriculados analisada possui praticamente a mesma proporção (Figura 21).

Figura 21 - Distribuição da amostra de evadidos que ingressaram através de prova por Nota Obtida



A análise estatística realizada na amostra revela alguns atributos que trazem indícios de evasão tendo como o mais relevante, a antecipação da efetivação da matrícula. Mesmo em atributos onde, pelo estudo realizado, não foram encontradas diferenças significativas entre a amostra e o montante de matriculados, é possível encontrar indícios de evasão, assim como, através da utilização de técnicas de mineração, é possível encontrar comportamentos comuns em conjuntos de atributos.

6.3.2.2 Padrão de entrada para mineração de dados

A verificação da Hipótese Alternativa H3 que afirma que a utilização do modelo proposto reduz o esforço da preparação dos dados para mineração, passa, necessariamente, pela etapa da transformação dos dados para a mineração de dados. Essa etapa consiste na organização desses dados num formato válido como entrada de dados para o algoritmo utilizado.

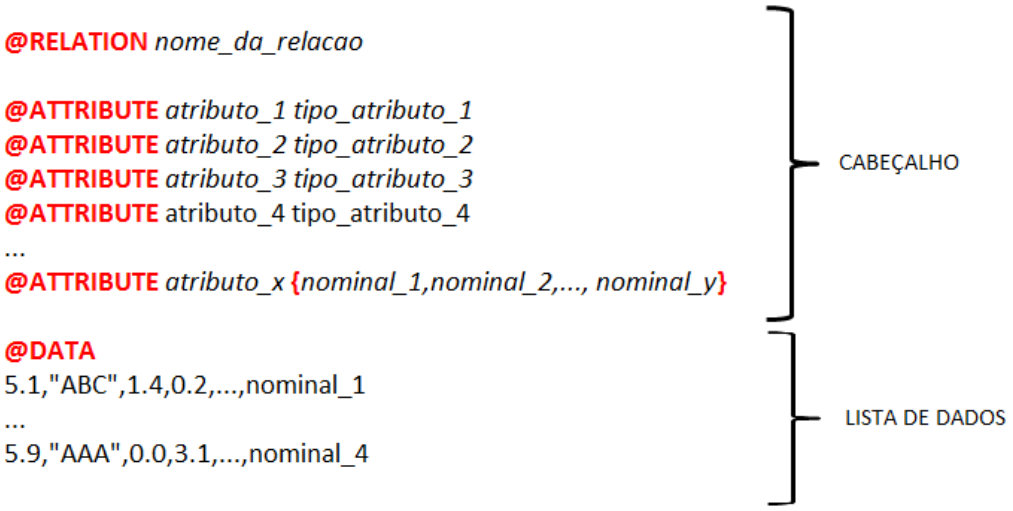
Para a execução das técnicas de mineração de dados nos casos analisados, foi utilizada a ferramenta Weka. Acrônimo para *Waikato Environment for Knowledge Analysis* ou, em português, Ambiente Waikato para Análise de Conhecimento, trata-se de uma ferramenta gratuita e de código aberto, desenvolvida e disponibilizada por pesquisadores da Universidade de Waikato, na Nova Zelândia, utilizada para minerar dados e transformá-los em conhecimento para o apoio em tomadas de decisão. O ambiente disponibiliza recursos para

análise de pré-processamento dos dados, para treinamento dos algoritmos e a sua aplicação em si, com uma grande variação de opções de algoritmos para classificação, agrupamento e associação.

O Weka requer como entrada para seus processos arquivos no formato “Atributo-Relação” ou simplesmente arff (do inglês *Attribute-Relation File Format*). São arquivos texto, obedecendo ao padrão ASCII, que descrevem uma lista de elementos que obedecem a um conjunto de atributos pré-estabelecido.

Cada arquivo arff é composto de duas seções distintas: um cabeçalho e uma lista de dados. O cabeçalho deve conter um nome para a relação em questão e a lista dos atributos utilizados, com seus respectivos tipos - numéricos, textuais (*strings*), datas ou rótulos nominais. A seção de dados apresenta a listagem de instâncias dos registros a serem analisados pelos algoritmos. A Figura 22 apresenta um esquema geral para um arquivo arff.

Figura 22 - Esquema padrão de um arquivo arff



Em vermelho, encontram-se as palavras de uso reservado @RELATION para identificar a relação, @ATTRIBUTE, para identificar cada atributo e @DATA para apontar o início da seção de dados. Atributos do tipo “valores nominais”, necessariamente, devem apresentar na sua declaração o seu conjunto de valores possíveis relacionados entre chaves.

Tarefas de associação e classificação trabalham, predominantemente, com valores nominais e tarefas de *clustering* com atributos numéricos. Dessa forma, decidimos trabalhar com arquivos distintos para cada um dos casos analisados. Em ambos os casos, os dados para os mesmos atributos (Periodo, Curso, TipoIngresso, IngressoNota,

IngressoAntecipaçãoMatrícula, Sexo, Idade, Cidade, BairroEstudo, PendenciasAcademicas, DisciplinasCursadas, PeriodosConcluidos e IndicadorEvasão) são utilizados para as tarefas de mineração, variando a forma com que são utilizados e, consequentemente os tipos dos atributos.

Os atributos para o agrupamento (Caso 1) foram declarados conforme a relação a seguir:

```
@attribute Periodo {20111,20112,20121}
@attribute Curso {LICENCIATURA,BACHARELADO,TECNOLOGICO}
@attribute TipoIngresso {VEST,ENEM,ME,TE,TI,PROUNI}
@attribute IngressoNota real
@attribute IngressoAntecipaçãoMatricula real
@attribute Sexo{1,2}
@attribute Idade real
@attribute Cidade {CAPITAL,INTERIOR}
@attribute BairroEstudo {CENTRO,OUTRO}
@attribute PendenciasAcademicas real
@attribute DisciplinasCursadas real
@attribute PeriodosConcluidos real
@attribute IndicadorEvasao {SIM,NAO}
```

Analogamente, os atributos para a classificação (Caso 2) foram declarados conforme a relação a seguir:

```
@attribute Periodo {20111,20112,20121}
@attribute Curso {LICENCIATURA,BACHARELADO,TECNOLOGICO}
@attribute TipoIngresso {VEST,ENEM,ME,TE,TI,PROUNI}
@attribute IngressoNota {SemNota,<6,>=6}
@attribute IngressoAntecipaçãoMatricula
{SemAntecipacao,AtéUmaSemana,AtéDuasSemanas,AtéUmMês,MaisDeUmMês}
@attribute Sexo{1,2}
@attribute Idade {Até25anos,de25a35anos,Maisque35anos}
@attribute Cidade {CAPITAL,INTERIOR}
@attribute BairroEstudo {CENTRO,OUTRO}
@attribute PendenciasAcademicas {0}
@attribute DisciplinasCursadas {0,1,2,3,4,5,6,7}
@attribute PeriodosConcluidos {0,1,2,3,4,5,6,7,8}
@attribute IndicadorEvasao {SIM,NAO}
```

Para gerar as listas de valores nominais dos atributos IngressoNota, IngressoAntecipacaoMatricula e Idade, foram determinadas faixas de valores indicando os

cortes desejados para os respectivos atributos. Os demais valores nominais foram obtidos através de consultas diretas nas bases de dados relacionando, de forma distinta, os valores existentes.

6.3.2.3 Extração de Dados sem o modelo proposto

A extração de dados para mineração, quando realizada diretamente nos bancos de dados dos sistemas em produção, demanda um conhecimento prévio de suas estruturas e relações ou, pelo menos, uma boa documentação que facilite a sua análise. Conhecer a estrutura dos bancos de dados em detalhes permite ao pesquisador identificar mais rapidamente a sua origem e definir a melhor estratégia para a obtenção dos dados.

Os dados dos atributos relacionados aos estudos de caso possuem origem em bancos de dados distintos na instituição analisada. No sistema acadêmico estão os dados cadastrais dos estudantes, assim como dados sobre o seu seguimento acadêmico no curso, enquanto que dados sobre o ingresso do estudante no curso encontram-se no sistema específico para os processos seletivos. A tabela 32 divide os atributos selecionados para a mineração de dados por sua origem.

Tabela 32 - Atributos para a mineração de dados por origem nos sistemas legados

Banco de Dados Acadêmico	**Banco de Dados de Processos Seletivos**
Periodo	Periodo
Curso	TipoIngresso
Sexo	IngressoNota
Idade	IngressoAntecipacaoMatricula
NomeCidade	
BairroEstudo	
PendenciasAcademicas	
DisciplinasEmCurso	
DisciplinasConcluidas	
PeriodosConcluidos	
IndicadorEvasao	

A dificuldade em reunir diretamente dados provenientes de bancos de dados distintos demanda, de imediato, a criação de um repositório intermediário para o tratamento e preparação dos dados extraídos. Esse repositório deverá ser formado pela união dos atributos das duas fontes de dados, acrescida de um atributo de identificação única do estudante. A tabela 33 apresenta a estrutura necessária para a tabela temporária.

Tabela 33 - Atributos da tabela temporária para a preparação de dados

IdEstudante
Periodo
Curso
Sexo
Idade
Cidade
BairroEstudo
PendenciasAcademicas
DisciplinasEmCurso
DisciplinasConcluidas
PeriodosConcluidos
IndicadorEvasao
TipoIngresso
IngressoNota
IngressoAntecipacaoMatricula

Alguns atributos como Sexo, Curso, Cidade, BairroEstudo e TipoIngresso, serão armazenados nesse banco intermediário da forma como estão organizados originalmente, sem nenhum tratamento.

A seleção direta dos dados de suas bases originais pode ser realizada com a utilização simples de consultas SQL diretamente nos seus bancos de dados, tendo seus resultados armazenados no repositório intermediário. O Anexo D, ao fim do documento, apresenta as consultas utilizadas para extrair esses dados.

A consulta SQL 1 foi utilizada para extrair os dados referentes aos atributos Periodo, Sexo, Curso, Cidade e BairroEstudo da base acadêmica e inserir na tabela temporária. Na consulta, são filtrados os dados de estudantes recém ingressados (`Al.Id_PeriodoLetivo = Al.Id_PeriodoLetivoIngresso`) dos períodos letivos desejados (`Al.Id_PeriodoLetivo IN (1, 2, 3)`). Os valores alfanuméricos referentes ao atributo Sexo (“F” ou “M”) são transformados em numéricos (1 ou 2) através da expressão `IIF(Pes.Sexo ="F",1,2)`. Os valores dos demais atributos são inseridos na tabela temporária com valor nulo até que as demais consultas os preencham.

Alguns dados da base do sistema de processo seletivo podem ser extraídos também através de consultas simples, conforme a consulta SQL 2 que seleciona os dados referentes ao atributo TipoIngresso, restritos aos processos seletivos dos períodos letivos desejados (`WHERE V.Vest_Id IN (1, 2, 3)`).

Os demais atributos necessários, por outro lado, não existem originalmente na forma desejada, necessitando de um tratamento de seus dados para que atendam ao propósito do trabalho:

Idade – Os dados do atributo Idade são calculados através da diferença, em anos, entre a data vigente e a data de nascimento dos estudantes.

IndicadorEvasao – Calculado a partir da informação da situação do estudante no curso – estudantes com matrícula "cancelada", "trancada" ou "em abandono" são considerados evadidos.

A consulta SQL 3 foi utilizada para extrair os dados referentes aos atributos Idade e IndicadorEvasao da base de dados acadêmica. Além do cálculo da Idade dos estudantes (`INT((Now() - P.Dat_Nascimento)/365)`), a consulta restringe os valores sobre a situação acadêmica dos alunos a um indicador que informa se ele evadiu ou não (`IIF(S.Des_Situacao="ATIVO","NÃO","SIM")`).

IngressoNota – Uma vez que avaliações distintas têm ponderações distintas, os valores armazenados nesse atributo devem conter um percentual representando a pontuação atingida sobre a pontuação máxima possível. Dessa maneira, o desempenho de um estudante em uma avaliação que vale 10 pontos pode ser comparado homogeneamente com o desempenho de outro estudante, de ENEM por exemplo, cuja pontuação máxima é de 1.000 pontos.

IngressoAntecipacaoMatricula – Com o objetivo de identificar a pontualidade do estudante mediante o prazo final de matrícula, o atributo é calculado pela subtração, em dias, entre a data final determinada para sua matrícula e a data em que a matrícula foi realizada. Dessa maneira, valores próximos a zero indicarão estudantes que tardaram em efetivar suas matrículas, enquanto que valores maiores apontarão para antecipações na formalização.

Os dados referentes aos atributos IngressoNota e IngressoAntecipacaoMatricula são calculados e extraídos através da consulta SQL 4. A expressão `IIF( C.Cand_TipoProva = "ENEM", C.Cand_Nota/10, C.Cand_Nota / 1000)` garante a homogeneidade nos valores das provas de vestibular e ENEM, que possuem ponderações distintas nos seus valores - com a expressão, garantimos que ambos sejam armazenados em valores proporcionais, na faixa de 0% a 100%. O cálculo da antecipação de matrícula é realizado subtraindo a data de realização da matrícula da data limite prevista para o encerramento das matrículas (`M.Marc_DataLimiteMatricula - C.Cand_DataMatricula`).

PendenciasAcademicas – Calculado pela contagem de disciplinas do histórico do estudante que pertencem, na matriz do curso, a um período anterior ao último período cursado e permanecem sem conclusão.

DisciplinasEmCurso – Calculado pela contagem de disciplinas inconclusas do histórico do estudante que pertencem, na matriz do curso, ao último período cursado.

DisciplinasConcluidas – Calculado pela contagem de disciplinas do histórico do estudante que pertencem, na matriz do curso, a um período anterior ao último período cursado e já foram concluídas.

PeriodosConcluidos – Calculado pela contagem de períodos do histórico do estudante, onde a totalidade de disciplinas esteja com situação concluída.

A geração dos dados calculados para a atualização dos atributos PeriodosConcluidos, DisciplinasEmCurso, DisciplinasConcluidas e PendenciasAcademicas requer a utilização de funções de agregação na sua consulta. A impossibilidade técnica do sistema gerenciador de banco de dados utilizado em realizar alterações de registros (UPDATE) que possuam funções de agregação demanda a criação de uma segunda tabela temporária para armazenar os resultados calculados e, depois, atualizar a tabela desejada. A tabela 34 apresenta a relação de atributos da tabela “temporaria2”, criada para armazenar os dados calculados e apoiar na atualização da tabela temporária.

Tabela 34 - Atributos da tabela temporaria2 para a preparação de dados

IdEstudante
Periodo
PeriodosConcluidos
DisciplinasEmCurso
DisciplinasConcluidas
PendenciasAcademicas

A consulta SQL 5a foi utilizada para extrair os valores dos atributos PeriodosConcluidos, DisciplinasEmCurso, DisciplinasConcluidas e PendenciasAcademicas da base de dados acadêmica, referentes aos estudantes recém ingressados nos períodos letivos desejados e inserir na tabela temporaria2. A consulta contabiliza as informações para cada estudante a partir das suas disciplinas cursadas ou em curso.

Os dados referentes à quantidade de períodos concluídos são calculados pela identificação do semestre anterior ao que cada estudante esteja cursando (`A.SemestreAtual - 1`). Dessa forma, estudantes de 3º semestre têm 2 períodos concluídos (o 1º e o 2º

semestres), os de 2º semestre têm 1 período concluído e os de 1º semestre não concluíram nenhum período.

Os atributos referentes às quantidades de disciplinas em curso e concluídas têm seus valores calculados através da contagem por estudante das disciplinas com status, respectivamente, “Em Curso” (`SUM(IIF(SM.Des_Situacao="Em Curso",1,0))`) e “Concluida” (`SUM(IIF(SM.Des_Situacao="Concluida",1,0))`).

Os valores referentes ao atributo PendenciasAcademicas também são obtidos pela contagem, por estudante, das disciplinas “não concluídas” (`IIF(SM.Des_Situacao<>"Concluida")`), situadas nas matrizes dos cursos em semestres anteriores ao atualmente cursado pelo estudante (`AC.Semestre_Atividade_Curricular<A.SemestreAtual`).

A atualização da tabela temporária a partir da “temporaria2” é realizada através da consulta SQL 5b.

A tabela temporária, devidamente preenchida, servirá de base para a geração dos arquivos arff que serão utilizados como entrada de dados para as aplicações de mineração nos estudos de caso.

Para o Caso 1, que analisa a aplicação de um algoritmo de *clustering* nos dados identificados utilizamos a consulta SQL 6 sobre a tabela temporária. As expressões `T.Curso="C01", T.Curso="C02"` e `T.Curso="C03"` são simplificações, para fins de apresentação, das listas de cursos dos três grupos – Licenciaturas, Bacharelados e Tecnológicos.

O Caso 2, por outro lado, necessita de valores nominais para a aplicação do algoritmo de associação. Dessa forma foi utilizada a consulta SQL 7.

Na consulta, os valores numéricos do atributo IngressoNota são transformados em três valores nominais representando, respectivamente, ingresso sem notas, notas menores do que 60%, e notas maiores ou iguais a 60% (`IIF(ISNULL(T.IngressoNota),"SEM NOTA", IIF(T.IngressoNota<0.6,"<60%",">=60%"))`).

Os valores numéricos referentes ao tempo de antecipação de matrículas dos estudantes são transformados nos valores nominais referentes a cinco faixas – sem antecipação (`T.IngressoAntecipacaoMatricula Is Null Or T.IngressoAntecipacaoMatricula=0`), até uma semana de antecipação (`T.IngressoAntecipacaoMatricula<=7`), até duas semanas de antecipação (`T.IngressoAntecipacaoMatricula<=15`), até 1 mês

(`T.IngressoAntecipacaoMatricula<=30`) e mais de 1 mês de antecipação (demais valores).

Da mesma forma, os valores referentes a idade dos estudantes são transformados em três faixas de valores – menos de 25 anos (`T.Idade<=25`), entre 25 e 35 anos (`T.Idade<=35`) e maiores de 35 anos (demais valores).

A extração dos dados para mineração, tanto para a aplicação de classificação quanto para o algoritmo de agrupamento, quando realizada diretamente a partir dos bancos de dados dos sistemas em produção, sem a utilização do modelo proposto, além de demandar um conhecimento prévio das entidades e seus relacionamentos, necessita que seja constituído um *data set* único, intermediário entre os arquivos arff e as fontes de dados, consolidando as informações das suas diversas origens.

Para a preparação dessa base temporária, conforme apresentado, foram necessárias cinco consultas SQL, sendo duas delas de baixa complexidade (consultas SQL 1 e 2) e três de média a alta complexidade (consultas SQL 3 a 5), fazendo uso de recursos de cálculos e agrupamento de registros.

A partir da base intermediária constituída e devidamente preenchida, a extração dos dados para a geração dos arquivos de entrada para os algoritmos de mineração utilizou, ainda, mais duas consultas simples, uma para cada aplicação (consulta SQL 6, para o algoritmo de agrupamento – caso 1; e consulta SQL 7, para o algoritmo de classificação – caso 2). Os dados obtidos dessas consultas foram preparados segundo o formato requerido para o aplicativo de mineração de dados.

Todo o processo, desde as primeiras análises das bases de dados disponíveis até a obtenção dos dois arquivos arff para as aplicações de mineração, descontadas as interrupções, contabilizou aproximadamente duas horas e meia conforme a tabela 35.

Tabela 35 - Demonstração do tempo de extração de dados sem o modelo proposto

Atividade	Tempo Aproximado
Análise dos bancos de dados disponíveis	00:40
Modelagem da tabela temporária	00:10
Consulta SQL 1 – Codificação	00:05
Consulta SQL 1 - Execução e transferência dos dados para a tabela temporária.	00:05
Consulta SQL 2 – Codificação	00:05
Consulta SQL 2 - Execução e transferência dos dados para a tabela temporária.	00:05
Consulta SQL 3 – Codificação	00:10
Consulta SQL 3 - Execução e transferência dos dados para a tabela temporária.	00:05
Consulta SQL 4 – Codificação	00:10
Consulta SQL 4 - Execução e transferência dos dados para a tabela temporária.	00:05
Consulta SQL 5 – Codificação	00:15
Consulta SQL 5 - Execução e transferência dos dados para a tabela temporária.	00:05
Consulta SQL 6 – Codificação	00:05
Consulta SQL 6 - Execução e geração do arff.	00:10
Consulta SQL 7 – Codificação	00:10
Consulta SQL 7 - Execução e geração do arff.	00:10
Tempo Total	**02:35**

Do tempo total executado, somente trinta e cinco minutos foram dedicados à geração dos arquivos arff, enquanto que a preparação da base de dados temporária, que foi utilizada como fonte principal para esses arquivos necessitou de duas horas.

6.3.2.4 Extração de Dados com o modelo proposto

A extração dos dados para os algoritmos de mineração do estudo de caso, utilizando o modelo proposto pressupõe a existência de um *data set* contendo os dados na estrutura do modelo, devidamente integrado com as bases de dados da instituição e, por consequência, permanentemente atualizados. Dessa maneira, as consultas sobre esse *data set*, descritas no Anexo E são suficientes para a geração dos arquivos arff de entrada para a mineração.

A consulta SQL 8, utilizada para a extração dos dados para o Caso 1, de aplicação do algoritmo de *clustering*, a partir do *data set* constituído considerando o modelo de dados proposto. Na consulta, os dados referentes aos períodos letivos desejados são filtrados através da expressão `P.IDPeriodo IN (1, 2, 3)` enquanto que a restrição aos recém ingressados é realizada pela expressão `M.IdPeriodoMatricula = E.IdPeriodoIngresso`. O atributo Idade é calculado pela diferença entre a data vigente e a data de nascimento do estudante (`INT((Now() - E.DataNascimento)/365)`).

A consulta SQL 9, quando aplicada ao *data set* preparado a partir do modelo proposto, possibilita a criação do arquivo arff de entrada para a aplicação do algoritmo de classificação, utilizado no caso 2.

Na consulta foram mantidas as mesmas faixas de valores para os atributos IngressoNota (“SemNota”, “<60%”, >=60%”), IngressoAntecipacaoMatricula (“SemAntecipacao”, “AtéUmaSemana”, “AtéDuasSemanas”, “AtéUmMês”, “MaisdeUmMês”) e Idade (“Até25Anos”, “de25a35Anos”, “Maisque35anos”) utilizadas na Consulta SQL 7, para extrair os dados para o Caso 2, a partir do *dataset* temporário, sem a utilização do modelo proposto. Da mesma forma, os mesmos filtros foram utilizados garantindo a utilização do mesmo conjunto de dados.

A geração dos arquivos arff para as aplicações de mineração de dados a partir do *data set* criado com referência no modelo de dados proposto utilizou apenas as duas consultas SQL apresentadas (consultas SQL 8 e 9, respectivamente, para os caso 1 e 2). O tempo decorrido entre a primeira análise da base de dados com sua documentação e a versão final dos dois arquivos arff para as aplicações de mineração foi de aproximadamente uma hora, conforme demonstrado na tabela 36.

Tabela 36 - Demonstração do tempo de extração de dados com o modelo proposto

Atividade	Tempo Aproximado
Análise do modelo proposto e sua documentação	00:15
Consulta SQL 8 – Codificação	00:05
Consulta SQL 8 - Execução e geração do arff.	00:10
Consulta SQL 9 – Codificação	00:15
Consulta SQL 9 - Execução e geração do arff.	00:10
Tempo Total	**00:55**

Vale destacar que apenas quinze minutos foram despendidos para análise e planejamento da extração dos dados, correspondente a 27% do tempo total.

Na próxima seção, uma análise é apresentada, comparando os esforços demandados na extração dos dados nos dois processos descritos.

6.3.2.5 Análise dos esforços de extração de dados

Conforme descrito nas seções anteriores, a geração dos arquivos de entrada para as aplicações de mineração de dados foi realizada através de dois processos distintos – o

primeiro, sem a utilização do modelo proposto, extraindo os dados a partir das bases originais; e o segundo, utilizando um *data set* construído a partir do modelo de dados proposto. Dessa maneira, quatro arquivos foram gerados ao todo, sendo dois arquivos gerados pelo primeiro processo – um para a aplicação de agrupamento (arquivo A) e um para a aplicação de classificação (arquivo B); e dois arquivos pelo segundo processo, também um arquivo para o agrupamento (arquivo C) e um para a classificação (arquivo D).

Tanto o par de arquivos destinados à aplicação de agrupamento (A e C), quanto os dois arquivos para a aplicação de classificação (B e D) foram produzidos idênticos entre si. Como os processos distintos geraram pares idênticos de arquivos, podemos afirmar que ambos são processos tecnicamente válidos para a preparação dos dados para mineração.

A comparação dos esforços requeridos para a extração de dados pelos dois processos pode ser realizada verificando, não apenas o tempo total necessário para as suas execuções mas também a quantidade de etapas realizadas e sua complexidade.

Do ponto de vista do tempo dedicado, o primeiro processo, sem a utilização do modelo, durou cerca de duas horas e meia para a sua conclusão enquanto que o segundo processo, que utilizou o modelo proposto, necessitou de apenas cinquenta e cinco minutos - apenas 20% do tempo do primeiro processo. Apesar de o tempo necessário para a geração dos arquivos ter sido praticamente o mesmo nos dois processos, a preparação dos *data sets* para a sua geração foi oito vezes menor no processo que utilizou o modelo do que no outro, sem a sua utilização. A Tabela 37 demonstra a comparação dos tempos entre os dois processos.

Tabela 37 - Comparação do tempo de extração de dados nos dois processos realizados

Atividade/Tempo	Sem o modelo proposto	Com o modelo proposto
Preparação do data set	02:00	00:15
Geração dos arquivos arff	00:35	00:40
TOTAL	02:35	00:55

Analisando as etapas realizadas para a extração dos dados e geração dos arquivos para a mineração, identificamos que o primeiro processo necessitou de dezesseis etapas, com a realização de oito operações de banco de dados – duas inserções (INSERT), quatro alterações de dados (UPDATE) e duas consultas (SELECT). A existência de estruturas de bancos de dados distintas foi determinante na complexidade do processo, uma vez que exigiu a criação de duas tabelas temporárias além da manipulação de treze tabelas distintas ao longo das operações de banco de dados.

A extração pelo segundo processo, baseado no modelo de dados proposto, necessitou somente de cinco etapas para a sua realização com a execução de apenas duas operações de consulta (SELECT), uma vez que o *data set* já se encontrava pré-estruturado para a realização de pesquisas na área de educação, facilitando o planejamento para a preparação dos dados para mineração. A tabela 38 apresenta um comparativo dos esforços entre os dois processos.

Tabela 38 - Comparação dos esforços necessários nos dois processos realizados

Atividade/Tempo	Sem o modelo proposto	Com o modelo proposto
Quantidade de Etapas Realizadas	16	5
Quantidade de Consultas Realizadas	8	2
Quantidade de Tabelas distintas envolvidas	15	5

A comparação entre os dois processos realizados aponta para vantagens quantitativas e qualitativas na utilização de um *data set* atualizado e estruturado segundo o modelo de dados proposto, principalmente nas etapas de planejamento da preparação de dados para mineração. A existência de uma estrutura de dados especificamente destinada a análises gerenciais do negócio da educação facilita a identificação das consultas necessárias à extração dos dados e, consequentemente, a geração dos arquivos para a mineração de dados.

A estruturação prévia e sistemática de uma base de dados para análises gerenciais, utilizando o modelo de dados propostos elimina diversas etapas na extração de dados para mineração, reduzindo o tempo necessário para a sua preparação de dados. Dessa maneira, fica confirmada a Hipótese Alternativa H3 do estudo de caso, que afirma que o esforço para a preparação dos dados para a aplicação de técnicas de mineração de dados a partir das tabelas dos bancos de dados da instituição de ensino ($\mathbf{E_{DataBase}}$) é maior do que o esforço a partir de um *data set* preparado de acordo com o modelo proposto ($\mathbf{E_{DataSet}}$).

Nas próximas seções, os arquivos gerados para a mineração de dados serão utilizados nos casos estudados, em uma aplicação de algoritmo de *clustering* (Caso 1) e em uma aplicação de algoritmo de classificação (Caso 2).

6.3.3 Aplicação de Algoritmos de Mineração de Dados

A hipótese alternativa H1 afirma que a utilização de um subconjunto de atributos do modelo proposto em aplicações de mineração de dados realmente propicia a identificação de indícios de evasão de estudantes. Segundo o planejamento realizado, o *data set* gerado a partir

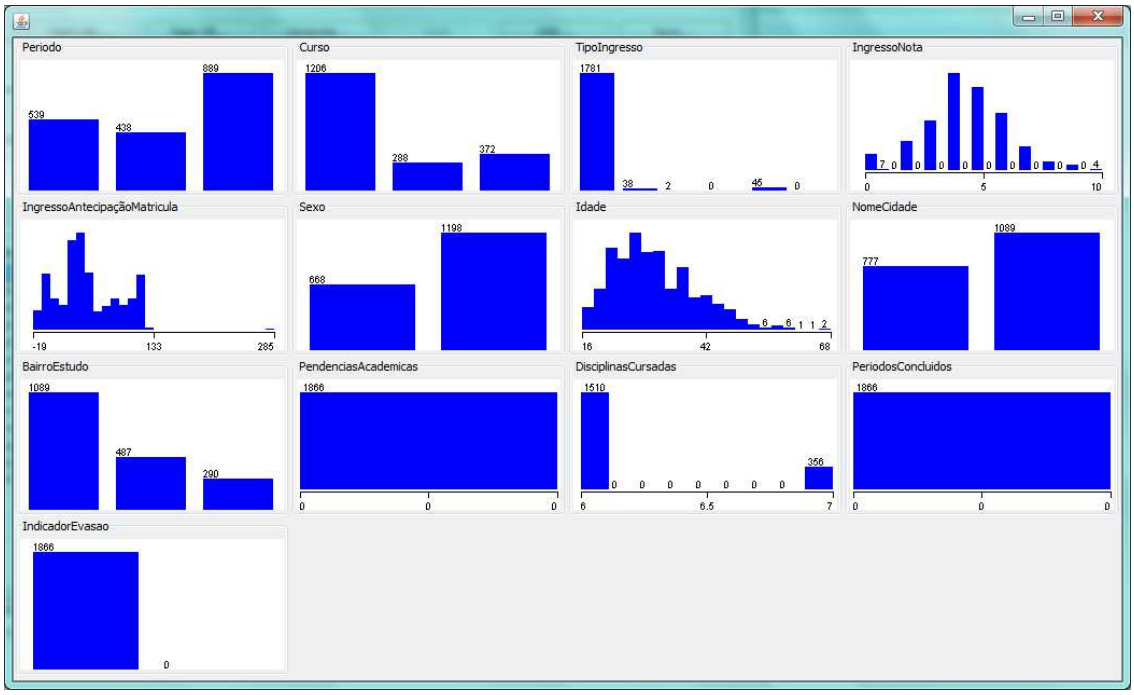
do modelo de dados proposto servirá de fonte para as entradas de dados utilizadas nos dois casos de aplicação de mineração de dados cujos resultados ($\mathbf{I_{Mineração}}$) serão comparados com os resultados da análise estatística direta ($\mathbf{I_{Análise}}$) apresentada na seção *6.3.2.1. Análise Estatística dos dados obtidos*. A identificação de resultados comuns às duas análises comprova a hipótese ($\mathbf{I_{Análise}} \cap \mathbf{I_{Mineração}} \neq \emptyset$).

6.3.3.1 Caso 1 – Aplicação de Algoritmo de Clustering

O Agrupamento ou *Clustering* identifica similaridades entres os valores dos atributos analisados e, a partir dessa análise, particiona a base de dados em grupos. Para a execução da técnica, no estudo de caso, foi selecionado o algoritmo *SimpleKMeans* que, a partir da indicação da quantidade (k) de *clusters* desejada, divide a base de dados de forma que a similaridade dos elementos de cada *cluster* seja alta e, entre os *clusters* seja baixa.

O arquivo de entrada de dados gerado para essa aplicação, descrito na seção anterior, foi carregado no WEKA onde algumas análises e considerações foram realizadas sobre a distribuição dos valores dos atributos e seu impacto na atividade. A Figura 23 apresenta as distribuições dos valores de cada atributo da base de dados carregada. Como pode ser observado, os atributos PendenciasAcademicas, PeriodosConcluidos e IndicadorEvasao apresentam apenas um valor, cada, em toda a base utilizada. Dessa forma, não possuem nenhuma interferência na criação dos agrupamentos.

Figura 23 - Representação gráfica da distribuição dos valores do arquivo de entrada para o caso 1



O algoritmo *simpleKmeans* apresenta algumas variáveis de conFiguração para a sua execução:

- *displayStdDevs* - Indica a exibição de desvios padrão dos atributos numéricos e contagens de atributos nominais. Seu valor padrão é *false*.
- *distanceFunction* – Determina a função distância a ser usada para comparação das instâncias. Como padrão, é utilizada a *weka.core.EuclideanDistance*.
- *dontReplaceMissingValues* – Indica se os valores faltantes devem ser substituídos pela média ou moda. A indicação padrão para esse parâmetro é *false*, permitindo a substituição dos valores ausentes.
- *fastDistanceCalc* – Indica a utilização de "pontos de corte" para acelerar o cálculo da distância. Possui valor inicial *false*.
- *initializeUsingKMeansPlusPlusMethod* – Determina a detecção dos centros dos *clusters* através do método probabilístico k-means++. O valor padrão para o parâmetro é *false*.
- *maxIterations* – Determina o número máximo de iterações. Sugere-se como valor padrão 500 iterações.
- *numClusters* – Determina o número de *clusters* a ser gerado. A indicação inicial aponta a geração de apenas dois agrupamentos.
- *preserveInstancesOrder* – Indica se a ordem original das instâncias deve ser preservada. Por padrão, o valor *false* indica que a ordem das instâncias pode ser modificada.
- *seed* - Referência para a utilização na geração de valores aleatórios.

Desses parâmetros, alteramos somente a indicação da quantidade de *clusters* a serem gerados (*numClusters*). Para determinar o melhor valor para o indicador, foram realizados alguns ensaios, alternando valores e observando resultados, sobretudo a dispersão dos *clusters* gerados. A utilização de dois *clusters* como sugere o valor padrão do Weka resulta numa distribuição dos registros na ordem de 42% e 58%. Por outro lado, ao utilizarmos 10 *clusters* obtemos valores percentuais da distribuição entre 4% e 18%, com uma variação entre 6 e 8 pontos percentuais do valor médio. A utilização de 15 *clusters* foi selecionada por evidenciar 3 grupos com o dobro do percentual médio da base, conforme as Figuras 24 e 25, que

apresentam o resultado gerado pelo Weka a partir da aplicação do algoritmo *simpleKMeans* na base de dados do Caso 1.

Figura 24 - Resultado da aplicação do algoritmo de Agrupamento na base de dados do Caso 1

```
Number of iterations: 35
Within cluster sum of squared errors: 857.4328575259917
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data          0          1          2          3          4          5          6
                             (1866)     (268)      (88)     (116)      (74)     (115)     (147)     (226)
=============================================================================================================================
Curso                     BACHARELADO BACHARELADO BACHARELADO BACHARELADO BACHARELADO BACHARELADO LICENCIATURA BACHARELADO
TipoIngresso               VESTIBULAR VESTIBULAR VESTIBULAR VESTIBULAR VESTIBULAR VESTIBULAR VESTIBULAR VESTIBULAR
IngressoNota                    4.418     4.5746     4.6818     5.1552     4.2838     4.0957     4.9184     5.2301
IngressoAntecipaçãoMatricula   49.9496    50.5485         28   104.9741        103     61.487    51.6871    43.4425
Sexo                                2          1          1          2          1          2          2          2
Idade                         31.0531     30.306    24.6023    25.4138    27.0946    40.9043    31.1293    27.3451
NomeCidade                   INTERIOR    CAPITAL   INTERIOR   INTERIOR   INTERIOR   INTERIOR   INTERIOR    CAPITAL
BairroEstudo                  CENTRO    IGUATEMI     CENTRO     CENTRO     CENTRO     CENTRO     CENTRO    IGUATEMI

Time taken to build model (full training data) : 0.23 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0       268 ( 14%)
 1        88 (  5%)
 2       116 (  6%)
 3        74 (  4%)
 4       115 (  6%)
 5       147 (  8%)
 6       226 ( 12%)
 7       112 (  6%)
 8        80 (  4%)
 9       161 (  9%)
10        63 (  3%)
11        59 (  3%)
12       230 ( 12%)
13        48 (  3%)
14        79 (  4%)
```

Figura 25 - Resultado da aplicação do algoritmo de Agrupamento na base de dados do Caso 1 – continuação

7	8	9	10	11	12	13	14
(112)	(80)	(161)	(63)	(59)	(230)	(48)	(79)
BACHARELADO	BACHARELADO	TECNOLOGICO	BACHARELADO	LICENCIATURA	BACHARELADO	BACHARELADO	BACHARELADO
VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR
3.3214	3.225	5.0807	1.9524	1.661	5.1304	6.5417	1.9241
57.7054	47.0125	40.7453	30.4762	39.3898	27.2174	57.6875	36.1013
2	1	2	1	2	2	1	2
40.125	42.4125	38.4037	25.8889	32.4237	25.5478	33.6667	27.0759
CAPITAL	INTERIOR	CAPITAL	INTERIOR	INTERIOR	INTERIOR	INTERIOR	INTERIOR
IGUATEMI	CENTRO	IGUATEMI	CENTRO	CENTRO	CENTRO	CENTRO	CENTRO

O *cluster* 0 , que abrange 14% dos dados, aponta evadidos de cursos de bacharelado da capital, do sexo feminino, com média de idade de 30 anos, que ingressaram por vestibular, com nota média de 4,5 e anteciparam sua matrícula em cerca de 2 meses. Já o *cluster* 6, com 12% de ocorrências, apresenta uma pequena variação em relação ao primeiro - evadidos de cursos de bacharelado, da capital, do sexo masculino, com média de idade de 27 anos, que

ingressaram por vestibular, com nota média de 5,2 e anteciparam sua matrícula em cerca de 1 mês e meio.

Chama a atenção também o *cluster* 12. Este também possui 12% de incidência, que aponta para evadidos do interior do Estado, do sexo masculino e média de idade de 25 anos, oriundos de cursos de bacharelado que ingressaram por vestibular com nota média de 5,1 e antecipação de cerca de 1 mês.

A separação dos *clusters* pela aplicação aponta para a predominância, entre os evadidos, de estudantes de cursos de Bacharelado, que ingressaram por Vestibular com notas relativamente baixas (entre 4 e 5). Predominam nos grupos, também, estudantes do sexo masculino, do interior do Estado, com idade média de 20 anos.

6.3.3.2 Caso 2 – Aplicação de Algoritmo de Classificação

A geração de árvores de decisão através de algoritmos de classificação na base de dados do Caso 2 permite identificar, hierarquicamente, os atributos que mais contribuem para a evasão, assim como quais as relações entre atributos e seus valores tem maior possibilidade de determinar a permanência dos estudantes. Foi utilizado para a técnica o algoritmo J48, que procura formar a árvore mais adequada sobre o conjunto de dados através da poda de regras, mantendo as que melhoram a sua eficiência.

A conFiguração da execução do algoritmo no Weka utiliza os seguintes parâmetros:

- *binarySplits* – Indica a utilização de divisões binárias nos atributos nominais (valor padrão=*false*).
- *collapseTree* – Indica se as partes que não diminuem erros de treinamento devem ser removidas (valor padrão=*true*).
- *confidenceFactor* – Determina o fator de confiança utilizado para o comprimento dos galho ou “poda” (valor padrão=0.25. Menores valores incorrem em podas maiores e galhos menores).
- *debug* - Indica se a ferramenta deve exibir informações adicionais para o console (valor padrão=*false*).
- *minNumObj* – Número mínimo de instâncias por folha (valor padrão=2).
- *reducedErrorPruning* – Indica a poda a ser utilizada. O valor *true* indica a utilização da “poda de erros reduzidos”, enquanto que *false* (padrão) aponta para a utilização da “poda C.4.5”.

- *numFolds* – Indica a quantidade de dados utilizada para a "poda de erros reduzidos" (valor padrão=3).
- *saveInstanceData* – Indica se os dados de treinamento gerados devem ser salvos para future visualização (valor padrão=*false*).
- *seed* – Determina o valor base para a randomização de dados ao utilizar a "poda de erros reduzidos" (valor padrão=1).
- *subtreeRaising* – Indica, se verdadeiro (padrão), se a poda deve considerar o crescimento de subárvores.
- *unpruned* – O valor *false* (padrão) determina que a poda deve ser realizada.
- *useLaplace* – Determina que a contagem de folhas deve ser suavizada com base em Laplace[1] (valor padrão=*false*).
- *useMDLcorrection* – O valor *true* (padrão) indica que a correção MDL[2] deve ser utilizada ao identificar divisões em atributos numéricos.

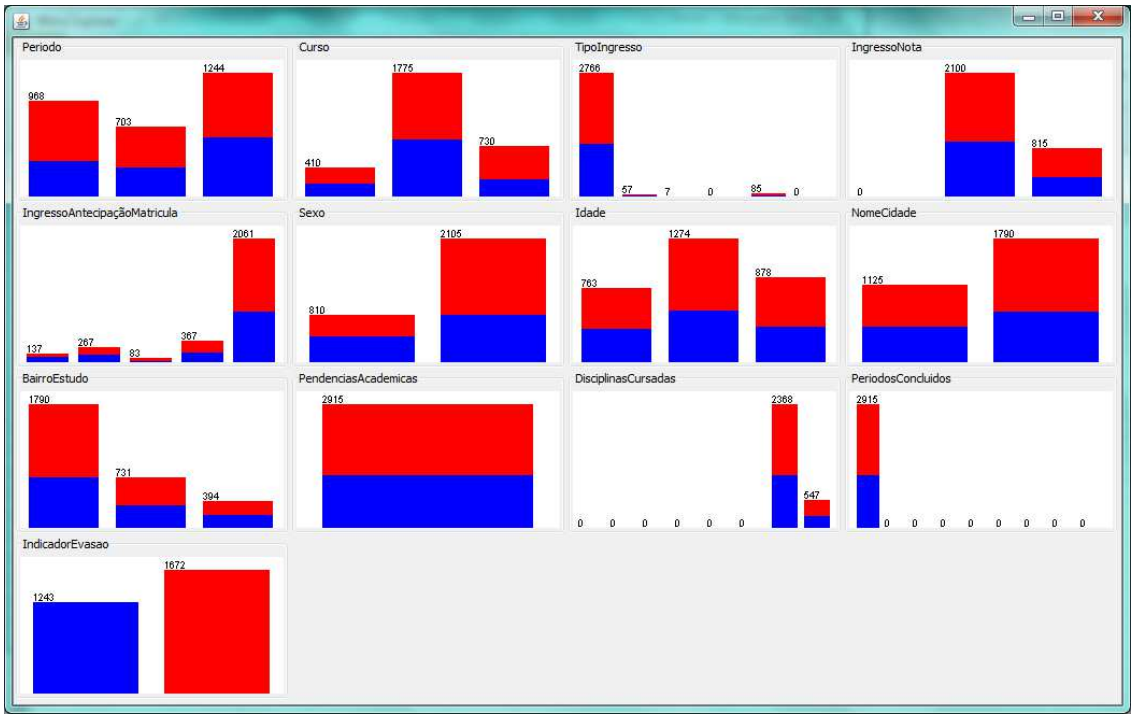
Para a aplicação da técnica, a base de dados gerada foi separada em dois arquivos distintos – o primeiro contendo 2/3 dos dados selecionados aleatoriamente, foi utilizado para o treinamento e geração da árvore modelo; e o segundo, contendo os demais dados, com o objetivo de testar a exatidão da árvore modelo gerada. A realização das duas etapas se justifica pela necessidade de utilização do modelo em valores desconhecidos e não apenas nos dados de que dispomos. Ao aplicar o modelo no conjunto de dados de teste, garantimos que a sua exatidão permanece a mesma para qualquer conjunto de dados (WITTEN *et al.*, 2005).

Composta por 2.915 instâncias (2/3 dos 4.372 registros), o conjunto de treinamento foi carregado no Weka. A distribuição dos dados pode ser verificada na Figura 26.

[1] "A suavização de Laplace (ou Laplace smoothing) é uma técnica geralmente utilizada para se evitar que cálculos probabilísticos resultem em zero por um devido fator nulo dentro de uma série em um produtório." (ALVES, F. F., 2012)

[2] Do ingles Minimum Description Length ou Tamanho Mínimo de Descrição, este método mede o tamanho de uma árvore de decisão por meio do número de bits necessários para codificar a árvore e determina árvores codificadas com menor quantidade de bits (ROKACH et al, 2008).

Figura 26 - Representação gráfica da distribuição dos valores do arquivo de entrada para o caso 2



Ao executar o algoritmo de classificação J48 na base carregada, com a indicação de utilizar o conjunto de treinamento (*Use training test*), o modelo da árvore foi gerado indicando uma exatidão de aproximadamente 65% conforme a Figura 27.

Figura 27 - Sumário do resultado do algoritmo J48 no conjunto de treinamento

```
=== Summary ===

Correctly Classified Instances        1880               64.494  %
Incorrectly Classified Instances      1035               35.506  %
Kappa statistic                          0.2285
Mean absolute error                      0.4493
Root mean squared error                  0.474
Relative absolute error                 91.8422 %
Root relative squared error             95.8351 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      99.2281 %
Total Number of Instances             2915
```

A execução do algoritmo no conjunto de testes, com base no modelo criado, apresentou uma exatidão de 61,5%, similar à do conjunto de treinamento, confirmando a operação. A Figura 28 apresenta o sumário dessa execução.

Figura 28 - Sumário do resultado do algoritmo J48 no conjunto de testes

```
=== Summary ===

Correctly Classified Instances         897               61.5649 %
Incorrectly Classified Instances       560               38.4351 %
Kappa statistic                          0.1682
Mean absolute error                      0.4696
Root mean squared error                  0.4963
Relative absolute error                 95.9596 %
Root relative squared error            100.3113 %
Coverage of cases (0.95 level)          98.8332 %
Mean rel. region size (0.95 level)      98.8675 %
Total Number of Instances             1457
```

Um grau de exatidão entre 60 e 65% indica uma árvore de decisão com necessidades de melhoria para a produção de resultados mais conclusivos sobre o tema analisado. A introdução de novos dados ou a utilização de um conjunto de atributos diferente, com refinamentos sucessivos do processo pode levar a melhores níveis de exatidão, mas, para efeito do estudo de caso atual, entendemos que o grau de exatidão obtido é suficiente para as análises desejadas de avaliação da hipótese formulada.

A Figura 29 apresenta uma representação simplificada da árvore de decisão gerada pelo algoritmo J48 a partir dos dados fornecidos. Dispostas em duas colunas, separadas pelo atributo-raiz (Sexo), estão apresentados somente os ramos e folhas que indicam a condição de evasão (SIM). A árvore completa pode ser verificada no documento anexo (Anexo D – Árvore de decisão completa gerada no estudo de caso).

Figura 29 - Representação simplificada da árvore de decisão gerada no Caso 2

Sexo = 1	Sexo = 2
| Periodo = 20111	| AntecipaçãoMatricula = SemAntecipacao
| | DisciplinasCursadas = 6	| | Periodo = 20111
| | | NomeCidade = CAPITAL: SIM (66.0/26.0)	| | | Curso = TECNOLOGICO: SIM (5.0/1.0)
| Periodo = 20112	| | | Curso = BACHARELADO: SIM (5.0/2.0)
| | AntecipaçãoMatricula = SemAntecipacao	| | Periodo = 20112
| | | BairroEstudo = CENTRO: SIM (9.0/2.0)	| | | BairroEstudo = IGUATEMI
| | | BairroEstudo = PARALELA: SIM (1.0)	| | | | Curso = TECNOLOGICO: SIM (2.0)
| | AntecipaçãoMatricula = AtéUmaSemana: SIM (77.0/23.0)	| | | | Curso = BACHARELADO: SIM (1.0)
| | AntecipaçãoMatricula = AtéDuasSemanas	| | | BairroEstudo = PARALELA: SIM (4.0)
| | | BairroEstudo = PARALELA: SIM (5.0/1.0)	| | Periodo = 20121: SIM (41.0/12.0)
| | AntecipaçãoMatricula = AtéUmMês	| AntecipaçãoMatricula = AtéUmaSemana
| | | BairroEstudo = IGUATEMI: SIM (12.0/2.0)	| | NomeCidade = CAPITAL
| | | BairroEstudo = PARALELA	| | | Idade = Até25anos

```
| | | | Idade = Até25anos: SIM (0.0)
| | | | Idade = de25a35anos: SIM (3.0)
| | | | Idade = Maisque35anos
| | | | | IngressoNota = >=6: SIM (3.0/1.0)
| | AntecipaçãoMatricula = MaisDeUmMês
| | | Idade = Até25anos
| | | | TipoIngresso = VEST
| | | | | IngressoNota = <6: SIM (3.0/1.0)
| | | | TipoIngresso = ENEM: SIM (3.0)
| | | | TipoIngresso = ME: SIM (0.0)
| | | | TipoIngresso = TE: SIM (0.0)
| | | | TipoIngresso = TI: SIM (0.0)
| | | | TipoIngresso = PROUNI: SIM (0.0)
| | | Idade = Maisque35anos
| | | | BairroEstudo = CENTRO
| | | | | IngressoNota = <6: SIM (3.0/1.0)
| | | | BairroEstudo = IGUATEMI: SIM (4.0/1.0)
| Periodo = 20121: SIM (354.0/138.0)
| | | | Curso = TECNOLOGICO: SIM (1.0)
| | | | Curso = LICENCIATURA
| | | | | BairroEstudo = PARALELA: SIM (2.0)
| | | Idade = de25a35anos
| | | | IngressoNota = SemNota: SIM (0.0)
| | | | IngressoNota = <6: SIM (17.0/5.0)
| | | | IngressoNota = >=6
| | | | | Curso = TECNOLOGICO: SIM (6.0/3.0)
| | | | | Curso = BACHARELADO: SIM (8.0/2.0)
| | | Idade = Maisque35anos
| | | | BairroEstudo = PARALELA
| | | | | Curso = TECNOLOGICO: SIM (2.0)
| | | | | Curso = BACHARELADO: SIM (3.0/1.0)
| AntecipaçãoMatricula = AtéUmMês
| | Periodo = 20112
| | | TipoIngresso = ENEM: SIM (5.0/1.0)
| | Periodo = 20121: SIM (25.0/8.0)
```

Analisando a árvore gerada, podemos inferir algumas situações de evasão indicadas. Identificamos, por exemplo, que estudantes do sexo feminino que ingressaram em 2011.2 com seis disciplinas na capital são, potencialmente, um grupo de evasão, uma vez que das 66 instâncias identificadas segundo essas características, apenas 26 não evadiram, ou seja, possuem 71,7% de ocorrência de evadidos.

Os atributos Período e AntecipacaoMatricula estão mais próximos da raiz, respectivamente, para instâncias do sexo feminino (1) e masculino (2), enquanto que Idade, TipoIngresso e IngressoNota se apresentam, predominantemente, mais próximos das folhas da árvore.

Chama a atenção também o fato de 81% (46 instâncias, de um total de 60) dos estudantes do sexo masculino que anteciparam suas matrículas em até duas semanas persistirem nos seus cursos, indicando um baixo índice de evasão nessa categoria.

6.3.4 Análise dos Resultados

A hipótese alternativa H1 formulada para o estudo de caso afirma que a utilização do modelo proposto em técnicas de mineração de dados propicia a identificação de indícios de

evasão em estudantes. A sua formulação indica que deve haver um conjunto de indícios comuns, detectados, tanto através da utilização de técnicas de mineração com o modelo de dados proposto, quanto através da observação estatística direta nos dados utilizados.

Um dos indícios de evasão identificados através da análise direta da base de dados, como relatado na seção *6.3.2.1. Análise Estatística dos dados obtidos*, encontra-se em cursos de Graduação Tecnológica, mesmo que discretamente, há uma pequena tendência a evadir em cursos dessa categoria do que em Bacharelados ou Licenciaturas. A análise da árvore de decisão gerada aponta cinco conjuntos de instâncias que contêm dados dessa categoria, todas com um alto índice de ocorrências de acertos nas regras, como pode ser notado na Tabela 39.

Tabela 39 - Ramos da árvore de decisão que incluem estudantes de Graduação Tecnológica

NÍVEL 1 (SEXO)	NÍVEL 2 (ANTECIPAÇÃO DA MATRÍCULA)	NÍVEL 3	NÍVEL 4	NÍVEL 5	OCORRÊNCIAS	% ACERTO
MASC	SemAntecipacao	Periodo = 20111			(5.0/1.0)	83,30%
MASC	SemAntecipacao	Periodo = 20112	BairroEstudo = IGUATEMI		(2.0)	100,00%
MASC	AtéUmaSemana	Cidade = CAPITAL	Idade = Até25anos		(1.0)	100,00%
MASC	AtéUmaSemana	Cidade = CAPITAL	Idade = de25a35anos	IngressoNota = >=6	(6.0/3.0)	66,70%
MASC	AtéUmaSemana	Cidade = CAPITAL	Idade = Maisque35anos	BairroEstudo = PARALELA	(2.0)	100,00%

Por outro lado, nas informações sobre evasão, extraídas a partir do resultado da aplicação do algoritmo de agrupamento, no caso 1, não foi encontrado nenhum destaque a respeito de cursos de Graduação Tecnológica e, pelo contrário, um grupo se evidenciou dos demais pelo seu percentual de incidência relacionado a cursos de Bacharelado. Apesar de, aparentemente contraditórios, esses resultados não são incompatíveis, uma vez que é possível, de acordo com a classificação, identificar situações com maior potencial de evasão para cursos de determinada categoria e, com a aplicação de outras técnicas, serem encontrados resultados que destacam outras situações.

A análise estatística direta apontou uma maior tendência a evadir em estudantes oriundos de processos de matrícula especial (portadores de diploma superior) e transferência externa, que apresentou uma proporção no grupo de evadidos (8,9%) maior do que na totalidade na amostra (0,2%), conforme apresentado na *Figura 23 Distribuição da amostra de evadidos por Forma de Ingresso*. A aplicação do algoritmo de Agrupamento, por outro lado, evidenciou que ingressantes por Vestibular possuem uma maior propensão a evadir. Isso ocorre porque, diferentemente da análise estatística, realizada atributo a atributo, a técnica do Agrupamento busca o comportamento dos dados a partir do conjunto dos atributos utilizados.

Ao analisar o impacto da antecipação das matrículas dos estudantes no fenômeno da evasão, porém, as três análises convergem quando apontam que estudantes que se matriculam com muita antecedência tendem a evadir mais que aqueles que o fazem em períodos mais próximos do encerramento do prazo. A análise direta indica que esse grupo, dentre os evadidos, tem uma incidência de 73,2%, diferentemente da incidência no universo de matriculados, com 57,1%. Os *clusters* de evadidos gerados na aplicação do algoritmo do caso 1 apontam uma média de 40 dias de antecipação e, na aplicação do algoritmo de classificação, no caso 2, o atributo aparece próximo do topo da árvore, determinando a evasão em 45 ocorrências, com incidência média de 70%.

Diversas inferências podem ser realizadas sobre potenciais fatores de evasão de estudantes ($I_{Mineração}$) a partir dos resultados das aplicações dos algoritmos dos casos 1 e 2 e, como descrito, alguns desses indícios convergem com as análises realizadas a partir dos dados estatísticos dos dados, antes da mineração ($I_{Análise}$). Para efeito de demonstração e comprovação da Hipótese Alternativa H1, apenas algumas análises foram realizadas sobre os resultados dos algoritmos de mineração, apontando, de maneira suficiente, que a utilização do modelo proposto em técnicas de mineração de dados proporciona a identificação de indícios de evasão de estudantes ($I_{Análise} \cap I_{Mineração} \neq \emptyset$). Futuras análises, com mais detalhamento sobre esses resultados e com refinamentos na execução dos algoritmos nas bases de dados construídas podem trazer mais e melhores conclusões sobre os impactos desses atributos nas decisões dos estudantes em evadir ou não de seus cursos.

6.4 CONCLUSÃO

Este capítulo apresentou o estudo de caso planejado para avaliar a aplicabilidade do modelo proposto para a utilização em análises de mineração de dados educacionais. O estudo, composto por uma hipótese nula (H0) e três hipóteses alternativas (H1, H2 e H3) buscou avaliar, num cenário real de uma instituição de ensino superior, as vantagens em utilizar um conjunto de dados construído sob um modelo de dados desenvolvido com esse fim específico.

Através das avaliações das hipóteses alternativas H2 e H3, as vantagens em utilizar o modelo de dados proposto nas etapas de preparação de dados para a mineração foram evidenciadas, na medida em que, tanto o esforço na identificação de atributos pertinentes à evasão de estudantes ($\mathbf{E_{Modelo}}$), quanto o esforço na construção dos arquivos de entrada para a mineração ($\mathbf{E_{DataSet}}$) com a utilização do modelo foram menores que os esforços para a realização dessas etapas sem o modelo proposto como base ($\mathbf{E_{Direto}}$ e $\mathbf{E_{DataBase}}$).

Na hipótese alternativa H1, a aplicabilidade do modelo foi posta à prova através da realização de dois casos práticos de mineração de dados educacionais utilizando os arquivos de entrada gerados a partir dos atributos do modelo proposto. A identificação de indícios de evasão através desse processo ($I_{Mineração}$), que converge com indícios encontrados pela análise estatística direta da base de dados ($I_{Análise}$), comprovou a validade da hipótese.

A confirmação das três hipóteses alternativas (H1, H2 e H3), consequentemente, permite refutar a hipótese nula (H0), segundo a qual a utilização de um subconjunto de atributos do modelo proposto não possui nenhum impacto na aplicação de técnicas de mineração de dados na identificação de indícios de evasão de estudantes, confirmando a aplicabilidade do modelo proposto.

O estudo realizado foi importante, também, no aspecto da análise do fenômeno da evasão de estudantes, trazendo alguns fatores que poderão indicar, em cursos da modalidade à distância, uma predisposição dos estudantes em evadir. Mesmo utilizando esse estudo como cenário para a comprovação das condições de aplicabilidade do modelo proposto, seus resultados podem ser utilizados como pontos de partida em pesquisas mais detalhadas sobre o assunto, levando a análises mais conclusivas sobre o tema.

7 CONSIDERAÇÕES FINAIS

O conceito da mineração de dados educacionais agrega os estudos e aplicações de técnicas de mineração de dados no apoio aos processos decisórios dos diversos atores envolvidos na oferta de cursos de graduação, além de concentrar a comunidade científica interessada no tema para o compartilhamento de experiências e sua consequente evolução.

Identificando a necessidade de uma maior homogeneização dos conceitos e do estabelecimento de padrões comuns para enriquecer esse intercâmbio de experiências, o trabalho atual teve por objetivo buscar essa convergência de conceitos através da proposição de um modelo de dados que agregasse os principais indicadores da área educacional.

7.1 RESULTADOS OBTIDOS

O trabalho atual tomou como ponto de partida a necessidade de investigar o fenômeno da evasão de estudantes em cursos de educação a distancia e buscar entender que fatores podem determinar um estudante a interromper seu curso. Foi realizada uma revisão da literatura sobre o tema, onde foram encontradas duas linhas de pesquisa. A primeira linha trabalhava questões conceituais sobre o comportamento de estudantes em seus cursos, apontando fatores que, em maior ou menor grau contribuíam para a sua permanência nos cursos. A segunda linha apontava uma concentração de trabalhos onde a mineração de dados surgia como forte aliada no apoio aos processos decisórios educacionais, em questões acadêmico-administrativas, como o combate à evasão, determinando, inclusive, a área de conhecimento chamada de Mineração de Dados Educacionais - MDE.

Iniciou-se, então, uma busca pelo estado da arte da mineração de dados educacionais e, mesmo tendo encontrado diversos trabalhos com objetivos e problemas convergentes, dificilmente suas terminologia e abordagens convergiam. Muitos focavam em questões específicas de suas instituições de ensino e relatavam suas experiências, mas poucas análises comparativas de trabalhos foram encontradas. Outro ponto de destaque encontrado foi a pouca importância dada nas aplicações de mineração de dados à etapa de preparação dos dados, sobretudo no que se refere às justificativas de escolha dos atributos utilizados.

A partir desses resultados, percebeu-se a necessidade de se iniciar um processo de homogeneização de terminologias e conceitos para MDE com a proposta de um modelo de dados para mineração que pudesse ser utilizado como referencial para futuros trabalhos. Orientado pelos modelos conceituais da literatura e por indicações de trabalhos sobre

mineração, foi identificado um conjunto de atributos devidamente justificados que, organizados em entidades de dados, compuseram o modelo proposto.

Uma vez concebido e documentado, o modelo de dados foi avaliado com a execução de um estudo de caso, onde foram avaliadas hipóteses sobre a sua aplicabilidade em casos reais, além do seu impacto na redução dos esforços na etapa de pré-processamento de dados para a mineração – seleção de atributos e preparação de dados.

7.2 CONTRIBUIÇÕES DA PESQUISA

A definição do modelo teve como principal desafio a busca por um conjunto de entidades e atributos que fossem pertinentes aos principais focos de análise de gestores de cursos e pesquisadores e, além disso, a sua organização de maneira que facilitasse a extração de dados na montagem de data sets para mineração de dados. A identificação, na literatura, de modelos conceituais do comportamento de estudantes lastreou essa seleção dos atributos e indicadores utilizados.

A utilização de dados reais na montagem de um conjunto de dados que, à luz do modelo proposto, validasse a sua aplicabilidade foi dificultada pela baixa predisposição das instituições de ensino em disponibilizar seus dados para a pesquisa, inseguras quanto à exposição de dados estratégicos ao mercado e à concorrência. Os dados somente puderam ser utilizados no estudo de caso, sob a condição de não identificar a sua origem – cursos ou instituições de ensinos.

A intenção de iniciar um processo de homogeneização de termos e conceitos com a proposição de um modelo de dados que sirva como referência para futuras pesquisas, mesmo passível de evoluções a partir dos pontos de vista de outros pesquisadores pode ser entendida como a maior contribuição do trabalho atual.

7.3 LIMITAÇÕES DO TRABALHO

As limitações do trabalho atual concentram-se, sobretudo, no processo de avaliação da aplicabilidade do modelo proposto:

- Foram utilizados, somente, dois algoritmos de mineração de dados – J48 de classificação e SimpleKMeans, de clusterização. Avaliações com outros algoritmos dessas e de outras técnicas podem reforçar a aplicabilidade do modelo.

- Um modelo de dados referencial deve ser passível de utilização em diferentes situações problemas do âmbito educacional. O estudo de caso se concentrou na análise de um único caso prático real – o problema da evasão de estudantes em cursos de graduação EAD.
- A utilização de bases de dados de somente uma instituição de ensino também restringe o leque de análises acerca das contribuições do modelo de dados em aplicações de mineração de dados.

7.4 TRABALHOS FUTUROS

Com o produto dessa dissertação, são sugeridas algumas pesquisas que podem vir a complementá-lo ou serem derivados do mesmo:

- Aplicações com algoritmos de mineração de dados diferentes dos utilizados no estudo de caso;
- Aplicações dos mesmos algoritmos utilizados com outras bases de dados, de instituições de ensino distintas;
- Utilização do modelo em outras situações-problema da área educacional, além do problema da evasão de estudantes como, por exemplo, a segmentação do público-alvo de cursos EAD, a relação dos objetos de aprendizagem no curso com o aprendizado do estudante, entre outros;
- Análise comparativa dos dados de estudantes de instituições de ensino distintas, consolidados num data set multi-institucional, à luz do modelo proposto.

Uma vez que o produto do trabalho atual apresenta uma sugestão de terminologia referencial para trabalhos de mineração de dados educacionais, espera-se que novas análises possam ser realizadas com a sua utilização, incluindo sugestões de evolução, tornando-o mais completo.

Entende-se também que, com uma convergência de terminologias, entidades e atributos, proporcionada pela utilização de um modelo de dados comum, seja possível a análise transversal comparativa de resultados de pesquisas distintas, produzindo mais e melhores informações para o apoio à tomada de decisões da área educacional.

REFERÊNCIAS

ABED. *Censo ABED 2011 – Relatório analítico da aprendizagem a distância no Brasil.* 2011. Disponível em: <http://www.abed.org.br/site/pt/midiateca/censo_ead/>. Acesso em 13 set. 2011.

ALVES, F. F. *Reconhecimento de Imagens 2D:* utilizando um modelo estatístico de formas. Santo André: Universidade Federal do ABC, 2012.

BAKER, R. S.; YACEF, K. *The State of Educational Data Mining in 2009:* a Review and Future Visions. 2009. Disponível em: <http://www.educationaldatamining.org/JEDM/images/articles/vol1/issue1/JEDMVol1Issue1_BakerYacef.pdf>. Acesso em 11 set.2011.

BAKER, R. S. J. de; ISOTANI, S.; CARVALHO, A. M. J. B. de. Mineração de dados educacionais: oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 2, 2011.

BATINI, C.; SCANNAPIECA, M. *Data Quality:* concepts, methodologies and tecniques. Springer-Verlag, 2006. ISBN: 3-540-33172-7.

BATISTA, Gustavo Enrique de A. P. A. B. *Pré-processamento de dados em aprendizado de máquina supervisionado.* Tese (Doutorado Ciência da Computação e Matemática Computação)- USP – São Carlos, 2003. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/publico/TeseDoutorado.pdf> Acesso em: 9 mar.2013.

CAMILO, Cássio; DA SILVA, João. *Mineração de dados:* conceitos, tarefas, métodos e ferramentas. [S.l]: Instituto de Informática, Universidade Federal de Goiás, 2009.

CAMPANA, *et al. Agentes para apoiar o acompanhamento das atividades em ambientes virtuais de aprendizagem*. Vitória–ES: Universidade Federal do Espírito Santo (UFES), 2006. Disponível em: <http://sbie2008.virtual.ufc.br/CD_ROM_COMPLETO/sbie_artigos_completo/Agentes%20para%20Apoiar%20o%20Acompanhamento%20das%20Atividades%20em.pdf>. Acesso em 23 jan.2012.

CESTARO, R.; PIVETTA, L. C. mineração de dados aplicada à identificação de alunos propensos à evasão no Ceulji/Ulbra de Ji-Paraná-Ro. *Ciência & Consciência,* v. 1, 2006.

CHEN, M.; HAN, J.; YU, P. S. *Data Mining:* an overview from database perspective. 1996. Disponível em: <http://dl.acm.org/citation.cfm?id=627794>. Acesso em 11 set. 2011.

CISLAGHI, R. *Um modelo de sistema de gestão do conhecimento em um framework para a promoção da permanência discente no ensino de graduação*. 2008. 273 f. Tese (Doutorado) - Universidade Federal de Santa Catarina, 2008.

DEKKER, G. W. *Predicting students drop out:* a case study. Holanda: Department of Electrical Engineering, Eindhoven University of Technology, 2009.

DIAS, M. M. *et al.* Aplicação de técnicas de mineração de dados no processo de aprendizagem na educação a distância. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 19., 2008. Fortaleza. *Anais...* p. 105-114, 2008.

DIAS, M. M. *Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados.* 2001. 212 f. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina – UFSC, Florianópolis, 2001.

DIMOKAS *et al*. *A Prototype System for Educational Data Warehousing and Mining*. 2008. Disponível em: <http://delab.csd.auth.gr/papers/PCI08dmna.pdf>. Acesso em 16 jun. 2012.

DOMINGUES, S. D. G.; GUILHERME, I. R. Uma ontologia para modelagem de conteúdo para ensino a distância. In: SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL, 2008, Rio de Janeiro. *Anais...* Universidade Federal Fluminense. Departamento de Ciência da Informação, Niterói, · Rio de Janeiro, 2008. Disponível em: <http://www.uff.br/ontologia/artigos/317.pdf>. Acesso em 19 nov.2011.

FAVERO, R. V. M.; FRANCO, S. R. K. Um estudo sobre a permanência e a evasão na Educação a Distância. *CINTED-UFRGS*, v. 4, n. 2, Dezembro, 2006. Disponível em: <http://www.cinted.ufrgs.br/renote/dez2006/artigosrenote/25103.pdf>. Acesso em: 23 jan. 2012.

FAYYAD, Usama; SHAPIRO, Gergory P.; SMYTH, Padhraic. *The KDD Process for Extracting Useful Knowledge from Volumes of Data.* ACM Press, New York, v. 39 , n. 11, p. 24-26, nov. 1996.

FERREIRA Jr. *et al.* Datawarehouse e Mineração de Dados nos institutos superiores da FAETEC. 2006. Disponível em: <http://www.faetec.rj.gov.br/ist-rio/images/pdf/02-2006.pdf>. Acesso em 11 jun. 2012.

FREITAS, C. M. D. S. *et al.* Introdução à Visualização de Informações. *RITA* , v. 7, n. 2, 2001.

FU, Y.; HAN, J. Meta-ruled-guided Mining of Association Rules in Relational Databases. 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.5283&rep=rep1&type=pdf>. Acesso em 11 set.2011.

GOEBEL, M.; GRUENWALD, L. A Survey of Data Mining and Knowledge Discovery Software Tools. *SIGKDD Explorations*, ACM SIGKDD, v. 1, p. 20, jun.1999.

GOMES, Alan Keller. *Análise do conhecimento extraído de classificadores simbólicos utilizando medidas de avaliação e de interessabilidade*. 2002. 151 f. Dissertação (Mestrado) Ciências de Computação e Matemática Computacional - Instituto de Ciências Matemáticas e

de Computação, Universidade de São Paulo, São Carlos, 2002. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-04072002-144610/>. Acesso em 9 mar. 2013.

HAJRA, A.; BIRANT, D., KUT, A. Improving Quality Assurance. In: *Education* with Web-based Services by Data Mining and Mobile Technologies, 2008.

HAN, Jiawei; KAMBER, Micheline. *Data Mining:* concepts and techniques. 2. ed. San Francisco: Morgan Kaufmann Publishers; Elsevier, 2006.

HAND, D.; MANNILA, H.; SMYTH, P. *Principles of Data Mining.* USA: Massachusets Institute of Technology, 2001.

KAMPFF, A. J. C. *Mineração de Dados Educacionais para Geração de Alertas em Avas.* 2009. 186 f. Tese (Doutorado)- Programa de Pós-graduação em Informática na Educação. Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

LIMA, L. M. *Mineração de dados utilizando algoritmos genéticos adaptativos.* 2009. 81 f. Monografia (Graduação). Ciências da Computação. Universidade Federal da Bahia – UFBA, Salvador, 2009.

MCCUBBIN, I. An Examination of Criticisms made of Tinto's 1975 Student Integration Model of Attrition. 2003. Disponível em: <http://www.psy.gla.ac.uk/~steve/localed/icubb.pdf>. Acesso em: 19 fev. 2013.

MESSAOUD, R. B. et.al. *Enhanced Mining of Association Rules from Data Cubes. 2006.* Disponível em: <http://eric.univ-lyon2.fr/~sabine/dolap06.pdf>. Acesso em: 12 set. 2011.

MUYINDA, P. *Distance_Education, Intech*: A free online edition of this book is available. 2012. Disponível em: <www.intechopen.com>. Acesso em: 26 set. 2012.

NASCIMENTO, H. A. D. D.; FERREIRA, C. B. R. Visualização de Informações: uma Abordagem Prática. A Universalidade da Computação: um agente de inovação e conhecimento. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25., São Leopoldo. *Anais...* São Leopoldo: UNISINOS, p. 1262-1312, jul. 2005.

OLSON, J. E. *Data Quality*: The accuracy dimension. San Francisco, CA: Morgan Kaufmann Publishers, 2003. ISBN: 1-55860-891-5.

ORR, K. Data Quality and Systems Theory. *Communications of the ACM*, v. 41, n. 2, fevereiro 1998.

PARKER, A. A Study of Variables That Predict Dropout from Distance Education, *International Journal of Educational Technology*, Gonzaga University, EUA, v.1, n.2, p1-10, Dec, 1999.

PARTHASARATHY, S. *Data mining at the crossroads: successes failures and learning from them*. 2007. Disponível em: <https://getinfo.de/app/Data-Mining-at-the-Crossroads-Successes-Failures/id/BLCP%3ACN068163648>. Acesso em: 18 set. 2011.

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. 2011. 153 f. Dissertação (Mestrado) – Computação Aplicada. Universidade do Vale do Itajaí, São José-SC, 2011.

PIMENTEL, E. P.; FRANÇA, V. F. D., OMAR, N. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - NCE - IM/UFRJ, 14., **Anais...** 2003.

PRETO, D.; SILVEIRA, S. R. Um estudo de caso da aplicação de mineração de dados em uma instituição de ensino superior. 2010. Disponível em: <http://ensino.univates.br/~cetec/wet/anais2010/C08-wet2010.pdf>. Acesso em 15 nov. 2011.

PYLE Dorian. *Data Preparation for Data Mining.* San Francisco: Editora Morgan Kaufmann, 1999. ISBN 1558605290, 9781558605299.

ROKACH, L. et al.“Series in Machine Perception and Artificial Intelligence. *Data Mining with Decision Trees* – Theory and Applications. ISBN-13 978-981-277-171-1, World Scientific Publishing Co. Pte. Ltd. v. 69, 2008.

ROMERO, C.; VENTURA, S. *Educational Data Mining:* a Survey from 1995 to 2005. 2006. Disponível em: <http://www.academia.edu/2662296/ Educational_data_mining_A_survey_from_1995_to_2005>. Acesso em 18 set. 2011.

ROMERO, C. et al. Data Mining Algorithms to Classify Students, Computer Science Department, Córdoba University, Spain. In: INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 1., 2007. *Proceedings.*… 2007.

ROMERO, C.; VENTURA, S.; GARCÍA, E. *Data Mining in Course Management Systems*: Moodle Case Study and Tutorial. 2007. Disponível em: <http://sci2s.ugr.es/keel/pdf/keel/articulo/CAE-VersionFinal.pdf>. Acesso em 18 set. 2011.

SFERRA, H. H.; CORRÊA, Â. M. C. J. Conceitos e Aplicações de Data Mining. *Revista de ciência & tecnologia,* v. 11, n. 22. p. 19-34, 2003.

SILVA, D. R. et al. *Using Data Warehouse and Data Mining Resources for Ongoing Assessment of Distance Learning*. 2002. Disponível em: <http://130.203.133.150/viewdoc/summary?doi=10.1.1.104.5146> . Acesso em 10 jun. 2012.

SODOVNIKOVA, D.; NIEDRNTE, L. *Using data warehouse resources for assessment of e-learning influence on university processes*. [S.l.]: [s.n.], 2005.

TAYI, G. T.; BALLOU, D. P. Examining Data Quality. *Communications of the ACM*, v. 41, n. 2, fevereiro 1998.

TINTO, V.; PUSSER, B. *Moving from theory to action building a model of institutional action for student success*. 2006. Disponível em: <http://web.ewu.edu/groups/academicaffairs/IR/NPEC_5_Tinto_Pusser_Report.pdf>. Acesso em 22 fev. 2013.

TORI, R.. *Avaliando Distâncias na Educação, Escola Politécnica.* São Paulo: USP, 2002.

TYLER-SMITH, K. Early Attrition among First Time eLearners: A Review of Factors that Contribute to Drop-out, Withdrawal and Non-completion Rates of Adult Learners undertaking eLearning Programmes. *Journal of Online Learning and Teaching*, Christchurch Polytechnic Institute of Technology, Christchurch, New Zealand, 2006. Disponível em: <http://jolt.merlot.org/Vol2_No2_TylerSmith.htm>. Acesso em: 22 abr. 2013.

WITTEN, I.H.; FRANK, E. *Data mining:* practical machine learning tools and techniques. 2. ed. San Francisco: Morgan Kaufmann Publishers; Elsevier, 2005.

ANEXO A – Convite aos pesquisadores para participação no Estudo de Caso

Prezado pesquisador,

Gostaria de convidá-lo a participar de um estudo que envolve o uso de mineração de dados educacionais. A partir de alguns modelos conceituais teóricos de pesquisas da área de educação e trabalhos sobre mineração de dados educacionais, foi proposto um modelo de dados que objetiva consolidar as principais entidades e atributos necessários em análises pedagógicas, acadêmicas e administrativas de atividades educacionais.

Para a avaliação do modelo proposto, estamos realizando um estudo de caso onde algumas medições deverão ser realizadas por profissionais de desenvolvimento de sistemas. Dessa forma, gostaríamos de contar com a sua colaboração para as atividades discriminadas a seguir.

1) Leitura das Orientações para a Realização do Trabalho que caracterizam o cenário do problema a ser analisado com a utilização dos modelos de dados.
2) Medição do horário de início do trabalho.
3) Análise do Modelo de Dados e Dicionário fornecido, no sentido de identificar as informações descritas nas Orientações.
4) Medição do horário de conclusão do trabalho.
5) Preenchimento do Perfil do Pesquisador no formulário anexo, descrevendo brevemente o seu perfil profissional. Essas informações servirão para caracterizar os participantes deste estudo de caso.
6) Preenchimento das informações de Aferição dos Resultados no formulário, descrevendo o esforço necessário e suas impressões sobre o trabalho.

É muito importante que você execute as etapas 3 a 5 sem interrupção para um resultado válido do trabalho.

Agradecemos imensamente a sua contribuição, ao passo em que nos comprometemos a compartilhar as conclusões da experiência realizada.

Péricles Nogueira Magalhães Júnior.

ANEXO B – Orientações aos pesquisadores para participação no Estudo de Caso

ORIENTAÇÕES PARA A REALIZAÇÃO DO TRABALHO

Contexto

Educação a distância pode ser definida como uma forma sistematicamente organizada de autodidatismo na qual o estudante se instrui a partir do material de estudo que lhe é apresentado com o acompanhamento e supervisão de tutores e professores. De maneira simplificada, alguns autores definem a educação a distancia como um formato de educação onde o professor está geograficamente distante dos seus estudantes.

O mercado da educação tem se tornado cada vez mais competitivo e global, procurando apoio em tecnologias e processos para atender às pressões populacionais, com necessidades de uma educação continuada. No Brasil, com a incorporação da modalidade da educação a distância na Lei de Diretrizes e Bases (Lei 9.394, de 20 de dezembro de 1996), transformando-a definitivamente em política pública de inclusão educacional, a oferta de cursos EAD em instituições de ensino superior públicas e privadas, regional e nacionalmente tem aumentado acima dos indicadores de crescimento econômico do país. Esse crescimento tem proporcionado, também, um aumento na preocupação com a qualidade do aprendizado dos seus estudantes e, consequentemente, pesquisas sistemáticas de teorias pedagógicas e tecnológicas que a suportem.

A retenção de estudantes nos seus cursos representa, atualmente, o maior problema de instituições de ensino superior, de acordo com o Censo ABED 2011. Diferentes conceitos de evasão encontrados na literatura, mas, de um modo geral, caracteriza-se por evasão a não continuidade de um estudante em um curso no qual havia se inscrito anteriormente, ou seja, a evasão é um ato de desistência. Uma evasão pode ser declarada, com o cancelamento ou trancamento da matrícula, ou silenciosa, com o simples abandono das atividades do curso.

De acordo com os respondentes do censo, as razões mais comuns apontadas para justificar a evasão estão relacionadas ao tempo de estudo (falta de tempo para estudar, 30,8%; acúmulo de atividades no trabalho, 24,5%; e viagens a trabalho com 9,1%), que constitui, juntamente com o aspecto financeiro, os principais argumentos utilizados pelas instituições para captar estudantes para os cursos da modalidade.

Análise

Entendendo que o fenômeno da evasão não acontece de forma instantânea, e sim gradual, a partir de certas características dos estudantes evadidos, pressupõe-se que, uma vez mapeadas essas condições, é possível identificar com antecedência predisposições a evadir e trabalhar institucionalmente na sua mitigação.

A proposta do presente trabalho é identificar no modelo fornecido atributos que possam caracterizar, isoladamente ou em conjunto, indícios de evasão de estudantes em cursos de graduação à distância. Para atingir o objetivo, solicitamos, na sua pesquisa:

1) A análise do modelo de dados disponibilizado, procurando, nas suas tabelas, os 10 principais atributos que, pelas explicações realizadas, podem estar relacionadas direta ou indiretamente, com a evasão de estudantes.
2) Registrar cada atributo identificado na seção correspondente do formulário anexo, com uma pequena justificativa sobre a sua escolha.

ANEXO C – Formulário para coleta de informações do Estudo de Caso

FORMULÁRIO - DADOS PARA COLETA				
PARTE 1 - PERFIL DO PESQUISADOR				
NOME:				
IDADE:		**ESCOLARIDADE:**		
TEMPO DE EXPERIÊNCIA PROFISSIONAL:				
Tem experiência profissional com modelagem de dados?			( ) SIM	( ) NÃO
Tem experiência profissional com mineração de dados?			( ) SIM	( ) NÃO
Já fez algum curso EAD ou já trabalhou com EAD?			( ) SIM	( ) NÃO
DATA E HORA DE INÍCIO DA ANÁLISE:				
PARTE 2 - ATRIBUTOS IDENTIFICADOS				
ENTIDADE	**ATRIBUTO**	**JUSTIFICATIVA**		

COMENTÁRIOS GERAIS		
PARTE 3 – ESFORÇO REQUERIDO		
HORA DE CONCLUSÃO:		**TEMPO DECORRIDO**
COMENTÁRIOS SOBRE O ESFORÇO		

ANEXO D – Consultas utilizadas na extração dos dados sem o modelo proposto.

Consulta SQL 1, para extrair os dados diretos da base de dados acadêmica:

```
INSERT INTO Temporaria
(IdEstudante, Periodo, Curso, Sexo, Idade, NomeCidade, BairroEstudo,
PendenciasAcademicas, DisciplinasEmCurso, DiscplinasConcluidas,
PeriodosConcluidos, TipoIngresso, IngressoNota,
IngressoAntecipacaoMatricula, IndicadorEvasao)
SELECT Al.Nu_Matricula AS IdEstudante, PL.Sigla_PeriodoLetivo AS
Periodo, Cur.Nom_Curso AS Curso, IIF(Pes.Sexo ="F",1,2) As Sexo,
NULL, Cid.Nom_Cidade AS NomeCidade, Cam.Bairro_Campus AS
BairroEstudo, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL
FROM Campus Cam INNER JOIN (
  (Cidade Cid INNER JOIN Pessoa Pes
    ON Cid.Id_Cidade = Pes.Id_Cidade) INNER JOIN
   (Periodo_Letivo PL INNER JOIN
    (Curso Cur INNER JOIN Aluno Al
       ON Cur.Id_Curso = Al.Id_Curso)
    ON PL.Id_PeriodoLetivo = Al.Id_PeriodoLetivo)
  ON Pes.Id_Pessoa = Al.ID_Pessoa)
ON Cam.Id_Campus = Al.Id_Campus
WHERE (Al.Id_PeriodoLetivo IN (1, 2, 3) AND
Al.Id_PeriodoLetivo=Al.Id_PeriodoLetivoIngresso);
```

Consulta SQL 2, para extrair os dados diretos da base de dados de Processos Seletivos

```
UPDATE Temporaria T INNER JOIN
( SELECT C.Cand_Nu_Matricula AS IdEstudante, C.Cand_TipoProva
  AS TipoIngresso, V.Vest_Apelido AS Periodo
  FROM tb_Vestibular AS V INNER JOIN
    tb_Candidato AS C ON V.Vest_Id = C.Vest_Id
  WHERE V.Vest_Id IN (1, 2, 3)
) Aux ON
T.IdEstudante = Aux.IdEstudante AND T.Periodo=Aux.Periodo
SET T.TipoIngresso=Aux.TipoIngresso;
```

Consulta SQL 3, para extrair e calcular Idade e Indicador de Evasão da base de dados de Acadêmica:

```
UPDATE Temporaria T INNER JOIN
(
  SELECT A.Nu_Matricula AS IdEstudante, PL.Sigla_PeriodoLetivo
  AS Periodo, INT((Now()-P.Dat_Nascimento)/365) AS Idade,
  IIf(S.Des_Situacao="ATIVO","NÃO","SIM") AS IndicadorEvasao
  FROM
  (    Situacao_Aluno S INNER JOIN
      (Pessoa P INNER JOIN Aluno A ON P.Id_Pessoa = A.ID_Pessoa
      ) ON S.Id_Situacao = A.Id_SituacaoAluno
  ) INNER JOIN Periodo_Letivo PL ON
     A.Id_PeriodoLetivo = PL.Id_PeriodoLetivo
  WHERE (((A.Id_PeriodoLetivo) IN (1,2,3)
  AND (A.Id_PeriodoLetivo)=[A].[Id_PeriodoLetivoIngresso]))
) AUX
ON T.IdEstudante=Aux.IdEstudante AND T.Periodo=Aux.Periodo
SET T.Idade = Aux.Idade,
T.IndicadorEvasao=Aux.IndicadorEvasao;
```

Consulta SQL 4, para extrair e calcular Nota e Antecipação de Matrícula da base de dados de Processos Seletivos:

```
UPDATE Temporaria T INNER JOIN
( SELECT C.Cand_Nu_Matricula AS IdEstudante,
  V.Vest_Apelido AS Periodo, IIf(C.Cand_TipoProva="ENEM",
  C.Cand_Nota/10, C.Cand_Nota/1000) AS IngressoNota,
  M.Marc_DataLimiteMatricula-C.Cand_DataMatricula AS
  IngressoAntecipacao FROM
  (  tb_Marcacao M INNER JOIN tb_Candidato C
     ON M.Marc_Id = C.Marc_Id
  ) INNER JOIN tb_Vestibular V ON M.Vest_Id = V.Vest_Id
  WHERE (((C.Vest_Id) IN (1,2,3)))
) AUX ON T.IdEstudante = AUX.IdEstudante
```

```
  AND T.Periodo=AUX.Periodo
SET T.IngressoNota = AUX.IngressoNota,
T.IngressoAntecipacaoMatricula = AUX.IngressoAntecipacao;
```

Consulta SQL 5a, para extrair os dados calculados da base de dados de Acadêmicos:

```
INSERT INTO Temporaria2 (IdEstudante, Periodo, PeriodosConcluidos,
DisciplinasEmCurso, DisciplinasConcluidas, PendenciasAcademicas)
SELECT A.Nu_Matricula, PL.Sigla_PeriodoLetivo,
A.SemestreAtual-1, SUM(IIf(SM.Des_Situacao="Em Curso",1,0)),
SUM(IIf(SM.Des_Situacao="Concluida",1,0)),
SUM(IIf(SM.Des_Situacao<>"Concluida" AND
AC.Semestre_Atividade_Curricular<A.SemestreAtual,1,0)) FROM
( Situacao_Matricula SM INNER JOIN
  ( ( Atividade_Curricular AC INNER JOIN Classe C ON
     AC.Id_Atividade_Curricular = C.Id_Atividade_Curricular
     ) INNER JOIN
    ( Aluno A INNER JOIN Matricula M ON
      A.Nu_Matricula = M.Nu_Matricula
    ) ON C.Id_Classe = M.Id_Classe
  ) ON SM.Id_Situacao = M.Id_Situacao_Matricula
) INNER JOIN Periodo_Letivo AS PL ON
A.Id_PeriodoLetivo = PL.Id_PeriodoLetivo
WHERE (((A.Id_PeriodoLetivo) In (1,2,3) And
(A.Id_PeriodoLetivo)=[A].[Id_PeriodoLetivoIngresso]))
GROUP BY A.Nu_Matricula, PL.Sigla_PeriodoLetivo, A.SemestreAtual-1;
```

Consulta SQL 5b, para atualizar os dados calculados na tabela temporária:

```
UPDATE Temporaria T INNER JOIN (
SELECT IdEstudante, Periodo, PeriodosConcluidos, DisciplinasEmCurso,
DisciplinasConcluidas, PendenciasAcademicas FROM Temporaria2) AUX ON
(T.Periodo = AUX.Periodo) AND (T.IdEstudante = AUX.IdEstudante) SET
T.PeriodosConcluidos = AUX.PeriodosConcluidos, T.DisciplinasEmCurso =
AUX.DisciplinasEmCurso, T.DisciplinasConcluidas =
AUX.DisciplinasConcluidas, T.PendenciasAcademicas =
AUX.PendenciasAcademicas;
```

Consulta SQL 6, para extrair os dados para o Caso 1, a partir do *dataset* temporário:

```
SELECT T.Periodo, IIF(T.Curso="C01","LICENCIATURA”, IIF(T.Curso=
"C02","BACHARELADO", IIF(T.Curso="C03","TECNOLOGICO"))) AS Curso,
T.TipoIngresso, T.IngressoNota, T.IngressoAntecipacaoMatricula,
T.Sexo, T.Idade, IIF(T.NomeCidade="SALVADOR","CAPITAL","INTERIOR") AS
Cidade, IIF(T.BairroEstudo="CENTRO","CENTRO","OUTRO") AS BairroEstudo,
T.PendenciasAcademicas, T.DisciplinasEmCurso, T.PeriodosConcluidos,
T.IndicadorEvasao
FROM Temporaria T;
```

Consulta SQL 7, para extrair os dados para o Caso 2, a partir do *dataset* temporário:

```
SELECT T.Periodo, IIF(T.Curso="C01","LICENCIATURA”, IIF(T.Curso=
"C02","BACHARELADO", IIF(T.Curso="C03","TECNOLOGICO"))) AS Curso,
T.TipoIngresso, IIF(ISNULL(T.IngressoNota),"SEM NOTA",
IIF(T.IngressoNota<0.6,"<60%",">=60%")) AS IngressoNota,
IIf(T.IngressoAntecipacaoMatricula Is Null Or
T.IngressoAntecipacaoMatricula=0, "SemAntecipacao",
    IIf(T.IngressoAntecipacaoMatricula<=7, "AtéUmaSemana",
IIf(T.IngressoAntecipacaoMatricula<=15,"AtéDuasSemanas",
IIf(T.IngressoAntecipacaoMatricula<=30,"AtéUmMês","MaisDeUmMês")))) AS
IngressoAntecipacaoMatricula, Sexo,
IIF(T.Idade<=25,"Até25Anos",IIF(T.Idade<=35,"de25a35Anos","Maisque35an
os")) AS Idade, IIF(T.NomeCidade="SALVADOR","CAPITAL","INTERIOR") AS
Cidade, IIF(T.BairroEstudo="CENTRO","CENTRO","OUTRO") AS BairroEstudo,
T.PendenciasAcademicas, T.DisciplinasEmCurso, T.PeriodosConcluidos,
IIF(T.IndicadorEvasao ="NÃO", 0, 1) AS IndicadorEvasao
FROM Temporaria AS T;
```

ANEXO E– Consultas utilizadas na extração dos dados com o modelo proposto.

Consulta SQL 8, para extrair os dados para o Caso 1, a partir do modelo proposto:

```
SELECT P.Ano & P.Sequencial AS Periodo, C.Tipo AS Curso,
M.TipoIngresso, M.NotaIngresso, M.AntecipacaoMatricula, E.Sexo,
INT((Now() - E.DataNascimento)/365) AS Idade,
IIF(I.Cidade="SALVADOR","CAPITAL","INTERIOR") AS Cidade,
IIF(I.Bairro="CENTRO","CENTRO","OUTRO") AS Bairro,
M.PendenciasAcademicas, M.DisciplinasEmCurso,
M.DisciplinasConcluidas, M.PeriodosConcluidos, M.SituacaoMatricula
FROM Instalacao I INNER JOIN
  (Curso C INNER JOIN
   (Periodo P INNER JOIN
     (Estudante E INNER JOIN Matricula M
      ON E.IdEstudante = M.IdEstudante)
    ON P.IdPeriodo = M.IdPeriodoMatricula)
  ON C.IdCurso = M.IdCurso)
ON I.IdInstalacao = M.IdInstalacao
WHERE P.IDPeriodo IN (1, 2, 3) AND M.IdPeriodoMatricula =
E.IdPeriodoIngresso;
```

Consulta SQL 9, para extrair os dados para o Caso 2, a partir do modelo proposto:

```
SELECT P.Ano & P.Sequencial AS Periodo, C.Tipo AS Curso,
M.TipoIngresso,
IIf(M.NotaIngresso IS Null,"SemNota",
IIf(M.NotaIngresso<0.6,"<60%",">=60%")) AS IngressoNota,
IIf(M.AntecipacaoMatricula Is Null Or M.AntecipacaoMatricula=0,
"SemAntecipacao",
IIf(M.AntecipacaoMatricula<=7, "AtéUmaSemana",
IIf(M.AntecipacaoMatricula<=15,"AtéDuasSemanas",
IIf(M.AntecipacaoMatricula<=30,"AtéUmMês","MaisDeUmMês")))) AS
IngressoAntecipacaoMatricula, [E].Sexo,
IIF(INT((Now()-E.DataNascimento)/365)<=25,"Até25Anos", IIF(INT((Now()-
E.DataNascimento)/365)<=35,"de25a35Anos", "Maisque35anos")) AS Idade,
IIF(I.Cidade="SALVADOR","CAPITAL","INTERIOR") AS Cidade,
IIF(I.Bairro="CENTRO","CENTRO","OUTRO") AS Bairro,
```

```
M.PendenciasAcademicas, M.DisciplinasEmCurso, M.DisciplinasConcluidas,
M.PeriodosConcluidos, IIF(M.SituacaoMatricula="Matriculado", 0, 1) AS
IndicadorEvasao
FROM Instalacao AS I INNER JOIN
  (Curso AS C INNER JOIN
    (Periodo AS P INNER JOIN
      (Estudante AS E INNER JOIN Matricula AS M
      ON [E].IdEstudante=M.IdEstudante)
    ON P.IdPeriodo = M.IdPeriodoMatricula)
   ON C.IdCurso = M.IdCurso)
ON I.IdInstalacao = M.IdInstalacao
WHERE P.IDPeriodo IN (1, 2, 3)
AND M.IdPeriodoMatricula = E.IdPeriodoIngresso;
```

ANEXO E – Árvore de Decisão completa gerada no estudo de caso

```
J48 pruned tree
------------------
Sexo = 1
|  Periodo = 20111
|  |  DisciplinasCursadas = 0: NAO (0.0)
|  |  DisciplinasCursadas = 1: NAO (0.0)
|  |  DisciplinasCursadas = 2: NAO (0.0)
|  |  DisciplinasCursadas = 3: NAO (0.0)
|  |  DisciplinasCursadas = 4: NAO (0.0)
|  |  DisciplinasCursadas = 5: NAO (0.0)
|  |  DisciplinasCursadas = 6
|  |  |  NomeCidade = CAPITAL: SIM (66.0/26.0)
|  |  |  NomeCidade = INTERIOR: NAO (116.0/47.0)
|  |  DisciplinasCursadas = 7: NAO (78.0/26.0)
|  Periodo = 20112
|  |  IngressoAntecipaçãoMatricula = SemAntecipacao
|  |  |  BairroEstudo = CENTRO: SIM (9.0/2.0)
|  |  |  BairroEstudo = IGUATEMI: NAO (3.0)
|  |  |  BairroEstudo = PARALELA: SIM (1.0)
|  |  IngressoAntecipaçãoMatricula = AtéUmaSemana: SIM (77.0/23.0)
|  |  IngressoAntecipaçãoMatricula = AtéDuasSemanas
|  |  |  BairroEstudo = CENTRO: NAO (9.0/1.0)
|  |  |  BairroEstudo = IGUATEMI: NAO (9.0/3.0)
|  |  |  BairroEstudo = PARALELA: SIM (5.0/1.0)
|  |  IngressoAntecipaçãoMatricula = AtéUmMês
|  |  |  BairroEstudo = CENTRO: NAO (15.0/5.0)
|  |  |  BairroEstudo = IGUATEMI: SIM (12.0/2.0)
|  |  |  BairroEstudo = PARALELA
|  |  |  |  Idade = Até25anos: SIM (0.0)
|  |  |  |  Idade = de25a35anos: SIM (3.0)
|  |  |  |  Idade = Maisque35anos
|  |  |  |  |  IngressoNota = SemNota: NAO (0.0)
|  |  |  |  |  IngressoNota = <6: NAO (4.0)
|  |  |  |  |  IngressoNota = >=6: SIM (3.0/1.0)
|  |  IngressoAntecipaçãoMatricula = MaisDeUmMês
|  |  |  Idade = Até25anos
|  |  |  |  TipoIngresso = VEST
```

```
| | | | | IngressoNota = SemNota: NAO (0.0)
| | | | | IngressoNota = <6: SIM (3.0/1.0)
| | | | | IngressoNota = >=6: NAO (2.0)
| | | | TipoIngresso = ENEM: SIM (3.0)
| | | | TipoIngresso = ME: SIM (0.0)
| | | | TipoIngresso = TE: SIM (0.0)
| | | | TipoIngresso = TI: SIM (0.0)
| | | | TipoIngresso = PROUNI: SIM (0.0)
| | | Idade = de25a35anos: NAO (25.0/6.0)
| | | Idade = Maisque35anos
| | | | BairroEstudo = CENTRO
| | | | | IngressoNota = SemNota: NAO (0.0)
| | | | | IngressoNota = <6: SIM (3.0/1.0)
| | | | | IngressoNota = >=6: NAO (2.0)
| | | | BairroEstudo = IGUATEMI: SIM (4.0/1.0)
| | | | BairroEstudo = PARALELA: NAO (4.0/1.0)
| Periodo = 20121: SIM (354.0/138.0)
Sexo = 2
| IngressoAntecipaçãoMatricula = SemAntecipacao
| | Periodo = 20111
| | | Curso = TECNOLOGICO: SIM (5.0/1.0)
| | | Curso = BACHARELADO: SIM (5.0/2.0)
| | | Curso = LICENCIATURA: NAO (2.0)
| | Periodo = 20112
| | | BairroEstudo = CENTRO: NAO (38.0/16.0)
| | | BairroEstudo = IGUATEMI
| | | | Curso = TECNOLOGICO: SIM (2.0)
| | | | Curso = BACHARELADO: SIM (1.0)
| | | | Curso = LICENCIATURA: NAO (3.0)
| | | BairroEstudo = PARALELA: SIM (4.0)
| | Periodo = 20121: SIM (41.0/12.0)
| IngressoAntecipaçãoMatricula = AtéUmaSemana
| | NomeCidade = CAPITAL
| | | Idade = Até25anos
| | | | Curso = TECNOLOGICO: SIM (1.0)
| | | | Curso = BACHARELADO: NAO (5.0/1.0)
| | | | Curso = LICENCIATURA
| | | | | BairroEstudo = CENTRO: NAO (0.0)
| | | | | BairroEstudo = IGUATEMI: NAO (3.0)
```

```
| | | | | BairroEstudo = PARALELA: SIM (2.0)
| | | Idade = de25a35anos
| | | | IngressoNota = SemNota: SIM (0.0)
| | | | IngressoNota = <6: SIM (17.0/5.0)
| | | | IngressoNota = >=6
| | | | | Curso = TECNOLOGICO: SIM (6.0/3.0)
| | | | | Curso = BACHARELADO: SIM (8.0/2.0)
| | | | | Curso = LICENCIATURA: NAO (5.0)
| | | Idade = Maisque35anos
| | | | BairroEstudo = CENTRO: NAO (0.0)
| | | | BairroEstudo = IGUATEMI: NAO (28.0/10.0)
| | | | BairroEstudo = PARALELA
| | | | | Curso = TECNOLOGICO: SIM (2.0)
| | | | | Curso = BACHARELADO: SIM (3.0/1.0)
| | | | | Curso = LICENCIATURA: NAO (2.0)
| | NomeCidade = INTERIOR: NAO (108.0/40.0)
| IngressoAntecipaçãoMatricula = AtéDuasSemanas: NAO (60.0/14.0)
| IngressoAntecipaçãoMatricula = AtéUmMês
| | Periodo = 20111: NAO (127.0/60.0)
| | Periodo = 20112
| | | TipoIngresso = VEST: NAO (116.0/34.0)
| | | TipoIngresso = ENEM: SIM (5.0/1.0)
| | | TipoIngresso = ME: NAO (0.0)
| | | TipoIngresso = TE: NAO (0.0)
| | | TipoIngresso = TI: NAO (0.0)
| | | TipoIngresso = PROUNI: NAO (0.0)
| | Periodo = 20121: SIM (25.0/8.0)
| IngressoAntecipaçãoMatricula = MaisDeUmMês: NAO (1481.0/540.0)
```